BAOU Education for All

Dr. Babasaheb Ambedkar Open University

(Established by Government of Gujarat)

NAAC ACCREDITED A++ GRADE 3.55/4

Regression Analysis and Predictive Models
MSCDS-201

Master of Science - Data Science (MSCDS)

2024

Regression Analysis and Predictive Models

Dr. Babasaheb Ambedkar Open University

Expert Committee

Prof. (Dr.) Nilesh Modi Professor and Director, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad	**(Chairman)**
Prof. (Dr.) Ajay Parikh Professor and Head, Department of Computer Science, Gujarat Vidyapith, Ahmedabad	**(Member)**
Prof. (Dr.) Satyen Parikh Dean, School of Computer Science and Application, Ganpat University, Kherva, Mahesana	**(Member)**
Prof. M. T. Savaliya Associate Professor and Head (Retired), Computer Engineering Department, Vishwakarma Engineering College, Ahmedabad	**(Member)**
Dr. Himanshu Patel Assistant Professor, School of Computer Science, Dr. Babasaheb Ambedkar Open University, Ahmedabad	**(Member Secretary)**

Course Writer

Dr. Hardik Soni
Professor & Director, Chimanbhai Patel Post-Graduate Institute of Computer Applications (MCA), Sardar Vallabhbhai Global University, Ahmedabad

Content Editor

Dr. Shivang M. Patel
Associate Professor, School of Computer Science,
Dr. Babasaheb Ambedkar Open University, Ahmedabad

Subject Reviewer

Prof. (Dr.) Nilesh Modi
Professor and Director, School of Computer Science,
Dr. Babasaheb Ambedkar Open University, Ahmedabad

ISBN- 978-81-984865-0-9

Printed and published by: Dr. Babasaheb Ambedkar Open University, Ahmedabad

BAOU Education for All
Dr. Babasaheb Ambedkar Open University
MSCDS-201

Regression Analysis and Predictive Models

Block 1: Foundations of Regression Analysis

Introduction

Regression analysis is one of the most powerful and widely used tools in statistics and data science, providing a framework for understanding relationships between variables and making data-driven predictions. Whether you are analyzing trends, building predictive models, or testing hypotheses, regression analysis serves as the foundation for many advanced analytical techniques. This block, *Foundations of Regression Analysis*, introduces you to the core principles of regression, starting with simple linear regression and gradually building your understanding of model evaluation, interpretation, and application.

In *Unit 1: Simple Linear Regression*, you will begin by exploring the fundamental concepts of regression analysis and its importance in statistical modeling. You will learn how to formulate a simple linear regression model, understand the key assumptions underlying it, and derive the Ordinary Least Squares (OLS) estimator. By the end of this unit, you will be able to solve the normal equations to obtain a closed-form solution for parameter estimation, laying the groundwork for more advanced topics.

Unit 2: Coefficients Calculation and Prediction builds on the concepts introduced in Unit 1, focusing on the practical application of the least squares method to calculate regression coefficients. You will learn how to interpret and visualize regression results, make predictions using the estimated regression equation, and validate the key assumptions of linear regression. Additionally, you will gain hands-on experience using R to compute regression coefficients, both manually and with built-in functions, ensuring you can apply these techniques in real-world scenarios.

In *Unit 3: Evaluating Model Fit*, you will delve into the critical task of assessing how well a regression model fits the data. You will learn to evaluate the alignment between observed data and model predictions, calculate and interpret the regression standard error, and understand the role of R-squared in measuring the proportion of variability explained by the model. These skills are essential for determining the accuracy and reliability of your regression models.

Finally, *Unit 4: Assessing the Strength of the Linear Relationship* focuses on understanding and interpreting the slope parameter in linear regression, which represents the relationship between the predictor and response variables. You will learn how to estimate and test the slope for significance, interpret p-values, and assess the strength of evidence for a linear association. This unit will equip you with the tools to evaluate the reliability of your model and draw meaningful conclusions from your analysis.

By the end of this block, you will have a solid understanding of the foundations of regression analysis, from model formulation and parameter estimation to evaluation and interpretation. Whether you are new to regression or looking to strengthen your foundational knowledge, this block will provide you with the skills and confidence to apply regression analysis effectively in your work.

Unit 1: Simple Linear Regression

Unit Structure

1.0 Learning Objectives

After going through this unit, you should be able to

- Understand the fundamental concepts of regression analysis.
- Explain the importance of regression analysis in statistical modelling.
- Understand a simple linear regression model
- Identify and understand the key assumptions underlying linear regression.
- Derive the Ordinary Least Squares (OLS) estimator
- Obtain Closed-form solution for parameter estimation: Normal equations

1.1 Introduction

Regression analysis plays a crucial role in both data science and managerial decision-making by quantifying the relationships between variables. It helps identify how independent (explanatory) variables influence a dependent (response) variable, such as understanding the impact of advertising expenditures on sales or predicting electricity demand based on daily temperatures. With its clear mathematical principles and high interpretability, regression analysis serves two main purposes: *explanatory analysis*, which explores how variables affect one another, and *predictive analysis*, which identifies the best combinations of variables for accurate forecasting. By providing a structured method to analyze these relationships, regression analysis is an essential tool for both data-driven insights and informed decision-making.

Regression analysis should not be viewed as a method for proving causation between variables. It can only show the extent to which variables are related. Any conclusions about cause and effect require the discernment of experts familiar with the specific context.

In regression analysis, the variable being predicted is called the response or dependent, denoted by y. The variables used to predict the response are referred to as explanatory or predictor or independent variables represented by x. For example, when examining the impact of advertising expenditures on sales, sales would be the response variable (y), and advertising expenditures would be the explanatory variable (x). In data science terms, y represents the target or output variable, while x represents the feature or input variables.

In this block, we explore the most basic form of regression analysis, which involves one independent variable and one dependent variable. Here, the relationship between the variables is represented by a straight line. It is called *simple linear regression*.

1.2 Uses of Regression

Regression models serve a variety of purposes in statistical analysis and data science, including the following:

1.2.1 Data Description

Regression analysis helps in understanding and describing the relationships between variables. For example, consider a study that examines the relationship between years of education and annual income. A regression model can describe how income tends to increase with additional years of education, providing a clearer picture of this relationship and helping to identify patterns in the data.

1.2.2 Parameter Estimation

Regression is used to estimate the parameters of the relationship between variables. In the education and income example, the regression model would estimate the coefficient that quantifies how much annual income is expected to increase for each additional year of education. This coefficient is crucial in understanding the strength and nature of the relationship, allowing for more precise interpretations.

1.2.3 Prediction and Estimation

One of the primary uses of regression models is to predict future outcomes based on current or past data. For instance, a company might use a regression model to forecast future sales based on historical sales data and marketing expenditure. By inputting the marketing expenditure, the model can estimate the expected sales, helping the company plan and make data-driven decisions.

1.2.4 Control

In experimental and observational studies, regression models help control confounding variables. For example, imagine a study analyzing the impact of exercise on weight loss, while also considering dietary habits. By including both exercise and dietary habits as independent variables in the regression model, researchers can isolate the effect of exercise on weight loss, ensuring that the observed relationship is not confounded by variations in dietary habits.

Regression models are indispensable tools in many disciplines, providing valuable insights and supporting data-driven decision-making. By understanding and leveraging these models, researchers and analysts can describe complex relationships, estimate key parameters, make accurate predictions, and control confounding factors.

Check Your Progress – 1

1. What is the primary purpose of regression analysis in data science and managerial decision-making?

(a)	To visualize data	(b)	To quantify the relationships between variables
(c)	To perform clustering	(d)	To clean the data

2. In regression analysis, what is the dependent variable also referred to as?

(a)	Predictor variable	(b)	Independent variable
(c)	Response variable	(d)	Confounding variable

3. Which of the following is NOT a use of regression models?

(a)	Data Description	(b)	Parameter Estimation
(c)	Prediction and Estimation	(d)	Data Encryption

4. In the context of regression analysis, what does the explanatory variable represent in a study examining the impact of advertising expenditures on sales?

(a)	Sales	(b)	Advertising expenditures
(c)	Confounding variables	(d)	Weight loss

1.3 Simple Linear Regression Model

The simple linear regression model establishes a linear relationship between a single explanatory variable (regressor) (x) and a response variable (y). It is expressed as:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

where:

- β_0 is the *intercept*, representing the value of (y) when ($x = 0$).
- β_1 is the *slope*, indicating how much (y) changes for a one-unit change in (x).
- ε is the *random error term*, capturing unobserved influences on (y).

Both β_0 and β_1 are constants to be estimated from the data and usually called *regression coefficients.*

1.3.1 Error Assumptions

In simple linear regression, the errors ε are assumed to:

1. Have a *mean of zero*: $\mathrm{E}[\varepsilon]=0$
2. Have *constant variance*: The error variance σ^2 remains the same across all values of x .
3. Be *uncorrelated*: The errors for different observations are independent of one another, meaning the error for one observation does not influence others.

1.3.2 Role of Variables

- The *regressor* (x) is treated as a fixed variable, controlled by the analyst, and measured with negligible error.
- The *response* (y) is considered a random variable, with its values depending on the distribution for each given value of (x).

1.3.3 Regression Equation

The equation that describes how the expected value of y, denoted $\mathrm{E}[y|x]$, is related to x is called the *regression equation*. Thus, using the error assumption (1), the regression equation for simple linear regression follows.

$$\mathrm{E}[y|x]=\beta_0+\beta_1 x \tag{1.2}$$

Moreover, the variance of y given x is constant and does not depend on x:

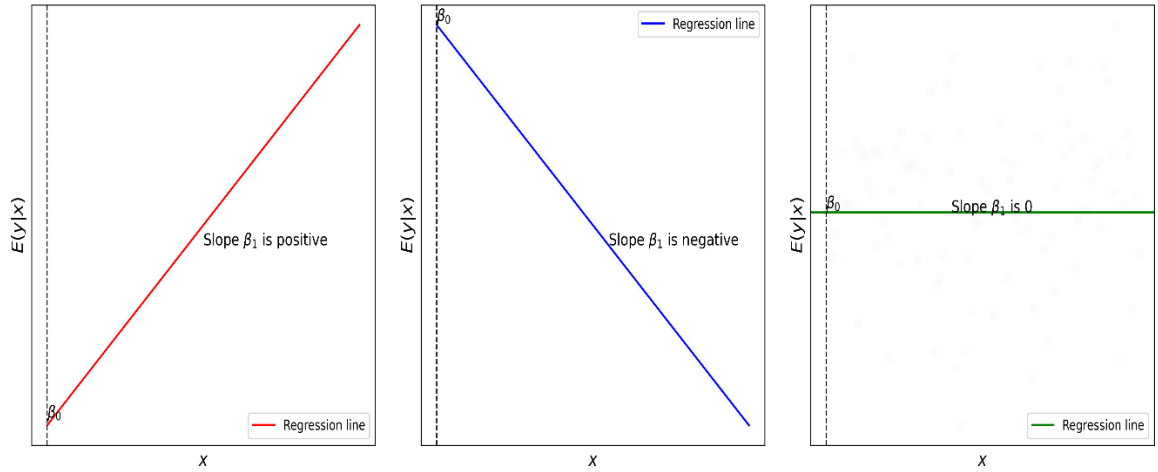
$$Var[y|x]=\sigma^2$$

which satisfies error assumption (2).

The graph of the simple linear regression equation is a straight line. The parameters β_1 and β_0 can be interpreted as follows.

- β_1 (*slope*): It indicates the expected change in the mean response y for a one-unit increase in the regressor x. For example, if $\beta_1 = 2$, it means that for each unit increase in x, the mean of y increases by 2.
- β_0 (*intercept*): It represents the mean value of y when $x = 0$. However, its practical interpretation is only meaningful if $x = 0$ is within the range of observed data. If $x = 0$ lies outside the data range, the intercept may not have a relevant or interpretable meaning.

Figure 1.1 illustrates different types of regression lines. In Panel A, the regression line indicates a positive relationship between y and x, where higher values of $\mathbb{E}[y|x]$ correspond to higher values of x. Panel B shows a negative relationship, with higher values of x leading to lower values of $\mathbb{E}[y|x]$. Panel C depicts a scenario where y is not related to x, meaning the mean value of y remains constant regardless of the value of x.



(a) Positive relationship (b) Negative relationship (c) No relationship

Figure 1.1: Possible Regression Lines in Simple Linear Regression

1.3.4 Estimated Regression Equation

If we knew the population parameters β_0 and β_1, we could use equation (1.2) to calculate the mean value of y for a given x. However, since these parameter values are typically unknown, they must be estimated from sample data. The sample statistics, also denoted as b_0 and b_1, are calculated to estimate the population parameters. By substituting these sample statistics into the regression equation, we obtain the *estimated regression equation* for simple linear regression. The equation can be given by

$$\hat{y} = b_0 + b_1 x \tag{1.3}$$

In general, $\hat{y}$ is the point estimator of $\mathbb{E}[y|x]$, the mean value of y for a given x.

Check Your Progress – 2

1. What does the intercept (β_0) in a simple linear regression model represent?

(a)	The value of the explanatory variable (x) when the response variable (y) is zero
(b)	The value of the response variable (y) when the explanatory variable (x) is zero
(c)	The slope of the regression line
(d)	The random error term

2. Which of the following is an error assumption in simple linear regression?

(a)	Errors have a mean of one
(b)	Errors have a constant variance across all values of x
(c)	Errors are correlated with one another
(d)	Errors depend on the value of y

3. In the regression equation $E(y|x) = \beta_0 + \beta_1 x$, what does β_1 represent?

(a)	The intercept	(b)	The error term
(c)	The change in the mean response y for a one-unit increase in the regressor x	(d)	The variance of y given x

4. In the context of regression analysis, what does the explanatory variable represent in a study examining the impact of advertising expenditures on sales?

(a)	To calculate the exact value of the population parameters β_0 and β_1
(b)	To predict the mean value of y for a given x using sample estimates
(c)	To describe the variance of the error term
(d)	To identify confounding variables in the data

1.4 Least Squares Method

The *Least Squares Method* is a standard approach used for finding the best-fitting line or curve to a set of data points. The method minimizes the sum of the squared differences between the observed values and the values predicted by the model. It is widely used in regression analysis to estimate the parameters of a model that describes the relationship between a dependent variable and one or more independent variables.

1.4.1 Least square criterion

The criterion for the least squares method is given by expression

$$\min \sum (y_i - \hat{y}_i)^2 \tag{1.4}$$

where, y_i = observed value of the dependent variable for the ith observation

$\hat{y}_i$ = predicted value of the dependent variable for the ith observation

To find the values of b_0 and b_1 that minimize the residual sum of squares (RSS), we take the partial derivatives of the RSS with respect to b_0 and b_1, and set them equal to zero. Substituting $\hat{y}$ from equation (1.3) into (1.4) yields,

$$RSS = \sum (y_i - b_0 - b_1 x_i)^2$$

The least squares estimators b_0 and b_1 must satisfy

$$\frac{\partial}{\partial b_0}\left(\sum (y_i - b_0 - b_1 x_i)^2\right) = -2\sum (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial}{\partial b_1}\left(\sum (y_i - b_0 - b_1 x_i)^2\right) = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0$$

Simplifying these two equations yields

$$n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{1.5}$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

Equations (1.5) are called *least square normal equations*. The solution of normal equations is

$$b_0 = \bar{y} - b_1 \bar{x} \tag{1.6}$$

and

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{1.4}$$

where, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Further, S_{xy} denotes covariance of x and y whereas S_{xx} denotes variance of x.

Note: An alternative formula for b_1 is

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

The least squares method yields an estimated regression equation that minimizes the sum of squared differences between the observed values of the dependent variable y_i and the predicted values $\hat{y}_i$. This criterion of least squares is employed to identify the equation that best fits the data. If an alternative criterion, such as minimizing the sum of absolute differences between y_i and $\hat{y}_i$, were used, a different equation would result.

In the next unit, we will walk through the step-by-step process of calculating regression coefficients (intercept and slope) using the example with a small dataset.

1.5 LET US SUM UP

This unit covered the basics of simple linear regression, which establishes a linear relationship between one independent variable (x) and one dependent variable (y) using the equation $y = \beta_0 + \beta_1 x + \varepsilon$. Regression analysis is used for describing data, estimating parameters, making predictions, and controlling confounding variables. The model's assumptions include errors having a mean of zero, constant variance, and no correlation. Parameters (β_0, β_1) are estimated from data to predict y values, enabling data-driven insights and decision-making. Least square method is introduced and closed form formula for the estimators are derived.

1.6 Check Your Progress: Possible Answers

Check Your Progress – 1

Question No.	Correct option
1.	(b)
2.	(c)
3.	(d)
4.	(b)

Check Your Progress – 2

Question No.	Correct option
1.	(b)
2.	(b)
3.	(c)
4.	(b)

1.7 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016

1.8 Assignment

(1) What are the two main purposes of regression analysis?
(2) How can regression models help in experimental and observational studies?
(3) Describe Regression equation in detail.
(4) What are the error assumptions in Regression model? Why is it important that the errors in a simple linear regression model have a mean of zero?

Unit 2: Coefficients Calculation and Prediction

Unit Structure

2.0 Learning Objectives

After going through this unit, you should be able to

- Understand how to apply the least squares method to obtain regression coefficients
- Gain insight into interpreting and visualizing regression results
- Make predictions based on estimated regression equation
- Identify and understand the key assumptions underlying linear regression.
- To obtain regression coefficients using R through both manual calculations and built-in functions.
- Utilize R packages and functions effectively for Linear Regression computations

2.1 Introduction

To understand the application of the least squares method, let's take a practical example. In this walk-through, we will perform a hands-on calculation of the regression parameters step-by-step. This process will help to clearly illustrate how the least squares method is used to determine the regression coefficients.

2.2 Calculation of Regression Coefficients

Example 2.1: Imagine we have a dataset detailing the **Rental Price** (in ₹100 per month) for offices located in the heart of Ahmedabad. A crucial element we're focusing on is the **Size** of these offices, expressed in square feet. Suppose ten observations on size and their respective rental price have been gathered and are presented in Table 2.1.

Figure 2.2 presents a scatter plot illustrating the office rentals dataset, with Rental Price depicted on the vertical (y) axis and Size on the horizontal (x) axis. The plot clearly demonstrates a strong linear relationship between these two variables: as Size increases, Rental Price similarly rises. The tentative assumption of the straight-line

model $y = \beta_0 + \beta_1 x + \varepsilon$ appears to be reasonable. Capturing this relationship within a model would allow us to achieve two significant outcomes. First, it would enable us to comprehend how office size influences rental price. Second, we could predict rental prices for office sizes not represented in the historical data. For instance, we could estimate the rental price for a 830-square-foot office. These insights would be extremely valuable for real estate agents setting rental prices for new properties.

Table 2.1: Dataset for Example 2.1

Location	Size	Rental price
1	500	320
2	550	380
3	620	400
4	630	390
5	660	380
6	700	410
7	770	480
8	880	600
9	920	570
10	1000	620

Figure 2.2: Scatter plot of the Size and Rental Price, Example 2.1

Some of the calculations necessary to develop the least squares estimated regression equation are shown in Table 2.2. With the sample of 10 locations, we have $n = 10$ observations. Because equations (1.6) and (1.7) requires $\bar{x}$ and $\bar{y}$ we begin the calculations by computing $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{7230}{10} = 723$$
$$\bar{y} = \frac{\sum y_i}{n} = \frac{4550}{10} = 455$$

Table 2.2: Calculation for least squares estimated regression equation for Example 1

Location i	Size x_i	Rental price y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	500	320	-223	-135	30105	49729
2	550	380	-173	-75	12975	29929
3	620	400	-103	-55	5665	10609
4	630	390	-93	-65	6045	8649
5	660	380	-63	-75	4725	3969
6	700	410	-23	-45	1035	529
7	770	480	47	25	1175	2209
8	880	600	157	145	22765	24649
9	920	570	197	115	22655	38809
10	1000	620	277	165	45705	76729
Totals	**7230**	**4550**			**152850**	**245810**
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Using equations (1.6) and (1.7) and the information in Table 2.2, we can compute the slope and intercept of the estimated regression equation for the given dataset. The calculation of the slope (b_1) proceeds as follows.

$$b_1 = \frac{\sum(x - x_i)(y - y_i)}{\sum(x - x_i)^2} = \frac{152850}{245810} = 0.62$$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x} = 455 - 0.62 \times 723 = 5.42$$

Thus, the estimated regression equation is

$$\hat{y} = 5.42 + 0.62\,x$$

In other words,

$$\text{Rental Price} = 5.42 + 0.62 \times \text{Size}$$

Figure 1.3 shows the graph of this equation on the scatter diagram.

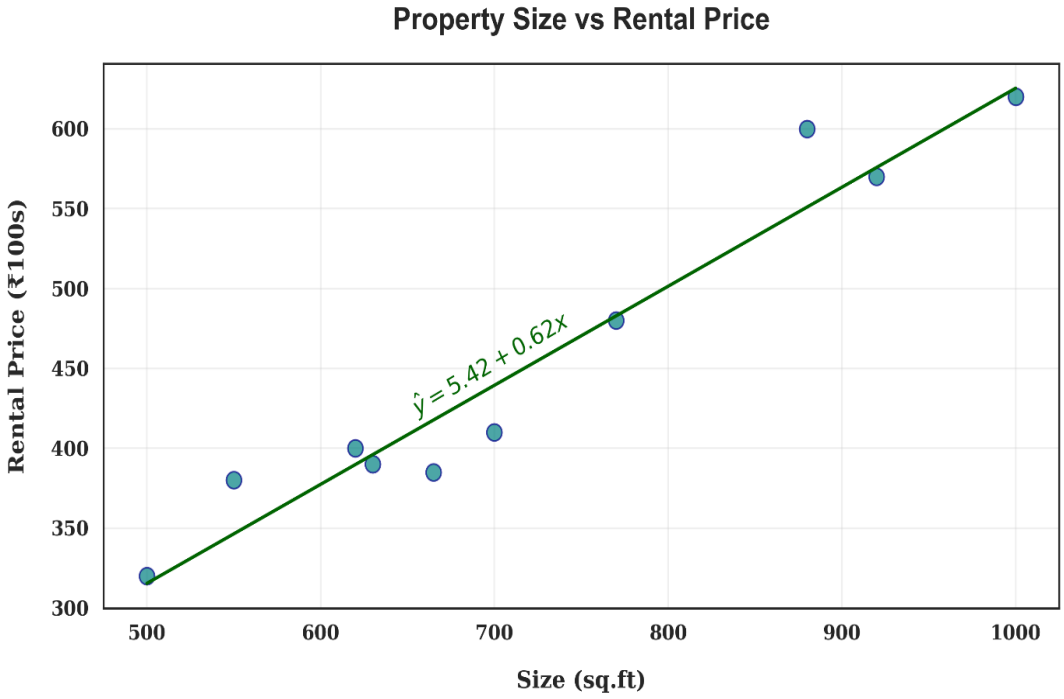


Figure 2.3: Graph of the estimated regression equation for Example 2.1

2.2.1 Interpretation for b_0 and b_1

The estimated y-intercept represents the expected response value when the predictor variable is zero. This interpretation is meaningful only if a predictor value of zero is reasonable for the situation being analyzed and if there is data with predictor values near zero. In this example, estimating the Rental Price when the floor Size is zero does not make sense. Moreover, we lack sample data close to zero. Therefore, it is not appropriate to interpret b_0 in practical terms. Henceforth, we will refer to b_0 as the estimated intercept, rather than the estimated y-intercept, since the latter is less relevant in regression analysis.

The slope estimates of $b_1 = 0.62$ has a clear practical interpretation. It indicates the slope of the linear relationship, meaning the expected change in the response variable for each 1-unit increase in the predictor variable. Specifically, we can state that the Rental Price is expected to increase by 0.62 for each 1-unit increase in Size. In other words, considering that Rental Price is measured in hundreds of rupees and Size is measured in square feet, we can expect the Rental Price to increase by ₹62 for each square foot increase in Size. It's crucial to mention the units of measurement for both the response and predictor variables when interpreting the slope in a simple linear regression model. This interpretation holds true only within the sample Size range, which is from 500 to 1000 square feet.

2.2.2 Predict the value of y when x is given

If we believe the least squares estimated regression equation adequately describes the relationship between x and y, it would seem reasonable to use the estimated regression equation to predict the value of y for a given value of x. For example, if we wanted to determine the expected rental price of the 830-square-foot office mentioned previously by simply plugging this value for SIZE into the model

$$\text{RENTAL PRICE} = 5.42 + 0.62 \times 830 = 521.53$$

It's important to be cautious when making predictions using regression for values of the independent variable (x) that fall outside the range of the data used to estimate the regression equation. We cannot be certain that the relationship holds true beyond the range of the data in the experiment.

Check Your Progress – 1

1. Assume you have noted the following prices for books and the number of pages that each book contains.

Book	A	B	C	D	E	F	G
Pages (x)	500	700	750	590	540	650	480
Price (y) (in \$)	70	75	90	65	75	70	45

   (a) Calculate $\bar{x}$ and $\bar{y}$.
   (b) Calculate $\Sigma(x_i - \bar{x})(y_i - \bar{y})$
   (c) Calculate $\Sigma(x_i - \bar{x})^2$
   (d) Develop the estimated regression equation by computing the values of b_0 and b_1.
   (e) Use the estimated regression equation to predict the value of y when x = 600.

2. Following are six observations collected in a regression study on two variables.

x	3	5	8	12	20	18
y	6	12	8	20	22	25

   (a) Calculate $\bar{x}$ and $\bar{y}$.
   (b) Calculate $\Sigma(x_i - \bar{x})(y_i - \bar{y})$
   (c) Calculate $\Sigma(x_i - \bar{x})^2$
   (d) Develop the estimated regression equation by computing the values of b_0 and b_1.
   (e) Use the estimated regression equation to predict the value of y when x = 15.

2.3 Properties of Fitted Regression Model

The least-squares fit possesses several notable properties:

1. In any regression model that includes an intercept term (β_0), the sum of the residuals is always zero.

$$\sum (y_i - \hat{y}_i) = 0$$

   This property directly follows from the first normal equation (Eq. 1.5). Note that rounding errors may affect this sum.
2. The sum of the observed values (y_i) is equal to the sum of the fitted values ($\hat{y}_i$). Symbolically,

$$\sum y_i = \sum \hat{y}_i$$

3. The least-squares regression line invariably passes through the centroid of the data, represented by the point ($\bar{x}$, $\bar{y}$).
4. The sum of the residuals weighted by the corresponding regressor variable value always equals zero, that is,

$$\sum x_i(y_i - \hat{y}_i) = 0$$

5. The sum of the residuals weighted by the corresponding fitted value always equals zero, that is,

$$\sum \hat{y}_i(y_i - \hat{y}_i) = 0$$

2.4 Streamlining Regression Analysis with R

Computing regression analysis manually is an incredibly time-consuming task. Luckily, using software such as R can significantly reduce this computational burden. This section will explain how employing R can streamline these analyses.

R is a versatile language specifically designed for statistical computing and graphics. It's highly valued by researchers and practitioners in Mathematics, Statistics, and Data Science. Essentially, R consists of numerous programs (or functions) that are organized into specialized packages (or libraries). These packages are developed by professional programmers and come with extensive help pages that elucidate each component of R. As an open-source and extensible software, R allows users to tailor it to their specific needs.

We assume that the learner has a basic understanding of R. If you are unfamiliar with R, we recommend visiting this tutorial which covers the basics and serves as a great starting point for beginners.

Let's now compute all the values shown in Table 2.2 using R. The following code snippet simplifies this computation:

```
# Data
size <- c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000)
rental_price <- c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)

# Calculate the means of X and Y
mean_x <- mean(size)
mean_y <- mean(rental_price)

# Calculate the slope (b1)
numerator <- sum((size - mean_x) * (rental_price - mean_y))
denominator <- sum((size - mean_x)^2)
b_1 <- numerator / denominator

# Calculate the intercept (b0)
b_0 <- mean_y - b_1 * mean_x

# Round the values to 2 decimal places
b_1 <- round(b_1, 2)
b_0 <- round(b_0, 2)

# Display the results
cat("Slope :", b_1, "\n")

## Slope : 0.62

cat("Intercept :", b_0, "\n")

## Intercept : 5.42

# The regression equation is:
cat("The regression equation is: Y =", b_0, "+", b_1, "* X\n")

## The regression equation is: Y = 5.42 + 0.62 * X
```

If you prefer to avoid manual calculations, you can use R's built-in functions to obtain the results directly. The following code snippet demonstrates this:

```
# Create the data frame
data <- data.frame(
  size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Fit the linear regression model
model <- lm(rental_price ~ size, data = data)

# Display the regression coefficients
coefficients(model)
```

```
## (Intercept)        Size
##   5.4228876   0.6218217
```

In the above code, we use the `lm()` function to fit a linear regression model, which predicts `rental_price` based on `size`. The `coefficients(model)` function then extracts and displays the regression coefficients.
As we have predicted the value of Rental Price for the Size = 830, we can use the following R code:

```
# Predict rental price for a single size
new_size <- data.frame(Size = 830)
prediction <- predict(model, newdata = new_size)

# Display the prediction
prediction
```

```
##        1
## 521.5349
```

R also allows us to predict more than one value for Size simultaneously. Here's how we can do it:

```
# Predict rental prices for new sizes
new_sizes <- data.frame(Size = c(750, 830, 980))
predictions <- predict(model, newdata = new_sizes)

# Display the predictions
predictions
```

```
##        1        2        3
## 471.7892 521.5349 614.8082
```

The following code will generate a scatter plot with a regression line:

```
# Create scatter plot
plot(data$size, data$rental_price,
     main = "Scatter Plot with Regression Line",
     xlab = "Size",
     ylab = "Rental Price",
     pch = 19, col = "red")  # Scatter plot with red points

# Add regression line
abline(model, col = "blue")  # Add blue regression line
```

This will generate a scatter plot with your data points and add a regression line to visualize the relationship between Size and Rental Price.

Check Your Progress – 2

Use R for the data provided in Problems 1 and 2 in **'Check Your Progress 1**' to achieve the following:

1. Perform step-by-step calculations to obtain regression coefficients and predictions.
2. Utilize the built-in R function `lm`().
3. Create scatter plots with regression line for both problems.

2.5 LET US SUM UP

In this unit, we provided an example to obtain regression coefficients using the least squares method. We explored how to apply this method and delved into interpreting and visualizing the results. You learned how to make predictions based on the estimated regression equation and gained insight into the key assumptions underlying linear regression. Additionally, we discussed the properties of the fitted regression model, enhancing your understanding of its behaviour and significance. We also demonstrated how to compute regression coefficients both manually and using R's built-in functions.

2.6 Check Your Progress: Possible Answers

Check Your Progress – 1

Answer: 1

(a) $\bar{x} = 601.43$ and $\bar{y} = 70.$

(b) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 6250$

(c) $\Sigma(x_i - \bar{x})^2 = 63085.72$

(d) $b_0 = 10.42$ and $b_1 = 0.099.$

(e) y = 69.86 when x = 600.

Answer: 2

(a) $\bar{x} = 11$ and $\bar{y} = 15.5.$

(b) $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 249$

(c) $\Sigma(x_i - \bar{x})^2 = 240$

(d) $b_0 = 4.0875$ and $b_1 = 1.0375.$

(e) y = 19.65 when x = 15.

2.7 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021

2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. R Introduction - W3Schools, https://www.w3schools.com/r/r_intro.asp.

2.8 Assignment

(1) What is the least squares method and why is it used in regression analysis?

(2) How would you use the estimated regression equation to make predictions for new data points? What are the potential risks or limitations of making predictions with a regression model?

(3) Oxygen consumption, also known as VO2 (Volume of Oxygen), is typically measured using a metabolic cart during physical activities like exercise. The following table presents a historical dataset collected by a space agency. The data illustrates the amount of oxygen an astronaut consumes during five minutes of intense physical activity. To simplify, we consider the astronaut's age as the variable that affects oxygen consumption.

ID	Age (Years)	Oxygen Consumption (Units)	ID	Age (Years)	Oxygen Consumption (Units)
1	37	44.39	7	46	28.17
2	42	47.34	8	37	31.22
3	41	37.99	9	43	44.72
4	43	30.83	10	38	54.85
5	44	37.85	11	43	39.84
6	48	27.07	12	46	36.42

(a) Use the least squares method to develop the estimated regression equation.

(b) Provide an interpretation of the slope of the estimated regression equation.

(c) Estimate the oxygen consumption for an astronaut who is 40 years old.

(d) Create scatter plot with regression line using R.

(e) Verify properties of fitted regression model stated in Section 2.3.

Unit 3 Evaluating Model Fit

Unit Structure

3.0 Learning Objectives

After going through this unit, you should be able to

- Understand the concept of *model fit* in the context of linear regression.
- Learn how to evaluate how closely the observed data (y-values) align with the fitted model's predictions ($\hat{y}$-values).
- Define and calculate *regression standard error* to assess the accuracy of the model's predictions.
- Understand the concept of $R\text{-}squared$ and its role in measuring how much of the variability in the response variable (y) is explained by the model.
- Learn to interpret the results of the regression standard error and $R\text{-}squared$ to assess model performance.

3.1 Introduction

Once a simple linear regression model has been fitted, it's important to assess how well it captures the relationship between the predictor variable (x) and the response variable (y). To determine the model's effectiveness, we need to evaluate the fit by answering two key questions: How well do the observed values of y align with the predictions made by the model? And how much of the variability in the data can the model explain? In this block, we'll

explore two crucial metrics that help answer these questions: the *regression standard error*, which tells us the average distance between observed and predicted values, and *R-squared*, which measures the proportion of variance in y explained by the model. These methods will provide insights into the reliability of the model for making predictions.

3.2 Regression standard error

Recall the least squares method used for estimating the regression parameters b_0 and b_1. The estimates b_0 and b_1 are the values that minimize the residual sum of squares (RSS),

$$RSS = \sum {e_i}^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

We can use this minimum value of RSS to determine how much (on average) the actual observed response values, y_i, deviate from the model-based fitted values, $\hat{y}_i$, by calculating the regression standard error, s:

$$s = \sqrt{\frac{RSS}{n-2}} \tag{3.1}$$

which is an estimate of the standard deviation of the random errors in the simple linear regression model. The residual sum of squares has $n - 2$ degrees of freedom, because two degrees of freedom are associated with the estimates b_0 and b_1 involved in obtaining $\hat{y}_i$. The quantity $RSS/(n-2)$is called the *mean square error*, which is often abbreviated MSE. The unit of measurement for s is the same as the unit of measurement for response variable y. The regression standard error is also known as *residual standard error* or *standard error of the estimate* or *root mean squared error*.

In Table 3.1 we show the calculations required to compute the residual sum of squares (RSS) for the RENTAL PRICE – SIZE example discussed in Unit 1. After computing and squaring the residuals for each location in the sample, we sum them to obtain RSS = 5804.54. Thus, RSS = 5804.54 measures the error in using the estimated regression equation $\hat{y} = 5.42 + 0.62\,x$ to predict rental price. The value of the residual standard error for the RENTAL PRICE – SIZE dataset is

$$s = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{5804.54}{10-2}} = \sqrt{725.57} = 26.94$$

This indicates average distance between actual rental price and estimated rental price is ₹2694. The R code snippets for obtaining the value of s is as follows:

```
# Get the summary of the model
model_summary <- summary(model)

# Extract the residual standard error (regression standard error)
regression_se <- model_summary$sigma
print(round(regression_se, 2))
```

```
   [1] 26.94
```

Table 3.1: Calculation of Squared Errors for Example 2.1 (Unit 2)

SIZE	*RENTAL PRICE*	*Fitted Values RENTAL PRICE = 5.42 + 0.62 × SIZE*	*Residuals (Errors)*	*Squared Errors*
x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
500	320	316.33	3.67	13.44
550	380	347.42	32.58	1061.14
620	400	390.95	9.05	81.86
630	390	397.17	-7.17	51.42
660	380	415.83	-35.83	1283.45
700	410	440.70	-30.70	942.37
770	480	484.23	-4.23	17.86
880	600	552.63	47.37	2244.29
920	570	577.50	-7.50	56.23
1000	620	627.24	-7.24	52.48
			RSS =	5804.54

A simple linear regression model is more effective when the observed y-values are closer to the fitted $\hat{y}$-values. Therefore, for a specific dataset, a smaller value of s is preferred over a larger one. The significance of "small" depends on the measurement scale of y, as both y and s share the same unit of measurement. Hence, s is most useful for comparing different models for the same response variable y. For instance, consider using FLOOR (number of floors) as an alternative predictor to SIZE. If we fit a simple linear regression model with RENTAL PRICE (in hundreds of rupees) and FLOOR, and find that the regression standard error is s = 30.52, it indicates that the observed RENTAL PRICE values deviate more (on average)

from the fitted RENTAL PRICE values in this model compared to the RENTAL PRICE–SIZE model (which had s = 26.94). This suggests that the random errors are larger, and consequently, the deterministic part of the RENTAL PRICE–FLOOR model is less accurately estimated on average. Therefore, we cannot determine the linear relationship between RENTAL PRICE and FLOOR as precisely as we can between RENTAL PRICE and SIZE.

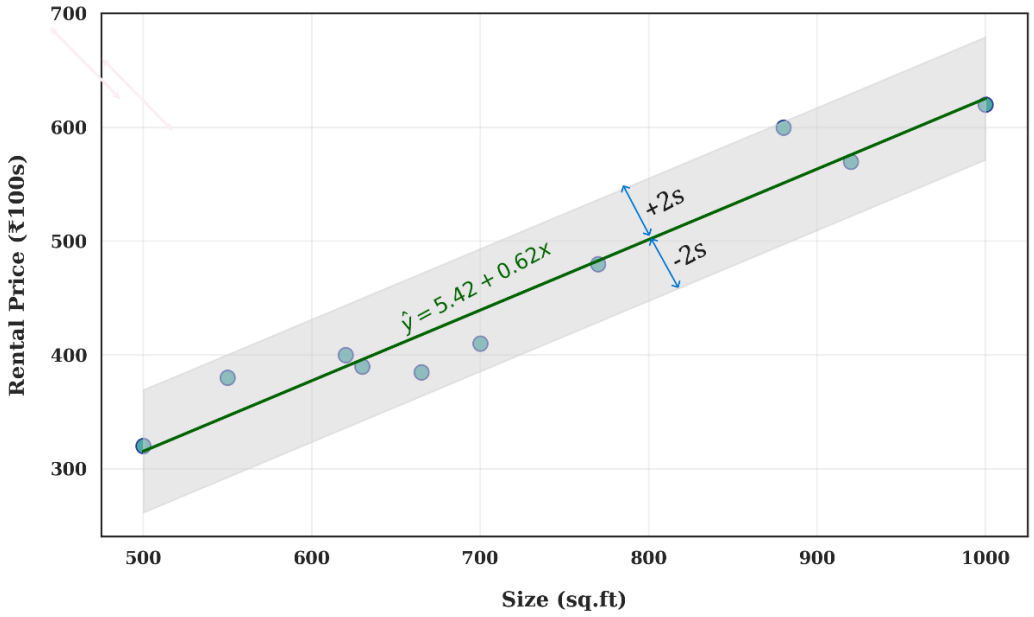


Figure 3.1: Regression line with a $\pm 2s$ band showing the 95% data range.

Using the Central Limit Theorem, another way to understand 's' is to multiply its value by 2, providing an approximate range for "prediction uncertainty." Specifically, about 95% of the observed y-values should fall within $\pm 2s$ of their predicted y-values. This means that using a simple linear regression model to predict y-values based on given x-values, we can expect an accuracy of about $\pm 2s$ at a 95% confidence level. In practice, approximately 95% of the data points in the scatterplot will lie within a vertical band of $\pm 2s$ from the regression line. It's reasonable to assume that unobserved data points will also typically fall within this range. Thus, when predicting an unknown y-value for a given x-value, it is likely to be within this band (See, Figure 3.1). This approximation can be improved by employing a more precise method for determining prediction intervals.

Check Your Progress – 1

1. Perform step-by-step calculations to compute *root mean squared error* data provided in Problems 1 and 2 in **'Check Your Progress 1'** of Unit 2.

3.3 Coefficient of Determination – R^2

To assess the fit of a simple linear regression model, we can compare it to a scenario where we have no knowledge of the predictor x. In this case, we only have a list of y-values. When predicting an individual y-value without a predictor, the sample mean $\bar{y}$ is the best estimate, as it is unbiased and has relatively low sampling variability. The difference $y_i - \bar{y}$ represents the error involved in using $\bar{y}$ to predict the y_i-value. We can gauge how well this univariate model fits the data by calculating the total sum of squares (TSS), which is the sum of the squared differences between the y_i-values and the sample mean $\bar{y}$, and defined as

$$\text{TSS} = \sum (\text{y}_\text{i} - \bar{\text{y}})^2$$

The sum at the bottom of the last column in Table 3.2 is the total sum of squares for RENTAL PRICE–SIZE dataset.

Table 3.2: Computation of total sum of squares for Example 2.1

SIZE	*RENTAL PRICE*	*Residuals (Errors)*	*Squared Errors*
x_i	y_i	$(\text{y}_\text{i} - \bar{\text{y}})$	$(\text{y}_\text{i} - \bar{\text{y}})^2$
500	*320*	*-135*	*18225*
550	*380*	*-75*	*5625*
620	*400*	*-55*	*3025*
630	*390*	*-65*	*4225*
660	*380*	*-75*	*5625*
700	*410*	*-45*	*2025*
770	*480*	*25*	*625*
880	*600*	*145*	*21025*
920	*570*	*115*	*13225*
1000	*620*	*165*	*27225*
		TSS =	100850

To measure how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$, another sum of squares is computed. This sum of squares, called *sum of squares due to regression*, denoted by, ESS (*Explained Sum of Squares*). This represents the variation in y-values (around their sample mean) that is "*explained*" by the simple linear regression model (see, Figure 3.2).

Figure 3.2 illustrates the relationship among these three sums of squares, presenting one of the most important results in statistics. The horizontal line represents the sample mean, $\bar{y}$, while the positively sloped line represents the estimated regression line, $\hat{y} = b_0 + b_1 x$.

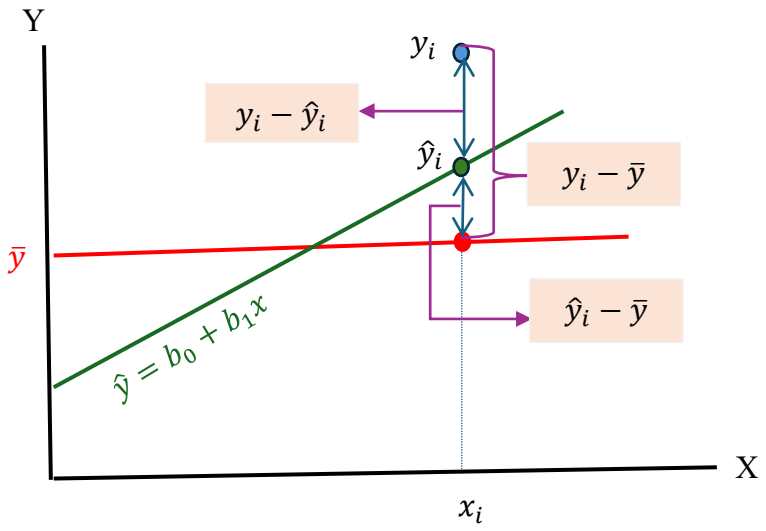


Figure 3.2: Deviations about the Estimated Regression Line and the Line $y = \bar{y}$.

This relation arises from the description of an observation as

$$\underbrace{y_j}_{\text{Observed}} = \underbrace{\hat{y}_j}_{\text{Fit}} + \underbrace{y_i - \hat{y}_i}_{\textit{Deviation}\text{ from fit}}$$

Subtracting $\bar{y}$ from both sides, we obtain

$$\underbrace{y_i - \bar{y}}_{\text{Deviation from mean}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{Deviation due to fit}} + \underbrace{y_i - \hat{y}_i}_{\text{Residual}}$$

Consequently, the total sum of squared deviations (TSS) in y can be divided into two components: the first, deviation due to fit, ESS, evaluates the effectiveness of x as a predictor of y, while the second, RSS, quantifies the prediction error.

To assess how much smaller the RSS is compared to the TSS, we calculate the proportional reduction from TSS to RSS. This is referred to as the *coefficient of determination*, or R^2 ("R-squared"):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

The *goodness-of-fit index*, R^2, represents the proportion of the total variability in the response variable y that can be attributed to the predictor variable x. A high R^2 value, close to 1, suggests that x accounts for a substantial portion of the variation in y. This index is known as the coefficient of determination because it indicates how much the predictor variable x determines or accounts for the response variable y.

For RENTAL PRICE–SIZE dataset, we already know RSS = 5804.54 and TSS = 100850. Therefore,

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{100850 - 5804.54}{100850} = 0.942$$

To interpret this number, it is standard practice to report the value as a percentage. In this case, we would conclude that 94.2% of the variation in RENTAL PRICE (about its mean) can be explained by a linear association between RENTAL PRICE and SIZE.

Since $0 \leq RSS \leq TSS$, the value of R^2 must fall between 0 and 1. Consider three possibilities:

1. If $RSS = TSS$, then $R^2 = 0$. This means that using x to predict y has not been effective, and we might as well predict y using the sample mean, $\bar{y}$, regardless of the value of x.
2. If $RSS = 0$, then $R^2 = 1$. In this case, using x allows us to predict y perfectly, with no random errors.
3. These extreme cases are rare in practice; typically, R^2 falls between 0 and 1, with higher R^2 values indicating better-fitting simple linear regression models.

3.3.1 Correlation Coefficient

We already know that the correlation coefficient serves as a descriptive measure of the strength of linear association between two variables, x and y. This coefficient can range from -1 to +1. A value of +1 signifies that the two variables are perfectly positively linearly related, meaning all data points lie on a straight line with a positive slope. Conversely, a value of -1 indicates a perfect negative linear relationship, with all data points on a straight line with a negative slope. If the correlation coefficient is close to zero, it suggests that x and y do not have a linear relationship.

If a regression analysis has been conducted and the coefficient of determination, R^2, has been calculated, we can use the algebraic relationship between the correlation coefficient, r, and the coefficient of determination, R^2 as follows:

$$\mathrm{r} = (\text{sign of } \mathrm{b}_1)\sqrt{\mathrm{R}^2}$$

where, b_1 is the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$.
For RENTAL PRICE–SIZE dataset, since R^2 = 0942 and b_1 has positive sign, the correction coefficient is $+\sqrt{0.942} = +0.97$. Hence, we would conclude that a strong positive linear association exists between x and y.

Both measures serve important purposes. The correlation coefficient indicates the strength and direction of any linear relationship between y and x, whereas the coefficient of

determination (R^2) is a more general concept. R^2 ranges from 0 to 1, while the correlation coefficient ranges from -1 to +1. While the correlation coefficient is confined to linear relationships between two variables, R^2 can be applied to both nonlinear relationships and those involving multiple independent variables. As a result, the coefficient of determination has a broader range of applicability.

The following code snippet gets the R-squared value from the `model_summary`, calculates the correlation coefficient by taking the square root of R-squared, and prints both values with a precision of two decimal places.

```
# Extract R-squared (coefficient of determination)
r_squared <- model_summary$r.squared

# Sample correlation coefficient (r) is the square root of R-squared
correlation_coefficient <- sqrt(r_squared)

# Print values with 2 digits precision
print(paste("R-squared:", round(r_squared, 2)))
```

```
   [1] "R-squared: 0.94"
```

```
print(paste("Correlation coefficient (r):", round(correlation_coefficient, 2)
))
```

```
   [1] "Correlation coefficient (r): 0.97"
```

Check Your Progress – 2

1. For the data provided in Problems 1 and 2 in **'Check Your Progress 1**' of Unit 2, Compute:
   (a) The coefficient of determination, R^2 and sample correlation coefficient.
   (b) Comment on the goodness of fit.

3.4 LET US SUM UP

The importance of model evaluation in linear regression lies in understanding how well the model captures the relationship between the predictor and response variables. Larger values of R-squared (R^2) indicate a better fit of the model to the data, meaning the observed values are more closely grouped around the regression line. Similarly, a smaller regression

standard error suggests that the model's predictions are more accurate, as the observed values are closer to the predicted values. Additionally, correlation coefficients are crucial for identifying linear associations between two variables, providing insight into the strength and direction of their relationship. Together, these metrics help assess the quality and reliability of a regression model's predictions.

3.5 Check Your Progress: Possible Answers

Check Your Progress – 1 & 2

We provide a solution for Problem Set 1 using R code. This code delivers a thorough overview of all key regression metrics and presents the results in an organized, user-friendly format, aligning well with manual calculations as discussed so far.

```
# Computing Model Metrics Including Residuals, Standard Error, and Display
ing All Relevant Metrics

# Step 1: Define the dataset
# We have two variables: Pages (x) and Price (y)
pages <- c(500, 700, 750, 590, 540, 650, 480)
price <- c(70, 75, 90, 65, 75, 70, 45)

# Step 2: Fit a linear model to the data
# We are fitting a linear regression model to predict Price based on Pages
model <- lm(price ~ pages)

# Step 3: Extract model summary
# This gives us R-squared, residual standard error, coefficients, and more
model_summary <- summary(model)

# Step 4: Prepare the table to show model details along with computed metr
ics
# Residual Sum of Squares (RSS)
RSS <- sum(residuals(model)^2)  # Sum of squared residuals

# Total Sum of Squares (TSS)
mean_price <- mean(price)  # Mean of the observed Price values
TSS <- sum((price - mean_price)^2)  # Sum of squared differences from the
mean
cat("Total Sum of Squares (TSS):", round(TSS, 2), "\n")

   Total Sum of Squares (TSS): 1100

# Step 5: Collect model metrics in a table
model_metrics <- data.frame(
  Metric = c("Residual Sum of Squares (RSS)", "Residual Standard Error (Si
```

```
gma)",
              "R-squared", "Intercept", "Slope"),
  Value = c(
    round(RSS, 2),  # RSS rounded to 2 decimal places
    round(model_summary$sigma, 2),  # Residual standard error (sigma)
    round(model_summary$r.squared, 2),  # R-squared from model summary
    round(model_summary$coefficients[1, 1], 2),  # Intercept
    round(model_summary$coefficients[2, 1], 2)   # Slope (coefficient of P
ages)
  )
)

# Step 6: Display the table with all metrics
cat("Model Summary and Key Metrics:\n")

   Model Summary and Key Metrics:

print(model_metrics)

                             Metric  Value
   1   Residual Sum of Squares (RSS) 480.80
   2 Residual Standard Error (Sigma)   9.81
   3                       R-squared   0.56
   4                       Intercept  10.42
   5                           Slope   0.10

# Step 7: Display the dataset along with predicted values (Fitted Values)
and Residuals
fitted_values <- round(fitted(model), 2)  # Fitted values rounded to 2 dec
imal places
residuals_values <- round(residuals(model), 2)  # Residuals rounded to 2 d
ecimal places

# Combine the dataset with fitted values and residuals into a table
results_table <- data.frame(
  Pages = pages,
  Actual_Price = price,
  Predicted_Price = fitted_values,
  Residuals = residuals_values
)

cat("\nDataset with Fitted Values and Residuals:\n")


   Dataset with Fitted Values and Residuals:

print(results_table)
```

```
  Pages Actual_Price Predicted_Price Residuals
1   500           70           59.95     10.05
2   700           75           79.77     -4.77
3   750           90           84.72      5.28
4   590           65           68.87     -3.87
5   540           75           63.91     11.09
6   650           70           74.81     -4.81
7   480           45           57.97    -12.97
```

3.6 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020

3.7 Assignment

(1) How does the standard error help in assessing the reliability of regression coefficients?
(2) Explain the role of the coefficient of determination (R^2) in evaluating a regression model.
(3) What does the correlation coefficient signify, and how does it differ from R^2?
(4) To further reinforce your understanding, you are encouraged to calculate all the metrics discussed in this unit for the dataset provided in Assignment 2.8 (Unit 2).

Unit 4 Assessing the Strength of the Linear Relationship

Unit Structure

4.0 Learning Objectives

After going through this unit, you should be able to

- Understand how the *slope* parameter (b_1) in linear regression represents the relationship between the predictor variable (x) and the response variable (y).
- Learn how to estimate and test the *slope parameter* for significance.
- Understand the concept of *statistical significance* and how to interpret p-values to determine if there is a meaningful linear relationship between x and y.
- Understand how to interpret the strength of evidence of a linear association and its implications for model reliability and prediction accuracy.

4.1 Introduction

After evaluating how well the model fits the data, it's essential to examine the relationship between the predictor variable (x) and the response variable (y). Is the linear relationship strong enough to make reliable predictions? To answer this question, we focus on the *slope* of the regression line, which indicates the nature of the relationship, and its *statistical significance*, which tells us whether this relationship is likely to be meaningful or due to chance. In this block, we will explore how to estimate and test the slope parameter, determine the strength of the linear association, and assess whether the model's predictor

can be trusted for making predictions. This step is crucial to ensure that the model not only fits well but is also backed by a solid and reliable relationship.

4.2 TESTS OF HYPOTHESES

A more formal approach to assess the usefulness of x as a predictor of y is through hypothesis testing of the regression parameter b_1. In a simple linear regression model, the expected value of y is a linear function of x, expressed as: $E(y|x) = \beta_0 + \beta_1 x$.

If $\beta_1 = 0$, the equation simplifies to $E(y|x) = \beta_0 + 0 \cdot x = \beta_0$, meaning the expected value of y does not depend on x. In this case, we would conclude that there is no linear relationship between x and y. On the other hand, if $\beta_1 \neq 0$, we would infer that x and y are related. Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.

To conduct this hypothesis test, we make the following assumptions: for each fixed value of x, the residuals (ε's) are assumed to be independent, normally distributed random variables with a mean of zero and a common variance σ^2. Under these assumptions, b_0 and b_1 are unbiased estimators of β_0 and β_1, respectively. Their variances are

$$Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x - \bar{x})^2}\right] \quad (1)$$

$$Var(b_1) = \frac{\sigma^2}{\sum(x - \bar{x})^2} \quad (4.2)$$

Furthermore, the *sampling distributions* of the least squares estimates b_0 and b_1 are normal with means β_0 and β_1 and variance as given in (4.1) and (4.2), respectively.

Replacing σ^2 in (4.1) and (4.2) by s^2, defined in Unit 3 eq. (3.1), we get unbiased estimates of the variances of b_0 and b_1. An *estimate* of the *standard deviation* is called the *standard error* (s.e.) of the estimate. Thus, the standard errors of b_0 and b_1 are

$$s_{b_0} = s\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x - \bar{x})^2}\right]} \quad (4.3)$$

$$s_{b_1} = \frac{s}{\sqrt{\sum(x - \bar{x})^2}} \quad (4.4)$$

RENTAL PRICE – SIZE dataset, $s = 26.94$. Hence, using $\sum(x - \bar{x})^2 = 245810$ (as shown in previous unit, Table --), we have

$$s_{b_1} = \frac{26.94}{\sqrt{245810}} = 0.054$$

as the estimated standard deviation of b_1.

4.2.1 t Test

With the sampling distributions of b_0 and b_1, we are now able to conduct a statistical analysis to assess the effectiveness of x as a predictor of y. Assuming normality, the t-Test is the appropriate test statistic for testing the null hypothesis $\mathbf{H}_0: \beta_1 = 0$ against the alternative hypothesis $\mathbf{H}_1: \beta_1 \neq 0$.

The *t* test for a significant relationship is based on the fact that the test statistic

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

follows a *t* distribution with $n - 2$ degrees of freedom. If the null hypothesis is true, then $\beta_1 = 0$ and $t = b_1/s_{b_1}$.

Let us conduct this test of significance for RENTAL PRICE – SIZE dataset at the $\alpha = 0.05$ level of significance. The test statistic is

$$t = \frac{b_1}{s_{b_1}} = \frac{0.62}{0.054} = 11.48$$

The appropriate critical value obtained from Student's t-distribution for $n - 2 = 10 - 2 = 8$ degree of freedom (df) and $\alpha = 0.05$ is 2.31.

Figure 4.1 shows a t-distribution with 8 degrees of freedom. With a significance level of $\alpha = 0.05$, the rejection regions are divided between both tails, with each tail representing $\frac{\alpha}{2} = 0.025$. The critical values are marked at -2.31 and 2.31, defining the boundaries of the acceptance region. Since our calculated t-value is 11.48, it falls well beyond the critical value of 2.31 ($t = 11.48 > 2.31$), placing it in the rejection region. Thus, the null hypothesis is rejected. This evidence is sufficient to conclude that a significant relationship exists between rental price and size.

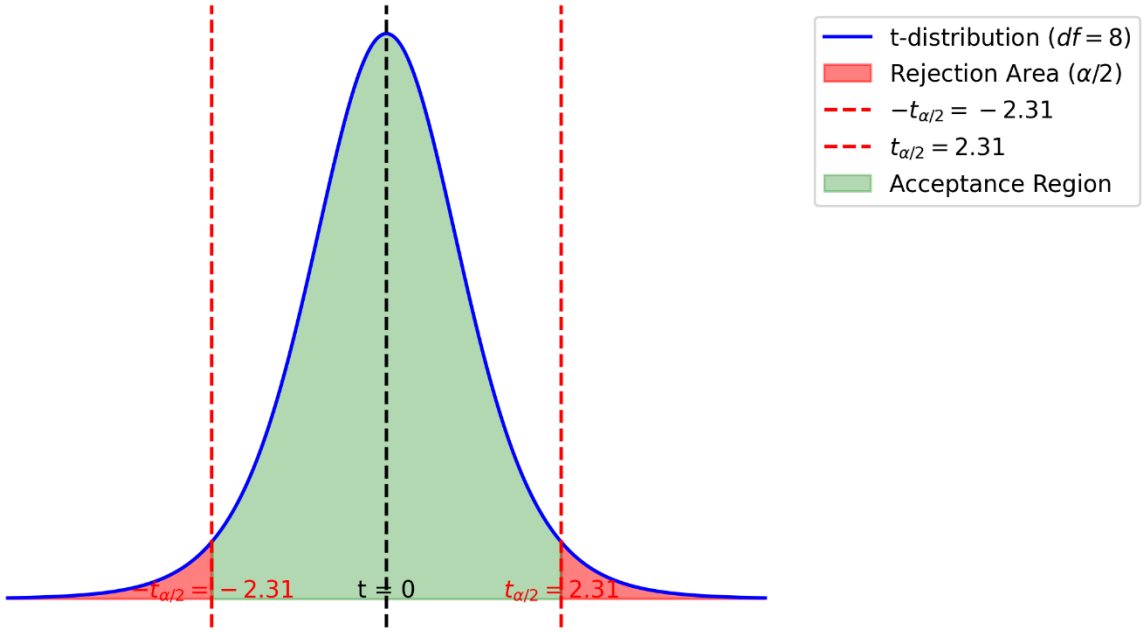


Figure 4.1: t-distribution (df = 8 and $\alpha = 0.05$) with critical values and rejection area

If we type **summary**(model) in the console, we will get

```
Call:
   lm(formula = Rental_Price ~ Size, data = data)

   Residuals:
       Min      1Q  Median      3Q     Max
   -35.825  -7.435  -5.698   7.702  47.374

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept)  5.42289   40.19353   0.135    0.896
   Size         0.62182    0.05433  11.445 3.07e-06 ***
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Residual standard error: 26.94 on 8 degrees of freedom
   Multiple R-squared:  0.9424, Adjusted R-squared:  0.9352
   F-statistic:   131 on 1 and 8 DF,  p-value: 3.072e-06
```

In the output above, after the function call, you'll find a summary of the residuals, followed by the Coefficients Table. This table presents the estimated regression coefficients (`5.42289` and `0.62182`) along with their respective standard errors (`40.19353` and `0.05433`). The first line below the Coefficients Table indicates that the Residual standard error is `26.94`, based on 8 degrees of freedom (n − 2).

***p*-value**: Another criterion to determine the significance is to compare p-values for the t-Test. The p-value is calculated and included in the output of the `lm()` function. In the **summary**(model) output above, the t-values and p-values for both the intercept and the slope are provided under the columns labeled "`t value`" and "`Pr(>|t|)`", respectively. It's important to note that the p-value for the slope is very close to zero (`3.07e-06`). Hence, the null hypothesis is rejected at practically any small significant level. This is indicated in the **summary**(model) output above by three stars. A p-value less than 0.05 (α), indicates that the predictor (Size in this case) is significantly associated with the response variable (Rental Price).

The entire procedure is illustrated in the following R code to perform a t-Test, obtain critical values, and extract t-values and p-values from the table, and interpret the results based on the significance level.

```r
# Creating a data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Rental_Price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Fit the linear regression model
model <- lm(Rental_Price ~ Size, data = data)

# Get the summary of the model
model_summary <- summary(model)

# Degrees of freedom
df <- nrow(data) - 2

# Critical value for two-tailed test at 5% significance level
alpha <- 0.05
critical_value <- qt(1 - alpha / 2, df)

# Print the critical value
cat("Critical t-value for 5% significance level: ", critical_value, "\n")
```

```
   Critical t-value for 5% significance level:  2.306004
```

```r
# Extract the coefficient (b1) and standard error for Size
coef_size <- coef(model)["Size"]
se_size <- model_summary$coefficients["Size", "Std. Error"]

# Calculate the t-statistic
t_stat <- coef_size / se_size

# Extract the p-value for Size
p_value <- model_summary$coefficients["Size", "Pr(>|t|)"]
```

```
# Print the results
cat("t-statistic for Size: ", t_stat, "\n")
```

```
   t-statistic for Size:  11.44528
```

```
cat("p-value for Size: ", p_value, "\n")
```

```
   p-value for Size:  3.072415e-06
```

```
# Determine if Size is significant at 5% significance level
if (abs(t_stat) > critical_value) {
  cat("Reject the null hypothesis: Size is significantly associated with Rent
al_Price.\n")
} else {
  cat("Fail to reject the null hypothesis: Size is not significantly associat
ed with Rental_Price.\n")
}
```

```
   Reject the null hypothesis: Size is significantly associated with Rental_P
rice.
```

Confidence Interval for β_1

The form of a confidence interval for β_1 is as follows:

$$b_1 \pm t_{\alpha/2} s_{b_1} \tag{4.5}$$

The confidence interval in (4.5) follows the standard interpretation: if we were to repeatedly sample the same size from the same values of x and construct 95% confidence intervals for the slope parameter for each sample, we would expect 95% of these intervals to contain the true value of the slope.

From earlier calculations, we see that a 95% confidence interval for β_1 is

$$b_1 \pm t_{\alpha/2} s_{b_1} = 0.62 \pm 2.31 \times 0.054 = (0.4965, 0.7471)$$

As the interval does not include zero, it suggests that Size has a statistically significant relationship with the Rental Price at the 95% confidence level. Thus, for each additional square foot Size, the Rental Price is expected to increase between ₹50 and ₹75.

The following R code snippet demonstrates how to obtain the lower and upper confidence limits.

```
# Compute the 95% confidence interval for the slope (Size)
conf_int <- confint(model, level = 0.95)

# Print the confidence interval for the slope
cat("95% Confidence Interval for the slope (Size): ", conf_int[2, ], "\n")
```

```
   95% Confidence Interval for the slope (Size):  0.4965366 0.7471069
```

Check Your Progress – 1

1. For the data provided in Problems 1 and 2 in '**Check Your Progress 1**' of Unit 2,
   (a) Test for a significant relationship by using the *t* test. Use α = .05.
   (b) What is your conclusion?

4.2.2 F Test

An F test, which relies on the F probability distribution, can also be used to test for significance in regression. When there is only one independent variable, the F test will yield the same result as the t-test. In other words, if the t-test shows that $\beta_1 \neq 0$ and indicates a significant relationship, the F test will also show a significant relationship. However, when there are multiple independent variables, only the F test can be used to test for an overall significant relationship.

The rationale for using the F test to assess the statistical significance of a regression relationship is based on generating two independent estimates of σ^2. We previously discussed how the Mean Squared Error (MSE) provides one estimate of σ^2. If the null hypothesis H_0: $\beta_1 = 0$ is true, dividing the sum of squares due to regression (ESS) by its degrees of freedom gives another independent estimate of σ^2. This estimate is referred to as the Mean Square due to Regression (MSR). If we consider the regression degrees of freedom equals to the number of independent variables in the model, MSR is given by

$$MSR = \frac{ESS}{\text{Number of independent variables}}$$

For RENTAL PRICE – SIZE dataset, there is one independent variable, $MSR = ESS = TSS - RSS = 100850 - 5804.54 = 95045.46$.

If the null hypothesis $H_0: \beta_1 = 0$ is true, both MSR and MSE are independent estimates of σ^2, and the ratio $\frac{MSR}{MSE}$ follows an F distribution, with the numerator degrees of freedom equal to one and the denominator degrees of freedom equal to $n - 2$. Therefore, when $\beta_1 = 0$, the ratio $\frac{MSR}{MSE}$ should be close to one. However, if the null hypothesis is false $(\beta_1 \neq 0)$, MSR will overestimate σ^2, causing the ratio $\frac{MSR}{MSE}$ to become inflated. Consequently, large values of $\frac{MSR}{MSE}$ lead to the rejection of H_0, indicating that the relationship between x and y is statistically significant.

Hence, the test statistic for RENTAL PRICE – SIZE dataset is

$$F = \frac{MSR}{MSE} = \frac{95045.46}{725.57} = 130.99$$

The *F* distribution table shows that with one degree of freedom in the numerator and 8 degrees of freedom in the denominator, F = 5.32 provides an area of 0.05 in the upper tail. Since our calculated F value is greater than critical value (130.99 > 5.32), the result is highly significant, leading to the rejection of the null hypothesis.

An ANOVA table can be used to summarize the results of the F test for significance in regression. The following code snippet produces the ANOVA table with the F test computations performed for RENTAL PRICE – SIZE dataset.

```
# Generate the ANOVA table
anova(model)
```

```
Analysis of Variance Table

Response: Rental_Price
          Df Sum Sq Mean Sq F value    Pr(>F)
Size       1  95045   95045  130.99 3.072e-06 ***
Residuals  8   5805     726
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the above output, `Sum Sq` indicates Sum of Squares whereas `Mean Sq` indicates Mean Squares. `Pr(>F)` shows the p-value associated with the F test. The ANOVA table confirms the same conclusion about the significance of the relationship as the t-test.

Check Your Progress – 2

1. For the data provided in Problems 1 and 2 in '**Check Your Progress 1**' of Unit 2,
   (a) Use the *F* test to test for a significant relationship. Use α = .05.
   (b) What is your conclusion?

4.3 LET US SUM UP

This unit emphasized assessing the linear relationship strength between the predictor (x) and the response (y). By examining the slope parameter (b_1), we understood its role in model predictions. Performing t-Tests and F-Tests provided insights into testing the slope's statistical significance, determining if the relationship is meaningful or by chance. These tests help evaluate the model's reliability and ensure the relationship between variables supports accurate predictions. Ultimately, this unit highlighted the importance of a statistically significant and trustworthy relationship for making informed decisions.

4.4 Check Your Progress: Possible Answers

Check Your Progress – 1 & 2

We provide a solution for Problem Set 1 using R code.

```
# Creating the Data Frame
data <- data.frame(
  pages = c(500, 700, 750, 590, 540, 650, 480),
  price = c(70, 75, 90, 65, 75, 70, 45)
)

# Performing Linear Regression
model <- lm(price ~ pages, data = data)

# Summary of the model
model_summary <- summary(model)

# t-Test p-value for the slope parameter (b₁)
t_test_p_value <- model_summary$coefficients[2, 4]

# Performing ANOVA for the model
anova_result <- anova(model)

# F-Test p-value
f_test_p_value <- anova_result$"Pr(>F)"[1]

# Set significance level
alpha <- 0.05

# Check significance of t-Test and F-Test
if (t_test_p_value < alpha & f_test_p_value < alpha) {
  cat("The relationship between pages and price is statistically significa
nt.\n")
  cat("t-Test p-value:", t_test_p_value, "\n")
  cat("F-Test p-value:", f_test_p_value, "\n")
} else {
  cat("The relationship between pages and price is not statistically signi
ficant.\n")
  cat("t-Test p-value:", t_test_p_value, "\n")
  cat("F-Test p-value:", f_test_p_value, "\n")
}
```

```
   The relationship between pages and price is not statistically significa
nt.
   t-Test p-value: 0.05204836
   F-Test p-value: 0.05204836
```

4.5 Further Reading

1. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
2. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
3. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

4.6 Assignment

1. Discuss how the F-test is used to assess the overall fit of a simple regression model.
2. Explain how the t-test is used to evaluate the significance of the slope (regression coefficient) in a simple regression model.
3. For the dataset provided in Assignment 2.8 (Unit 2), answer the following questions:

(a) Does the *t* test indicate a significant relationship between Oxygen consumption and the age? what is your conclusion? Use α = .05.

(b) Test for a significant relationship using the *F* test. What is your conclusion? Use α = .05.

(c) Show the ANOVA table for this data.

Block 2: Multiple Regression and Model Diagnostics

Introduction

Regression analysis is a fundamental tool in data science, widely used to uncover relationships between variables and make data-driven predictions. While simple linear regression provides a foundation, real-world problems often require more sophisticated techniques to account for multiple predictors and ensure model accuracy. This block, *Multiple Regression and Model Diagnostics*, delves into the intricacies of multiple linear regression, hypothesis testing, and model diagnostics, equipping you with the tools to build, evaluate, and refine robust regression models.

In *Unit 1: Multiple Linear Regression*, you will explore the principles and formulation of multivariable linear regression. You will learn how to develop models using multiple numerical predictors, interpret the coefficients, and evaluate model fit using metrics like R-squared and Adjusted R-squared. This unit lays the groundwork for understanding how multiple variables interact to influence an outcome.

Unit 2: Testing for Significance focuses on hypothesis testing in the context of multiple regression. You will learn about the F-test, which assesses the overall significance of the model, and the t-test, which evaluates the significance of individual predictors. Through practical examples, such as the Office Rental Price Example, you will gain hands-on experience in applying these tests to real-world data.

Finally, *Unit 3: Model Diagnostic and Residual Analysis* emphasizes the importance of validating regression models. You will learn about residuals, the key assumptions of linear regression, and how to use residual plots and statistical measures like leverage and Cook's Distance to diagnose model issues. By the end of this unit, you will be proficient in using statistical software (e.g., R) to perform comprehensive regression diagnostics, ensuring your models are both accurate and reliable.

Together, these units provide a comprehensive understanding of advanced regression techniques, empowering you to tackle complex data analysis challenges with confidence. Regardless of your background or experience, this block will enhance your ability to build, interpret, and validate sophisticated regression models in real-world scenarios.

Unit 1 Multiple Linear Regression

Unit Structure

1.0 Learning Objectives

After going through this unit, you should be able to

- Understand the principles and formulation of multivariable linear regression.
- Develop a multivariable linear regression model using multiple numerical predictors.
- Interpret the coefficients of a multivariable model.
- Evaluate the fit of a regression model using metrics such as R-squared, and Adjusted R-squared.

1.1 Introduction

The simple two-variable regression models we have examined so far represent the most basic form of regression analysis. However, real-world scenarios are usually more intricate. Therefore, in this unit, we extend the model to introduce multiple linear regression models. The principles of regression models and regression equations discussed in the previous block also apply to multiple regression.

1.2 Multiple Linear Regression Model

The multiple linear regression model describes an algebraic relationship between a response variable and one or more predictor variables.

- Y represents the response variable, which may also be referred to as the dependent, outcome, or output variable. This variable should be quantitative, meaning it should have meaningful numerical values.
- (X_1, X_2, ...) are the predictor variables, also called independent, input, or explanatory variables, or covariates. In this unit, we assume that these variables are also quantitative.

Consider a sample of n sets of observations of $(X_1, X_2, \dots, Y)$, represented as $(x_{1\mathrm{i}}, x_{2\mathrm{i}}, \dots, y_{\mathrm{i}})$ for $i = 1,2, \dots, n$, where i indexes each observation in the sample. The simple linear regression model discussed in Block 1 is a special case where there is only one predictor variable. If there are p predictor or explanatory variables, the observations are usually represented as in Table 1.1

Table 1.1: Notation Used in Multiple Regression Analysis

Observation Number	Response (Y)	Predictors			
		X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

In most cases, the response variable Y is easily identifiable—it typically "responds" in some way to changes in the values of the predictor variables $(X_1, X_2, \dots)$. If the model accurately represents the relationship between Y and the predictors, knowing the values of the predictors allows us to predict corresponding values of the response variable.

The multiple linear regression model can be written as:

$$Y|(X_1, X_2, \dots) = E\big(Y|(X_{1\mathrm{i}}, X_{2\mathrm{i}}, \dots)\big) + e_{\mathrm{i}} \quad \text{for} \quad i = 1,2, \dots, n$$

Where,

- The vertical bar (|) indicates "given."
- $E\big(Y|(X_{1\mathrm{i}}, X_{2\mathrm{i}}, \dots)\big)$ is the expected value of **Y** given the values of the predictor variables.
- e_{i} is random error

The deterministic part of the model, $E\big(Y|(X_1, X_2, \dots)\big)$, is expressed as:

$$E\big(Y|(X_1, X_2, \dots)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

The regression equation, which represents this relationship, typically includes only the deterministic part. The regression parameter β_0 is the intercept, which is the value of Y when all predictor variables are zero. The regression parameter β_1 represents the change in Y for a 1-unit increase in X_1, while holding all other predictors constant. Similarly, β_2 shows the change in Y for a 1-unit increase in X_2, and so on.
The equation:

$$E\big(Y|(X_1, X_2, \dots)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

is known as the *regression equation* and represents the linear plane of best fit for the data.

1.3 Estimated Multiple Regression Equation

In practice, the parameters $\beta_0, \beta_1, \beta_2, \dots$ are estimated using sample data. From a random sample, we calculate sample statistics $b_0, b_1, b_2, \dots$, which serve as estimates for the corresponding population parameters $\beta_0, \beta_1, \beta_2, \dots$.
The estimated multiple regression equation, based on the sample data, is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots$$

Where: - $\hat{Y}$ is the predicted (or fitted) value of Y. - $X_1, X_2, \dots$ are the predictor variables. - The estimated parameters $b_0, b_1, b_2, \dots$ represent the intercept and the coefficients for the predictors.

In summary, the estimated regression equation provides an algebraic expression of the linear relationship between the response variable and the predictors, enabling us to make predictions about **Y** given specific values of the predictors.

1.4 Least Squares Criterion

In Block 1, we applied the least squares method to derive the estimated regression equation that most closely approximates the linear relationship between the dependent and independent variables. This same method is utilized to formulate the estimated multiple regression equation. The least squares criterion is restated as follows:

$$\min \sum (y_i - \hat{y}_i)^2 \qquad (1.1)$$

According to expression (1.1), the least squares method utilizes sample data to determine the values of $b_0, b_1, b_2, \dots$ that minimize the sum of squared residuals, which are the deviations between the observed values (y_i) and the predicted values ($\hat{y}_i$) of the dependent variables.

In Block 1, we presented formulas for calculating the least squares estimators b0 and b1 for the simple linear regression equation $\hat{y} = b_0 + b_1 x$. For smaller data sets, these formulas allowed us to manually compute b_0 and b_1. However, in multiple regression, deriving the formulas for the regression coefficients $b_0, b_1, b_2, \dots$ requires matrix algebra, which is beyond the scope of this text. Therefore, we will focus on using statistical software

like R to obtain the estimated regression equation and other relevant information in multiple regression. The main emphasis will be on interpreting the output from the R code, rather than on performing the multiple regression calculations manually.

1.5 Example – Estimating Rental Prices

Recall the *Rental Price–Size* example from Block 1, where we examined the relationship between the rental price and the size of office spaces. we continue with the same example and expand the analysis by investigating whether adding a new predictor—the *Floor* of the building where the office is located— can further explain the variation in rental price. The dataset with the *Floor* of the building added, are shown in Table 1.2.

Table 1.2: Dataset for Example 1.1

Location	Size	Floor	Rental price
1	500	4	320
2	550	7	380
3	620	9	400
4	630	5	390
5	660	8	380
6	700	4	410
7	770	10	480
8	880	12	600
9	920	14	570
10	1000	9	620

To model the rental price based on the size and floor of office locations in the population represented by this sample, we use a multiple linear regression model:

$$\mathrm{E}\,[\text{Rental Price}|(\text{Size}, \text{Floor})] = \mathrm{b}_0 + \mathrm{b}_1 \times \text{Size} + \mathrm{b}_2 \times \text{Floor}$$

These sample values are stored in an R object named 'data'. The computations for the estimated regression coefficients are illustrated using R code, and their values are displayed.

```
# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Floor = c(4, 7, 9, 5, 8, 4, 10, 12, 14, 9),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)
```

```r
# Fit the regression model
model <- lm(Rental_price ~ Size + Floor, data = data)

# Display only the coefficients
coefficients(model)
```

```
## (Intercept)        Size       Floor
##  15.5862798   0.5537922   4.7587783
```

Hence, the fitted regression equation is as follows:

$$\widehat{\text{Rental Price}} = b_0 + b_1 \times \text{Size} + b_2 \times \text{Floor} \; = \; 15.59 \; + 0.55 \times \text{Size} + 4.76 \times \text{Floor}$$

Note that this association holds only over the range of sample predictor values, that is, Size from 500 to 1000 square feet and Floor from 4 to 14.

The values b_1 = 0.55 and b_2 = 4.76 can be used together to determine how changes in both size and floor level affect the rental price. For instance, since the rental price is expressed in hundreds of rupees, an office with a size of 830 square feet located on the 10th floor would have a rental price of ₹100 × (0.55 × 830 + 4.76 × 10) = ₹50,410.

1.5.1 Interpretation of Coefficients

The interpretation of regression coefficients in a multiple regression equation often causes confusion. In simple regression, the equation represents a line, while in multiple regression, the equation represents a plane (when there are two predictors) or a hyperplane (when there are more than two predictors). As a result, the value of b_1 differs between simple and multiple regression. For instance, in our example, b_1 = 0.62 in simple regression, but b_1 = 0.55 in multiple regression. This means that for every 1 square foot increase in office space, the rental price is expected to increase by ₹0.55 when the number of floors remains constant. More specifically, we would expect the rental price to rise by ₹55 for each square foot increase in Size, assuming the number of floors is held constant. Similarly, b_2 = 4.76 indicates the change in rental price for a 1-unit increase in the number of floors, while keeping all other predictor variables constant. In other words, for each additional Floor, the rental price is expected to increase by ₹476, assuming the other variables remain unchanged.

From this, we can conclude that simple and multiple regression coefficients are not equal unless the predictor variables are uncorrelated. In observational data, predictor variables are rarely uncorrelated. However, in experimental settings, researchers often design the experiment to ensure that predictor variables are uncorrelated because they control the values of the predictors. Therefore, in experimental samples, it's possible for the

explanatory variables to be uncorrelated, which would make the simple and multiple regression coefficients identical in that particular sample.

1.5.2 MULTIPLE COEFFICIENT OF DETERMINATION

Once the linear model has been fitted to the data set, it's important to evaluate the adequacy of the fit. The discussion in Section 3.3 of Block 1 is relevant here. All the material extends naturally to multiple regression and thus will not be repeated.

The multiple coefficient of determination, denoted R^2, is computed as follows:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squares and RSS is the residual sum of squares. R^2 is the proportion of variation in Y(about mean) "explained" by a linear association between Y and $(X_1, X_2, \dots,)$.
We can obtain this value from the following command.

```
# Extract Multiple R-squared (Multiple Coefficient of Determination)
round(summary(model)$r.squared,4)

## " Multiple R-squared: 0.9535"
```

So, with the Multiple R-squared value being 0.9535, it indicates that approximately 95.35% of the variance in the rental price is predictable from the Size and Floor variables.

In Section 3.3 of Block 1, we see that the R^2 (R-squared) value for the regression equation with a single predictor variable, Size, is 94.2%. This means that 94.2% of the variation in rental price is explained by the regression equation. When an additional predictor variable, number of floors, is included, the R^2 value increases to 95.35%. Generally, R^2 tends to increase as more independent variables are added to the model.

This increase occurs because adding predictor variables typically reduces the prediction errors, which in turn lowers the residual sum of squares (RSS). Since ESS = TSS – RSS, a smaller RSS leads to a larger ESS, thus increasing the R^2 (R-squared) value. Note that $R^2 = ESS / TSS$, where ESS is the explained sum of squares and TSS is the total sum of squares.

When a variable is added to a model, R^2 tends to increase, even if the added variable is not statistically significant. To account for the number of independent variables in the model, the adjusted R^2 is used. This adjustment helps prevent overestimating the effect of adding a variable on the explained variability in the regression equation. Analysts often prefer using the adjusted R^2 to get a more accurate representation of the model's

explanatory power. The formula for the adjusted R^2 is calculated using n (the number of observations) and p (the number of independent variables) as follows:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad (1.2)$$

With n = 10 and p = 2, for our example, we have,

$$R_a^2 = 1 - (1 - 0.9535)\frac{10-1}{10-2-1} = 0.9403$$

In practice, we can obtain the value for adjusted R^2 directly from the following command for any particular multiple linear regression model.

```
# Extract Multiple R-squared (Multiple Coefficient of Determination)
round(summary(model)$adj.r.squared,4)

## "Adjusted R-squared:  0.9403 "
```

If the value of R^2 is low and the model includes many independent variables, the adjusted coefficient of determination may become negative. In such cases, the adjusted R_a^2 is effectively treated as zero.

Check Your Progress – 1

Note to students: The exercises in this and subsequent units are designed to be solved using computer software. It is highly recommended to use appropriate software packages to perform data analysis and complete the exercises effectively.

1. Consider the following data for a dependent variable Y and two independent variables, X1 and X2.

X1	25	30	47	51	40	36	74	51	76	59
X2	17	12	10	16	5	12	7	19	16	13
Y	112	94	108	178	94	117	170	175	211	142

(a) Develop an estimated regression equation relating Y to X1. Predict Y if X1 = 50.

(b) Develop an estimated regression equation relating Y to X2. Predict Y if X2 = 11.

(c) Develop an estimated regression equation relating Y to X1 and X2. Predict Y if X1 = 50 and X2 = 11.

(d) Compute R^2 for the model estimated in part (a), (b), and (c). Comments on the goodness of fit.

(e) Compute R_a^2.

(f) Do you prefer the multiple regression results? Explain.

1.6 LET US SUM UP

This unit provides a comprehensive introduction to multivariable linear regression. By the end, you will have a proper understanding of the key principles and formulation behind this technique. You'll be able to develop a regression model using multiple numerical predictors, interpret the model's coefficients, and assess its fit with important metrics like R-squared and Adjusted R-squared. This foundational knowledge will empower you to apply multivariable linear regression effectively in real-world data analysis.

1.7 Check Your Progress: Possible Answers

Check Your Progress – 1

Solution for Problem Set 1 using R code.

```
# Given Data
X1 <- c(25, 30, 47, 51, 40, 36, 74, 51, 76, 59)
X2 <- c(17, 12, 10, 16, 5, 12, 7, 19, 16, 13)
Y <- c(112, 94, 108, 178, 94, 117, 170, 175, 211, 142)
data <- data.frame(X1, X2, Y)

# Regression model relating Y to X1
model1 <- lm(Y ~ X1, data = data)
# summary(model1)
# Display only the coefficients
print(coefficients(model1))

## (Intercept)          X1
##   45.059369    1.943571

# Predicting Y when X1 = 50
pred_Y_X1_50 <- predict(model1, data.frame(X1 = 50))
pred_Y_X1_50

##        1
## 142.2379

# Regression model relating Y to X2
model2 <- lm(Y ~ X2, data = data)
# summary(model2)
# Display only the coefficients
print(coefficients(model2))

## (Intercept)          X2
##   85.217102    4.321488
```

```r
# Predicting Y when X2 = 11
pred_Y_X2_11 <- predict(model2, data.frame(X2 = 11))
pred_Y_X2_11
```

```
##        1
## 132.7535
```

```r
# Regression model relating Y to both X1 and X2
model3 <- lm(Y ~ X1 + X2, data = data)
# summary(model3)
# Display only the coefficients
print(coefficients(model3))
```

```
## (Intercept)          X1          X2
##  -18.368268    2.010185    4.737812
```

```r
# Predicting Y when X1 = 50 and X2 = 11
pred_Y_X1_50_X2_11 <- predict(model3, data.frame(X1 = 50, X2 = 11))
pred_Y_X1_50_X2_11
```

```
##        1
## 134.2569
```

```r
# Calculate R-squared and Adjusted R-squared for each model
r_squared <- c(summary(model1)$r.squared,
               summary(model2)$r.squared,
               summary(model3)$r.squared)

adj_r_squared <- c(summary(model1)$adj.r.squared,
                   summary(model2)$adj.r.squared,
                   summary(model3)$adj.r.squared)

# Create a data frame to display the results
results <- data.frame(
  Model = c("Model 1: Y ~ X1", "Model 2: Y ~ X2", "Model 3: Y ~ X1 + X2"),
  R_squared = r_squared,
  Adjusted_R_squared = adj_r_squared
)

# Display the results
print(results)
```

```
##                   Model R_squared Adjusted_R_squared
## 1      Model 1: Y ~ X1 0.6600351          0.6175395
## 2      Model 2: Y ~ X2 0.2215265          0.1242173
## 3 Model 3: Y ~ X1 + X2 0.9255251          0.9042466
```

Model 3 explains much more of the variation in the dependent variable (Y) compared to the simpler models with just X1 or X2. The higher Adjusted R-squared value in Model 3 suggests that the inclusion of both predictors (X1 and X2) leads to a better model.

1.8 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
4. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

1.9 Assignment

1. Explain the difference between simple and multiple regression.
2. How do you interpret the coefficients in a multiple regression model?
3. Let's consider adding another independent variable to our office rental price dataset: the broadband rate available at the office (in Mbps). The full dataset is presented in the table below.

Table 1.3: Office Rental Price Dataset Including Broadband Rate

Location	Size	Floor	Broadband Rate	Rental price
1	500	4	80	320
2	550	7	500	380
3	620	9	70	400
4	630	5	240	390
5	660	8	1000	380
6	700	4	80	410
7	770	10	70	480
8	880	12	500	600
9	920	14	80	570
10	1000	9	240	620

(a) Obtain estimated multiple regression equation.
(b) Compute and interpret R^2 and R_a^2.
(c) Do you prefer the multiple regression results? Explain.

Unit 2 Testing for Significance

Unit Structure

2.0 Learning Objectives

After completing this unit, you will be able to

(a) Understand the purpose and application of the F-test in multiple regression for assessing overall significance.
(b) Learn how to use the t-test to evaluate the significance of individual independent variables in a multiple regression model.
(c) Distinguish between the roles of the F-test and t-test in multiple regression analysis.
(d) Gain practical experience by applying both the F-test and t-test to a real-world example, such as the Office Rental Price Example.

2.1 Introduction

Once we have estimated the parameters in a multiple regression model, two important questions arise:

1. What is the overall adequacy of the model?
2. Which specific predictors are important in explaining the variation in the dependent variable?

To answer these questions, we use several hypothesis testing procedures. These formal tests rely on the assumption that random errors are independent and follow a normal distribution with a mean of zero and constant variance.

In this unit, we will explore how to use significance tests in multiple regression to assess both the overall fit of the model and the significance of individual predictors. We will cover the use of the F-test for overall significance and the t-test for assessing the importance of each predictor in the model.

2.2 F-Test

The F-test helps to determine whether there is a significant relationship between the dependent variable and the entire set of predictor variables. We will refer to this as the *test for overall significance*.

Suppose that our population multiple linear regression model has *p* predictor *X*-variables:

$$E\left(Y|\left(X_1, X_2, \dots X_p\right)\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p$$

The hypotheses for the *F* test involve the parameters of the multiple regression model.

1. *Null hypothesis:* $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
2. *Alternative hypothesis:* H_1:at least one of $\beta_1, \beta_2, \cdots \beta_p$ is not equal to zero
3. *Calculate test statistic:*

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

The first formula gives us some understanding of how the hypothesis test works. As outlined in Section 4.2.2 of Block 1, the difference between the Total Sum of Squares (TSS) and the Residual Sum of Squares (RSS), or TSS − RSS, is known as the regression sum of squares. When this value is small compared to RSS, it implies that the predictors $\left(X_1, X_2, \dots X_p\right)$ do little to reduce the random errors between the actual Y-values and the predicted $\hat{Y}$-values. In such cases, using the sample mean, $\bar{y}$, as the model could be equally effective. The F-statistic will be small, likely falling outside the rejection region, making the null hypothesis more likely to be true. Conversely, if the regression sum of squares (ESS) is large relative to the RSS, the predictors significantly reduce the random errors between the actual Y-values and the predicted $\hat{Y}$-values. This suggests that at least one of the predictors should be included in the model. Under these circumstances, the F-statistic will be large, likely falling in the rejection region, which supports the alternative hypothesis.

The second formula enables the calculation of the global F-statistic by using the value of R^2. It demonstrates how a high R^2 value tends to result in a large F-statistic, indicating a strong relationship between the predictors and the response variable, and vice versa.

4. *Set significance level* α: 5%
5. *Look up a critical value or a p-value using an F-distribution:*

- *Critical value*: A particular percentile of the F-distribution with p numerator degrees of freedom and $n - p - 1$ denominator degrees of freedom;
- *p-value*: The area to the right of the global F-statistic for the F-distribution with p numerator degrees of freedom and $n - p - 1$ denominator degrees of freedom;

6. *Make decision:*
   - If the F-statistic falls in the rejection region, or the p-value is less than the significance level, then we reject the null hypothesis.
   - If the F-statistic does not fall in the rejection region, or the p-value is greater than or equal to the significance level, then the null hypothesis cannot be rejected.
7. *Interpret in the context of the situation:*
   Rejecting the null hypothesis (H_0) gives us sufficient statistical evidence to conclude that at least one of the parameters is not zero. This means that at least one of the predictors (X1, X2, ..., Xp) has a linear association with the response variable, Y, and their overall relationship is significant. On the other hand, failing to reject the null hypothesis suggests that none of the predictors (X1, X2, ..., Xp) are linearly associated with Y, and we do not have enough evidence to confirm a significant relationship between the predictors and the response variable.

Let us apply the *F* test to Office Rental Price multiple regression problem. With two independent variables, the hypotheses are written as follows:

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \text{and/or } \beta_2 \text{ is not equal to zero}$$

The following R code demonstrates the entire procedure for performing an F-Test. It includes obtaining critical values, computing the F-statistic using two different formulas, and extracting and displaying F-values and p-values. and interpreting the results based on the significance level.

```
# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Floor = c(4, 7, 9, 5, 8, 4, 10, 12, 14, 9),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Fit the regression model
model <- lm(Rental_price ~ Size + Floor, data = data)

# Calculate the total sum of squares (TSS)
TSS <- sum((data$Rental_price - mean(data$Rental_price))^2)

# Calculate the residual sum of squares (RSS)
RSS <- sum(residuals(model)^2)
```

```r
# Get the R-squared value from the model
R_squared <- summary(model)$r.squared

# Number of predictors (p) and number of observations (n)
p <- length(coef(model)) - 1  # excluding the intercept
n <- nrow(data)

# Compute the F-statistic using the formula1
F_statistic_1 <- ((TSS - RSS) / p) / (RSS / (n - p - 1))

# Compute the F-statistic using the formula2
F_statistic_2 <- (R_squared / p) / ((1 - R_squared) / (n - p - 1))

# Display the F-statistic

cat("F-statistic using formula 1 :", round(F_statistic_1, 4), "\n")

## F-statistic using formula 1 : 71.8154

cat("F-statistic using formula 2 :", round(F_statistic_2, 4), "\n")

## F-statistic using formula 2 : 71.8154

# Get the summary of the regression model
model_summary <- summary(model)
# Extract the F-statistic value and its degrees of freedom (df)
F_statistic <- model_summary$fstatistic[1]


# Get the degrees of freedom for the model and residuals
df1 <- model_summary$fstatistic[2]  # Degrees of freedom for the model (p)
df2 <- model_summary$fstatistic[3]  # Degrees of freedom for the residuals (n
- p - 1)

# Significance level (alpha)
alpha <- 0.05

# Critical value for the F-distribution at 5% significance level
F_critical <- qf(1 - alpha, df1, df2)

# Calculate the p-value for the F-statistic
p_value <- 1 - pf(F_statistic, df1, df2)

# Display the results
cat("F-statistic:", round(F_statistic, 4), "\n")

## F-statistic: 71.8154

cat("Critical value:", round(F_critical, 4), "\n")

## Critical value: 4.7374
```

```r
cat("p-value:", round(p_value, 4), "\n")
```

```
## p-value: 0
```

```r
# Decision based on critical value and p-value approach
if (F_statistic > F_critical) {
  cat("Reject the null hypothesis using the critical value approach. The mode
l is significant.\n")
} else {
  cat("Fail to reject the null hypothesis using the critical value approach.
The model is not significant.\n")
}
```

```
## Reject the null hypothesis using the critical value approach. The model is
significant.
```

```r
if (p_value < alpha) {
  cat("Reject the null hypothesis using the p-value approach. The model is si
gnificant.\n")
} else {
  cat("Fail to reject the null hypothesis using the p-value approach. The mod
el is not significant.\n")
}
```

```
## Reject the null hypothesis using the p-value approach. The model is signif
icant.
```

Table 2.1 illustrates the general analysis of variance (ANOVA) table, presenting the F test results for a multiple regression model. The F test statistic's value, found in the last column, can be compared to F_α. This comparison involves p degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator to draw the conclusion of the hypothesis test.

Table 2.1: Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Value
Regression	ESS	p	$MSR = (TSS - RSS)/p$	$F = \frac{MSR}{MSE}$
Residual	RSS	$n-p-1$	$MSE = RSS/(n-p-1)$	
Total	TSS	$n-1$		

The `model` object created in the above code using the `lm()` function contains all the necessary information to construct an ANOVA table. However, it does not generate the table automatically. For a more comprehensive and well-organized display of regression results,

including coefficients and ANOVA tables, the ols_regress() function from the **olsrr** package can be used. To understand this better, try running the following code:

```
# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Floor = c(4, 7, 9, 5, 8, 4, 10, 12, 14, 9),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)
# Fit the regression model
model <- lm(Rental_price ~ Size + Floor, data = data)


if(!require("olsrr")) install.packages("olsrr")

ols_regress(model)
```

The first two tables of the output are:

```
                          Model Summary
-----------------------------------------------------------------
R                       0.976       RMSE                21.649
R-Squared               0.954       MSE                468.662
Adj. R-Squared          0.940       Coef. Var            5.687
Pred R-Squared          0.913       AIC                 97.878
MAE                    16.899       SBC                 99.088
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                                ANOVA
--------------------------------------------------------------------
                 Sum of
                Squares        DF    Mean Square      F          Sig.
--------------------------------------------------------------------
Regression     96163.377        2      48081.688    71.815    0.0000
Residual        4686.623        7        669.518
Total         100850.000        9
--------------------------------------------------------------------
```

2.3 t - Test

Once it is determined that at least one of the predictors plays a significant role, the next step is to identify which specific predictor(s) contribute meaningfully to the model. Introducing a new variable into a regression model will inevitably increase the regression sum of squares while decreasing the residual sum of squares. However, it is crucial to assess whether this increase is substantial enough to justify adding the new predictor. Additionally, incorporating extra predictors raises the variance of the predicted value $\hat{y}$, making it important to include only those variables that meaningfully explain the response. Furthermore, adding an irrelevant predictor could lead to an increase in the residual mean square, potentially diminishing the overall effectiveness of the model. Hence, a separate t-test is conducted for each independent variable in the model. These tests are referred to as *tests for individual significance*.

The hypotheses for testing the significance of any individual regression coefficient, such as β_j, are

$$H_0: \beta_j = 0, H_1: \beta_j \neq 0$$

If $H_0: \beta_j = 0$ is not rejected, then this indicates that the predictor x_j can be removed from the model. The test statistic for this hypothesis is

$$t = \frac{b_j}{s_{b_j}}$$

where s_{b_j} is the estimate of the standard deviation of b_j. The value of s_{b_j}will be provided by the computer software package. The null hypothesis is rejected if $|t| \geq t_{\alpha/2,n-p-1}$ or if p-value $\leq \alpha$.

To illustrate the procedure, consider the Office Rental Price data given in Unit 1 of Block 2. Suppose we wish to assess the value of the predictor variable x_2 (floor) given that the predictor x_1 (size) is in the model. The hypotheses are

$$H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$$

The t statistics can be computed as

$$t = \frac{b_2}{s_{b_2}} = \frac{4.759}{3.683} = 1.292$$

This result indicates that the regressor x_2 (floor) does not significantly contribute to the model. The p-value associated with this t statistic is `0.237`, (From Table 2.2), which is greater than the significance level, leading to the conclusion that Floor is not a significant predictor of Rental Price.

Moreover, the confidence interval **Lower = -3.950, Upper = 13.467** for Floor contains zero, meaning there is a possibility that the true value of the coefficient could be zero. This further supports the conclusion that Floor may not significantly contribute to predicting Rental Price.

Table 2.2: Statistical Significance for Rental Price

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	15.586	39.403		0.396	0.704	-77.587	108.759
Size	0.554	0.074	0.865	7.470	0.000	0.379	0.729
Floor	4.759	3.683	0.150	1.292	0.237	-3.950	13.467

On the other hand, the t-statistic for Size is quite large (**t = 7.470)**, which suggests that the coefficient for Size is significantly different from zero. The p-value for Size is very low (significantly less than 0.05), which means that Size is statistically significant in predicting the Rental Price. With 95% confidence, we can say that the true value for the Size coefficient lies between 0.379 and 0.729. Since this interval does not contain zero, the coefficient is significantly different from zero.
In short, the Size variable is a significant predictor of Rental Price, while the Floor variable isn't as important in this case.

Practical Considerations: Even though **Floor** is not statistically significant in this case, you may want to consider if there are practical or domain-specific reasons why it should remain in the model. For example, in some real-world situations, the number of floors might still have a theoretical importance (e.g., buildings with more floors might still generally have higher rental prices due to other factors, like views or prestige).

Check Your Progress – 1

1. Refer to the data presented in "Check Your Progress 1" in Unit 1 of this Block.

X1	25	30	47	51	40	36	74	51	76	59
X2	17	12	10	16	5	12	7	19	16	13
Y	112	94	108	178	94	117	170	175	211	142

(a) Verify the values for MSR and MSE.
(b) Use an *F* test and a .05 level of significance to determine whether there is a relationship among the variables.
(c) Use α = .05 to test the significance of β_1. Should X1 be dropped from the model?
(d) Use α = .05 to test the significance of β_2. Should X2 be dropped from the model?

2.4 LET US SUM UP

In this unit, we explored how to conduct significance tests in multiple regression analysis. We discussed the use of the F-test to determine whether there is a significant relationship between the dependent variable and all independent variables in the model (overall significance). Once overall significance is established, we then use the t-test to evaluate the individual contributions of each independent variable to the model (individual significance). Through the Office Rental Price Example, we demonstrated the practical application of both the F-test and t-test, highlighting their distinct purposes and how they complement each other in multiple regression analysis.

2.5 Check Your Progress: Possible Answers

Check Your Progress – 1

Solution for Problem Set 1 using R code.

```
# Given Data
X1 <- c(25, 30, 47, 51, 40, 36, 74, 51, 76, 59)
X2 <- c(17, 12, 10, 16, 5, 12, 7, 19, 16, 13)
Y <- c(112, 94, 108, 178, 94, 117, 170, 175, 211, 142)
data <- data.frame(X1, X2, Y)

# Regression model relating Y to both X1 and X2
model <- lm(Y ~ X1 + X2, data = data)
ols_regress(model)
```

```
                         Model Summary
 ----------------------------------------------------------------
 R                       0.962       RMSE                10.634
 R-Squared               0.926       MSE                113.075
 Adj. R-Squared          0.904       Coef. Var            9.072
 Pred R-Squared          0.880       AIC                 83.659
 MAE                     8.107       SBC                 84.870
 ----------------------------------------------------------------
  RMSE: Root Mean Square Error
  MSE: Mean Square Error
  MAE: Mean Absolute Error
  AIC: Akaike Information Criteria
  SBC: Schwarz Bayesian Criteria
```

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	14052.155	2	7026.077	43.496	1e-04
Residual	1130.745	7	161.535		
Total	15182.900	9			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-18.368	17.972		-1.022	0.341	-60.864	24.128
X1	2.010	0.247	0.840	8.134	0.000	1.426	2.595
X2	4.738	0.948	0.516	4.995	0.002	2.495	6.981

2.6 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
4. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

2.7 Assignment

1. Explain how the F-test is used to assess the overall significance of a regression model.
2. What does a significant F-test imply about the relationship between the predictors and the response variable?
3. Discuss how the t-test is used to evaluate the significance of individual predictors in a regression model.
4. For the dataset provided in Assignment 1.9 (Unit 1), answer the following questions:
   (a) Formulate the null and alternative hypotheses, perform *F* test with a .05 level of significance to determine whether there is a relationship among the variables.
   (b) Formulate the null and alternative hypotheses and perform a t-test with a 0.05 level of significance to determine whether the predictor variables are significant in the regression model.

Unit 3: Model Diagnostic and Residual Analysis

Unit Structure

3.0 Learning Objectives

After reading this unit, you should be able to:

- Define residuals and explain their significance in regression analysis.
- Understand the key assumptions of linear regression and their role in model accuracy.
- Identify and interpret various types of residuals, including their implications for model diagnostics.
- Learn how to use residual plots to visually diagnose model issues.
- Identify influential observations using leverage and Cook's Distance.
- Use statistical software (e.g., R) to perform comprehensive regression diagnostics.

3.1 Introduction

In the previous units, we explored the fundamental aspects of regression analysis, including:

- The construction of regression models,
- The least squares method,
- The interpretation of R^2 as a measure of model fit, and
- Hypothesis testing using t-tests and F-tests in simple and multiple regressions.

These tools help us evaluate the relationships between variables and assess the statistical significance of our model. However, assessing a model's reliability goes beyond statistical significance—we must also examine whether the model meets key assumptions and provides accurate predictions.

Residual analysis plays a crucial role in this process. By analyzing the differences between observed and predicted values, we can diagnose potential issues such as:

- *Non-linearity* (when a linear model does not adequately describe the relationship),
- *Heteroscedasticity* (when error variance is not constant), and
- *Outliers* or *influential* points that can distort model estimates.

Understanding residuals and their properties allows us to refine regression models and ensure their validity. By the end of this unit, you will be equipped with the skills to assess your regression model using residual analysis and make necessary improvements.

3.2 The Regression Assumptions

The major **assumptions** that we have made thus far in our study of regression analysis are as follows:

3.2.1 Linearity Assumption:

The relationship between dependent and independent variables must be **linear**:

$$E\left(Y\middle|(X_1, X_2, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{3.1}$$

which implies that the i^{th} observation, for $i = 1, 2, \cdots, n$, can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{3.2}$$

A scatter plot of Y versus X can help assess linearity in the case of simple linear regression. Checking for linearity in multiple regression is more challenging due to the high dimensionality of the data. If non-linear patterns exist, transformations (e.g., logarithmic, polynomial regression) may be needed.

3.2.2 Assumptions About the Errors

In a linear regression model, the error terms $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ are assumed to follow specific statistical properties to ensure valid inferences. These assumptions include:

(a) **Normality of Errors:** The error terms are expected to be normally distributed. This assumption is particularly important for hypothesis testing and confidence interval estimation. Assessing normality is more challenging when predictor variables are not replicated, but it can be evaluated using graphical methods such as histograms, Q-Q plots, or statistical tests.

(b) **Zero Mean:** The average of the error terms should be zero, ensuring that the regression model does not systematically overestimate or underestimate the dependent variable.

(c) **Constant Variance (Homoscedasticity):** The error terms should have a uniform variance (σ^2) across all levels of the independent variables. If this condition is violated, it results in heteroscedasticity, where residual variance changes with predictor values, potentially leading to inefficient estimates and biased inference. Methods such as residual plots can help to detect this issue.

(d) **Independence of Errors:** The errors should be independent of one another, meaning that the value of one error term should not be influenced by another. If errors are correlated (as often seen in time-series data), it results in **autocorrelation**, which can affect the efficiency of regression estimates. Autocorrelation is commonly tested using the Durbin-Watson statistic.

3.2.3 Assumptions About the Predictors:

In regression analysis, certain assumptions about the predictor variables $(X_1, X_2, \dots, X_p)$ must be met to ensure valid and reliable model estimates. These assumptions include:

(a) **Non-Random Predictors:** The predictor variables are typically assumed to be fixed, meaning their values are predetermined rather than random. This assumption holds in experimental settings where researchers control the predictor values. However, in observational studies, predictors are often random, and as a result, statistical inferences must be interpreted conditionally based on the observed data.

(b) **Measurement Accuracy:** It is assumed that predictor variables are measured without error. In practice, measurement errors can occur, particularly in fields like social sciences, where precise measurements are difficult to obtain. Such errors can distort the residual variance, affect the multiple correlation coefficient, and introduce bias in regression coefficients. The extent of this impact depends on factors like the standard deviation of measurement errors and correlations between

variables. Although correcting for measurement errors is complex, their effects should be considered when interpreting regression results.

(c) **No Perfect Multicollinearity:** The predictor variables should not be perfectly correlated with each other. If two or more predictors exhibit a near-perfect linear relationship, it leads to multicollinearity, which makes it difficult to estimate individual regression coefficients accurately. This problem affects the stability of the regression model and can inflate standard errors, leading to unreliable statistical inferences. Techniques such as **Variance Inflation Factor (VIF) analysis** and **principal component analysis (PCA)** can help detect and address multicollinearity.

While the first two assumptions are often difficult to verify directly, they influence how regression results should be interpreted. Ensuring that predictor variables are appropriately chosen and free from severe multicollinearity helps maintain the integrity of the regression model.

Check Your Progress – 1

1. Which assumption helps detect non-constant variance?

(a)	Independence of Errors	(b)	Homoscedasticity
(c)	Zero Mean	(d)	Normality of Errors

2. What does multicollinearity refer to?

(a)	Error terms having unequal variance	(b)	Predictor variables being perfectly correlated
(c)	Model overfitting	(d)	Regression model instability

3. Which method can help detect multicollinearity?

(a)	Durbin-Watson Test	(b)	Variance Inflation Factor (VIF)
(c)	Q-Q Plot	(d)	Residual Plot

3.3 Various Types of Residuals

Residuals are a key diagnostic tool in regression analysis, used to assess model performance and identify potential violations of regression assumptions. By analyzing residuals, we can uncover underlying model deficiencies that may not be evident from summary statistics alone. Different types of residuals serve distinct purposes in model diagnostics, offering valuable insights for improving the model and better understanding the data.

3.3.1 Ordinary Least Squares (OLS) Residuals

The most basic type of residual is the **Ordinary Least Squares (OLS) residual**, which is the difference between observed values (y_i) and the predicted values ($\hat{y}_i$) from a regression model:

$$e_i = y_i - \hat{y}_i$$

These residuals provide an initial assessment of how well the model fits the data:

- If the residuals are *small and randomly distributed*, it indicates a good model fit.
- If residuals show a *pattern*, it suggests that the model is not adequately capturing the data structure.
- The *sum of residuals is always zero* in a correctly specified linear regression model.

However, one issue with OLS residuals is that their *variance is not constant* across all observations due to differences in leverage (some points have more influence on the model than others). This issue is addressed using standardized and studentized residuals.

3.3.2 Standardized Residuals

Standardized residuals are often used in residual analysis. They are calculated by subtracting the mean and dividing by the standard deviation. For residuals from the least squares method, the mean is zero, so each residual is divided by its standard deviation to standardize it. Standardizing is useful as it makes residuals comparable across different observations despite their non-constant variance in OLS. The *standardized residual* is given by:

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}$$

where:

- σ is the estimated standard deviation of the residuals.
- p_{ii} is the **leverage value**, which indicates how much influence an observation has on its own fitted value.

Standardized residuals are useful in identifying **outliers**—values greater than ±2 or ±3 may indicate extreme observations. However, standardized residuals still depend on the overall estimate of error variance and may not always be the best indicator of influential points. This leads us to *studentized residuals*, which provide a better approach.

Leverage

Leverage is quantified using the **hat matrix**, which transforms the observed values into fitted values. The regression model estimates fitted values as:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

An alternative way to express this is:

$$\hat{y}_i = p_{i1} y_1 + p_{i2} y_2 + \cdots + p_{in} y_n$$

where p_{ij} represents elements of the **hat matrix**. These elements depend only on the predictor variables and not on the response variable.

In simple linear regression, the leverage values are computed as:

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2}$$

For multiple regression, leverage values are the diagonal elements of the **hat matrix** H:

$$H = X(X^T X)^{-1} X^T$$

where H is the projection matrix, and each diagonal element p_{ii} represents the leverage of observation i.

3.3.3 Studentized Residuals (Internally Studentized Residuals)

To further refine residual analysis, we use *studentized residuals*, also known as *internally studentized residuals*. These are calculated as follows:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}$$

where $\hat{\sigma}$ is the estimated standard deviation of the residuals.

Unlike standardized residuals, studentized residuals account for variability in each observation more effectively. They are better suited for detecting *outliers* in regression analysis. However, since they still depend on the dataset as a whole, a more robust approach is to use *externally studentized residuals*.

3.3.4 Externally Studentized Residuals (Deleted Studentized Residuals)

Externally studentized residuals, also called *deleted studentized residuals or R - student* , improve upon internally studentized residuals by removing the effect of the observation being evaluated. The formula is:

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - p_{ii}}}$$

where $\hat{\sigma}_{(i)}$ is the standard deviation estimated **without** the i^{th} observation.

Check Your Progress – 2

1. What is the main drawback of OLS residuals?

(a)	They are difficult to compute
(b)	They do not account for leverage
(c)	They cannot detect outliers
(d)	They always have a non-zero sum

2. How do studentized residuals differ from standardized residuals?

(a)	They use the mean of residuals
(b)	They account for variability in each observation
(c)	They ignore leverage values
(d)	They are the same as OLS residuals

3. What does a high externally studentized residual indicate?

(a)	A strong linear relationship	(b)	A significant outlier
(c)	A normally distributed residual	(d)	An insignificant predictor variable

3.4 Visualizing Residuals

Visual inspection of residuals is a crucial diagnostic technique for assessing the assumptions of regression models. Residual plots are graphical tools used to evaluate the adequacy of a regression model and detect violations of key assumptions such as non-linearity, heteroscedasticity, and outliers. These plots are typically generated using statistical software (e.g., R, Python, SPSS) and should be examined routinely in regression analysis. In this section, we will explore common residual plots and their interpretation.

To make meaningful interpretations, **internally studentized residuals** are often used because they have constant variance.

3.4.1 Normal Probability Plot (Q-Q Plot)

The *normal probability plot* is a graphical tool used to assess whether the error term in a regression model follows a normal distribution. This plot is developed using the concept of *normal scores*, which are based on the expected values of ordered statistics from a standard normal distribution.

Concept of Normal Scores

1. Suppose you repeatedly draw random samples of size 10 from a standard normal distribution (mean = 0, standard deviation = 1).
2. For each sample, order the values from smallest to largest.
3. The smallest value in each sample is called the *first-order statistic*.
4. Statisticians have determined that for a sample size of 10, the expected value of the first-order statistic is -1.55. This expected value is called a *normal score*.

5. For a sample of size $n = 10$, there are 10 order statistics and 10 corresponding normal scores (see the R code in Example section).
6. In general, for a dataset with n observations, there will be n order statistics and n normal scores.

Constructing the Normal Probability Plot

1. Order the standardized residuals from smallest to largest.
2. Pair each ordered standardized residual with its corresponding normal score.
3. Plot the normal scores on the **horizontal axis** and the ordered standardized residuals on the **vertical axis**.

Interpreting the Plot

- If the standardized residuals are approximately normally distributed, the plotted points should cluster closely around a **45-degree line** passing through the origin.
- Deviations from this line indicate departures from normality:
    - **S-shaped or curved patterns** suggest skewness.
    - **Heavy tails** indicate the presence of outliers or extreme values.

3.4.2 Residuals vs. Predictor Variables Plot

A **residual plot against the independent variable $\mathbf{X}$** is a graph where the independent variable values are plotted on the horizontal axis, and the corresponding residual values are plotted on the vertical axis. Each residual is represented by a point on the plot.

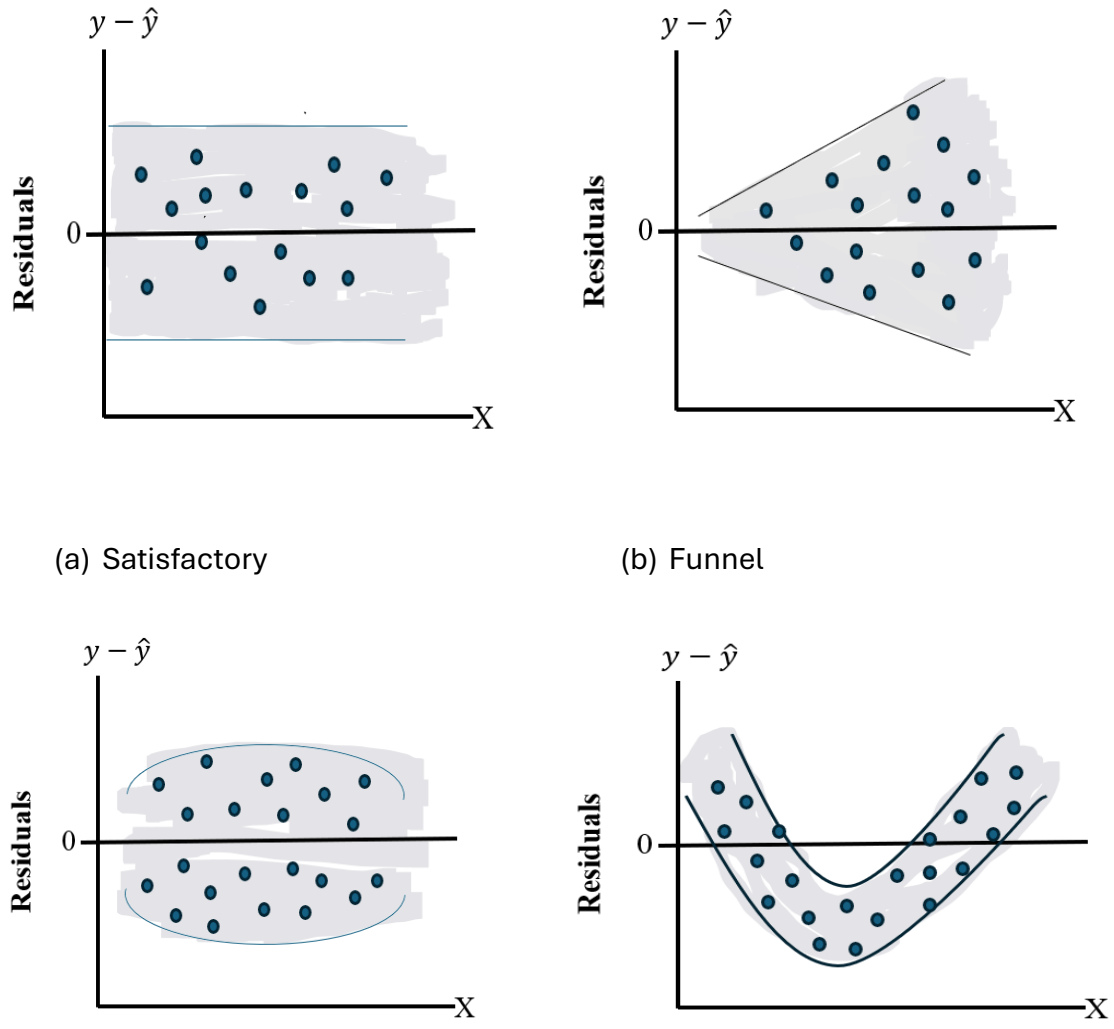
Before interpreting this plot, let's examine some common patterns that might appear in any residual plot. Three examples are illustrated in **Figure 3.1**.

1. **Ideal Case (Panel a)**:
   If the variance of the error term (ε) is constant for all values of the predictor variable (x) and the regression model accurately represents the relationship between the variables, the residual plot should resemble a **horizontal band of points**. This indicates that the model assumptions are satisfied.

2. **Non-Constant Variance (Panel b)**:
   If the variance of ε is not the same for all x values—for example, if there is greater variability around the regression line for larger x values—the residual plot might display an **outward-opening funnel pattern**. This suggests that the assumption of constant variance is violated. An **inward-opening funnel pattern** could also occur, indicating that the variance increases as y decreases.

3. **Double-Bow Pattern (Panel c)**:
   This pattern often appears when y represents a proportion between 0 and 1. The variance of a binomial proportion is typically greater near 0.5 than near 0 or 1. To address this, **variance-stabilizing transformations** (e.g., log or square root) of the response variable are often used.

4. **Nonlinear Relationship (Panel d)**:
   If the residual plot shows a **curved pattern**, it suggests that the assumed regression model does not adequately capture the relationship between the variables. In such cases, consider using a **curvilinear regression model** or a **multiple regression model** with additional terms (e.g., quadratic or interaction terms).



(a) Satisfactory (b) Funnel

(c) Double bow (d) Non-linear

Figure 3.1: Patterns for residual plots

3.4.3 Residuals vs. Fitted Values Plot

These plots often exhibit patterns like those shown in Figure 3.1, but the horizontal axis represents $\hat{y}$ (the predicted values) instead of the predictor variable X. Therefore, the pattern

observed in this residual plot is the same as that in the residual plot against the independent variable X. For simple linear regression, both the residual plot against X and the residual plot against $\hat{y}$ exhibit the same pattern. However, for multiple regression analysis, the residual plot against $\hat{y}$ is more commonly used because there are multiple independent variables involved.

Check Your Progress – 3

1. What is the purpose of a residual plot?

(a)	To check model
(b)	To predict new observations
(c)	To calculate regression coefficients
(d)	To standardize residuals

2. What does a funnel shape in a residual plot indicate?

(a)	Homoscedasticity	(b)	Heteroscedasticity
(c)	Multicollinearity	(d)	Normality

3. A Q-Q plot is used to assess:

(a)	Linearity assumption	(b)	Normality of residuals
(c)	Homoscedasticity	(d)	Independence of errors

3.5 Leverage and Influence

An observation is considered *influential* if its removal—either individually or in combination with a few others (e.g., two or three)—causes substantial changes in the fitted regression model. Such changes may include significant shifts in the estimated coefficients, fitted values, t-tests, or other key regression outputs. While the deletion of any data point will generally cause some changes in the model, we are particularly interested in identifying those points whose removal has a disproportionately large impact (i.e., they exert undue influence).

3.5.1 Identifying Influential Observations

Influential observations can often be identified from a scatter diagram when only one independent variable is present. An influential observation may:

1. Be an *outlier* (an observation with a y value that deviates substantially from the trend).
2. Correspond to an x value far away from its mean.
3. Result from a combination of the two (a somewhat off-trend y value and a somewhat extreme x value).

Observations with extreme values for the independent variables are called *high leverage points*. The leverage of an observation, denoted as h_i, measures how far the values of the independent variables are from their mean values. High leverage points have the potential to influence the regression results significantly.

- A common rule of thumb is to flag observations as influential if their leverage exceeds the threshold:

$$h_i > \frac{2(p+1)}{n}$$

  where:
  - p = number of independent variables,
  - n = number of observations.
- For example, in the Rental Price dataset with $p = 2$ independent variables and $n = 10$ observations, the critical leverage value is:

$$\frac{2(2+1)}{10} = 0.6$$

Influential observations caused by an interaction of large residuals and high leverage can be particularly difficult to detect. Diagnostic procedures are available to account for both factors when determining whether observation is influential. One such measure is **Cook's D statistic**, which quantifies the influence of an observation by considering both its residual and leverage.

Influence is measured using **Cook's Distance**:

$$D_i = \frac{r_i^2}{p+1} \times \frac{p_{ii}}{1-p_{ii}}$$

where:
- r_i is the studentized residual,
- p_{ii} is the leverage value,
- $p + 1$ is the number of parameters in the model.

A high leverage point with a small residual is not necessarily problematic. However, a high leverage point with a large residual can disproportionately affect the regression model.

Check Your Progress – 4

1. An observation with high leverage and large residuals is:

(a)	A regular data point	(b)	An influential observation
(c)	A low-impact observation	(d)	Always an outlier

2. According to the rule of thumb, when is an observation flagged as influential based on its leverage value?

(a)	When its leverage value exceeds 1
(b)	When its leverage value is less than $2(p+1)\ /\ n$
(c)	When its leverage value exceeds $2(p+1)\ /\ n$
(d)	When its leverage value is exactly equal to $2(p+1)\ /\ n$

3. Which of the following measures is used to quantify the influence of an observation by considering both its residual and leverage?

(a)	Studentized Residual
(b)	Mean Square Error
(c)	R-squared Value
(d)	Cook's Distance

3.6 Practical Application – Estimating Rental Prices

3.6.1 Dataset Description

The dataset consists of 10 observations with the following variables: - **Size**: Property size in square feet. - **Floor**: Floor level of the property. - **Broadband Rate**: Internet speed in Mbps. - **Rental Price**: Monthly rental price in ₹100s.

Table 3.1: Office Rental Price data

Location	Size	Floor	Broadband Rate	Rental price
1	500	4	80	320
2	550	7	500	380
3	620	9	70	400
4	630	5	240	390
5	660	8	1000	380
6	700	4	80	410
7	770	10	70	480
8	880	12	500	600
9	920	14	80	570
10	1000	9	240	620

We begin by fitting a multiple linear regression model to predict rental price based on property size, floor level, and broadband rate.

R Code for Model Fitting

```r
# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Floor = c(4, 7, 9, 5, 8, 4, 10, 12, 14, 9),
  Broadband_Rate = c(80, 500, 70, 240, 1000, 80, 70, 500, 80, 240),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Fit the regression model
model <- lm(Rental_price ~ Size + Floor + Broadband_Rate, data = data)
```

3.6.2 Residual Analysis

Residual analysis helps us assess whether the regression assumptions are satisfied. We will generate and interpret the following residual plots:

R Code for Residual Plots

```r
# Generate residual plots
par(mfrow = c(2, 2))  # Arrange plots in a 2x2 grid
plot(model, which = 1)  # Residuals vs. Fitted
plot(model, which = 2)  # Normal Q-Q Plot
plot(model, which = 3)  # Scale-Location Plot (for heteroscedasticity)
plot(model, which = 5)  # Residuals vs. Leverage
```

3.6.3 Leverage and Influence Diagnostics

Leverage Values

- **Purpose**: Identify observations with extreme predictor values.
- **Rule of Thumb**: Observations with leverage, $h_i > \frac{2(p+1)}{n}$ are high leverage points.

Cook's Distance

- **Purpose**: Measure the influence of an observation on the regression model.
- **Rule of Thumb**: Observations with Cook's Distance, $D_i > 1$ are influential.

R Code for Leverage and Cook's Distance

```r
# Compute leverage values
leverage_values <- hatvalues(model)

# Compute Cook's Distance
cooks_distance <- cooks.distance(model)

# Create a diagnostics table
diagnostics_table <- data.frame(
  Predicted_Values = round(predict(model), 2),
```

```
  OLS_Residuals = round(residuals(model), 2),
  Standardized_Residuals = round(rstandard(model), 2),
  Studentized_Residuals = round(rstudent(model), 2),
  Leverage = round(leverage_values, 2),
  Cooks_Distance = round(cooks_distance, 2)
)

# Display the diagnostics table
print(diagnostics_table)
```

3.6.4 Testing Regression Assumptions Numerically

Multicollinearity

Variance Inflation Factor (VIF) checks if predictors are highly correlated.

```
library(car)
vif(model)
```

- **VIF > 10**: High multicollinearity, consider removing or transforming variables.

Homoscedasticity

Breusch-Pagan test checks for constant variance in residuals.

```
library(lmtest)
bptest(model)
```

- **p-value < 0.05**: Heteroscedasticity present, consider transformations.

Normality of Residuals

Shapiro-Wilk test assesses normality.

```
shapiro.test(residuals(model))
```

- **p-value < 0.05**: Residuals are not normally distributed.

Autocorrelation

Durbin-Watson test detects correlated residuals.

```
library(lmtest)
dwtest(model)
```

- **p-value < 0.05**: Presence of autocorrelation.

3.7 LET US SUM UP

Residual analysis is essential for verifying regression model assumptions and improving model accuracy. By using different types of residuals and graphical tools, we can detect potential issues and refine our regression approach. By understanding leverage and its relationship with residuals, we can improve the robustness and accuracy of our regression models. Standardized and studentized residuals, along with Cook's Distance, help in detecting influential points.

Type of Residual	Formula	Key Features	Best Use Case
OLS Residuals	$e_i = y_i - \hat{y}_i$	Basic residuals; sum to zero.	Initial model evaluation.
Standardized Residuals	$z_i = \frac{e_i}{\sigma\sqrt{1-p_{ii}}}$	Mean = 0, Variance = 1.	Outlier detection.
Internally Studentized Residuals	$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$	More accurate than standardized residuals.	Detecting large residuals.
Externally Studentized Residuals	$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}$	Removes observation's influence.	Identifying influential points.

The normal probability plot is particularly useful for identifying subtle departures from normality that might not be apparent in summary statistics. Residual Vs. Fitted values plot is: (a) Satisfactory: The residuals are randomly scattered around the horizontal axis (zero), showing no obvious pattern. (b) Funnel: The residuals form a pattern that widens (or narrows) like a funnel, indicating increasing (or decreasing) variance. (c) Double Bow: The residuals form a bow-like pattern on both sides of the plot, suggesting an issue with the model fit. (d) Nonlinear: The residuals display a systematic, curved pattern, indicating the model is missing a nonlinear relationship. By following the steps outlined in the Rental Price dataset, you can assess the validity of your regression model, identify potential issues, and refine the model for better accuracy and reliability.

3.8 Check Your Progress: Possible Answers

Check Your Progress – 1

Question No.	Correct option
1.	(b)
2.	(b)
3.	(b)

Check Your Progress – 2

Question No.	Correct option
1.	(b)
2.	(b)
3.	(b)

Check Your Progress – 3

Question No.	Correct option
1.	(a)
2.	(b)
3.	(b)

Check Your Progress – 4

Question No.	Correct option
1.	(b)
2.	(c)
3.	(d)

3.9 Further Reading

1. Introduction to Linear Regression Analysis 6th Edition, Montgomery, Peck, Vining, Wiley Publication, February 2021
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
4. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

3.10 Assignment

1. Explain the importance of residual analysis in regression modeling.
2. Describe the various types of residuals.
3. Why is homoscedasticity important in regression analysis? How can it be checked?
4. Explain the significance of Q-Q plots in diagnosing regression models.
5. Describe common patterns in residual plots and their implications.
6. Discuss the significance of leverage in identifying influential observations.
7. Apply these techniques to following simulated dataset and evaluate the assumptions numerically.

Sr No.	Y	X1	X2	X3	Sr No.	Y	X1	X2	X3
1	0.18495	0.00298	0.30133	0.44701	26	1.12376	0.95081	0.44695	0.66161
2	0.73951	0.91095	0.53212	0.06239	27	0.95351	0.49113	0.87509	0.45171
3	0.28833	0.11112	0.8525	0.29337	28	0.18831	0.00112	0.80255	0.39571
4	0.03079	0.19965	0.90261	0.18615	29	0.62951	0.89237	0.45912	0.31139
5	0.48922	0.45139	0.70902	0.70707	30	-0.59231	0.34513	0.70645	0.39591
6	1.14236	0.95071	0.44683	0.66169	31	0.78144	0.71133	0.77547	0.28786
7	-0.02754	0.38579	0.19394	0.27554	32	0.01493	0.25256	0.47111	0.32147
8	0.69319	0.31847	0.96495	0.03485	33	0.93451	0.29061	0.6849	0.19582
9	0.18895	0.00298	0.30133	0.44701	34	-0.89132	0.59231	0.46128	0.01562
10	-0.33177	0.94081	0.87203	0.06583	35	0.51261	0.89706	0.13126	0.34616
11	0.13493	0.55456	0.80561	0.57634	36	0.51261	0.89706	0.13126	0.34616
12	0.68922	0.19839	0.72309	0.47101	37	0.68822	0.25128	0.71845	0.37461
13	-0.19753	0.30482	0.5702	0.39127	38	0.48131	0.29325	0.48327	0.30517
14	1.01132	0.08762	0.17431	0.89149	39	0.56913	0.29367	0.78345	0.37128
15	-0.59812	0.19463	0.80648	0.01294	40	-0.11831	0.18319	0.55498	0.29172
16	0.48161	0.39124	0.72539	0.38562	41	0.48451	0.29251	0.48371	0.30521
17	0.69213	0.52227	0.97097	0.71516	42	0.72253	0.49163	0.61742	0.22139
18	-0.49172	0.30482	0.80745	0.39127	43	-0.11831	0.17231	0.80153	0.03127
19	0.14801	0.25867	0.38172	0.30494	44	0.35631	0.68372	0.70419	0.31562
20	0.14893	0.29235	0.48073	0.40191	45	1.12376	0.95081	0.44695	0.66161
21	0.41918	0.58372	0.80549	0.40172	46	0.95351	0.49113	0.87509	0.45171
22	0.15639	0.29215	0.48022	0.40162	47	0.18831	0.00112	0.80255	0.39571
23	1.11132	0.09762	0.07131	0.95195	48	0.62951	0.89237	0.45912	0.31139
24	0.16935	0.28385	0.68363	0.32419	49	-0.59231	0.34513	0.70645	0.39591
25	-0.49763	0.38362	0.82087	0.05891	50	0.78144	0.71133	0.77547	0.28786

Block 3: Data Transformations and Qualitative Predictors

Introduction

Regression analysis is an essential tool for understanding and predicting relationships between variables. However, real-world data often does not meet the assumptions required for effective regression modeling. To address this, *data transformations* and the *handling of qualitative predictors* play a critical role in improving the accuracy and interpretability of regression models.

In *Unit 1: Transforming Predictor Variables*, we will explore how to apply various transformations to predictor variables. Transformations such as natural logarithms, polynomial, and reciprocal transformations are used to address issues like skewness, non-linearity, and unequal variance. These adjustments help to ensure that the model can accurately capture relationships between the dependent and independent variables.

Unit 2: Advanced Transformations focuses on more complex transformations that apply to both the response and predictor variables. This unit will cover methods such as the natural logarithm transformation, Box-Cox transformations, and scaling techniques like centering and standardization. These techniques are essential in addressing non-linearity, heteroscedasticity, and skewed distributions, which can impact model assumptions and predictions.

In *Unit 3: Transforming Qualitative Predictors*, we dive into how to incorporate categorical predictors into regression models, particularly by transforming qualitative predictors into dummy variables. Dummy variables are used to represent qualitative factors, such as gender or payment method, allowing them to be analyzed within a regression framework. We will also explore how to interpret interactions between numerical and categorical predictors and assess the combined effects of these variables on model performance.

By the end of this block, you will gain a deep understanding of how to transform both quantitative and qualitative predictors, ensuring that your regression models are more robust, interpretable, and effective in capturing real-world relationships.

Unit 1 Transforming Predictor Variables

Unit Structure

1.0 Learning Objectives

After completing this unit, you should be able to

- Apply natural logarithm transformations to skewed predictor variables in multiple linear regression to improve symmetry, normality, and model effectiveness.
- Use polynomial transformations to capture non-linear relationships and enhance model performance.
- Understand and apply the principle of preserving hierarchy in polynomial regression by including all lower-order terms up to the highest significant power.
- Implement reciprocal transformations on predictor variables to improve model effectiveness.
- Conduct an extensive review of model performance using metrics such as R-squared, Adjusted R-squared, and standard error to evaluate and compare models.

1.1 Introduction

Data often requires transformation before analysis to meet objectives like ensuring linearity, achieving normality, or stabilizing variance. Model building in regression involves creating a regression equation to define the relationship between dependent and independent variables, addressing violations of regression assumptions through data transformation. By adjusting the scale or metric of the response or regressor variables, issues such as unequal variance can be resolved. It's common to fit linear regression models to transformed variables rather than the original ones. This unit explores when transformations are needed, the types of transformations available, and how to analyze transformed data. While illustrated using simple regression, transformations in multiple regression demand more effort and care when some predictors require transformation and others do not.

1.2 Need for Transformation

A model is linear if its parameters are linear, even if predictors are nonlinear. For example, each of the four following models:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 \log X + \varepsilon$$

$$Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$$

because the model parameters β_0, β_1 and β_2 enter linearly. On the other hand,

$$Y = \beta_0 + e^{\beta_1 X} + \varepsilon$$

is a nonlinear model because the parameter β_1 does not enter the model linearly. To meet the assumptions of the standard regression model, we sometimes need to work with transformed variables instead of the original ones. Transformations may be necessary for a variety of reasons:

1. Nonlinear Models with Inherent Linear Properties:

   In the case of the exponential model, we can apply a transformation to the variables, which allows us to perform regression analysis using the general linear model. The exponential model is given by the following regression equation:

   $$E(y) = \beta_0 \, \beta_1^x$$

   This equation is appropriate when the dependent variable y changes by a constant percentage as x increases, rather than by a fixed amount.

We can transform this nonlinear regression equation to a linear regression equation by taking the natural logarithm of both sides of equation

$$log\big(E(y)\big) = log(\beta_0) + x\ log(\beta_1)$$

This shows that $log\big(E(y)\big)$ and x are linearly related, allowing us to use standard regression methods. Despite the original variables having a nonlinear relationship, the transformation makes the relationship between the transformed variables linear. Thus, transformation is used to achieve the linearity of the fitted model.

2. The response variable Y, which is being analyzed, may have a probability distribution where its variance is related to the mean. If the mean depends on the predictor variable X, the variance of Y will change as X changes, resulting in non-constant variance. Under these conditions, the distribution of Y is typically non-normal. Non-normality can affect the validity of standard significance tests, which rely on the assumption of normality (though this issue is less significant with large sample sizes). When the error terms have unequal variance, the estimates remain unbiased, but they are no longer optimal in terms of having the smallest variance. To address this, we often transform the data to ensure normality and constant error variance. These transformations are typically chosen to stabilize the variance (variance-stabilizing transformations), and it is a fortunate coincidence that these transformations also tend to normalize the data effectively.
3. There is no prior theoretical or probabilistic basis to suggest that a transformation is needed. The need for a transformation becomes evident when we examine the residuals from fitting a linear regression model using the original variables.

Transformations can help address these issues by altering the scale or shape of the data, making the relationship more linear, stabilizing the variance, and improving the normality of residuals.

1.3 Natural Logarithm Transformation for Predictors

In statistical modeling and data science, it is common to encounter predictors (independent variables) that are highly skewed or have a wide range of values. Such variables can lead to issues like heteroscedasticity, non-linearity, or undue influence of outliers. One common technique to address these issues is *natural logarithm transformation*. This transformation can help stabilize variance, make the data more normally distributed, and improve the interpretability of the model.

The natural logarithm transformation is particularly useful when:

1. The predictor has a right-skewed distribution.
2. The predictor spans several orders of magnitude (e.g., income, population size).

3. The relationship between the predictor and the response variable is multiplicative rather than additive.

The natural logarithm transformation applies the `log()` function to the predictor variable. For a predictor X, the transformed variable is:

$$X_{\text{transformed}} = \log(X)$$

Here, `log()` refers to the natural logarithm (base e).

1.3.1 Example in R

Let's walk through an example using R. We'll use the `mtcars` dataset, which is included in R by default. Suppose we want to model the relationship between a car's weight (`wt`) and its miles per gallon (`mpg`). We'll apply a natural logarithm transformation to the `wt` variable to improve the linearity of the relationship.

Step 1: Load the Data

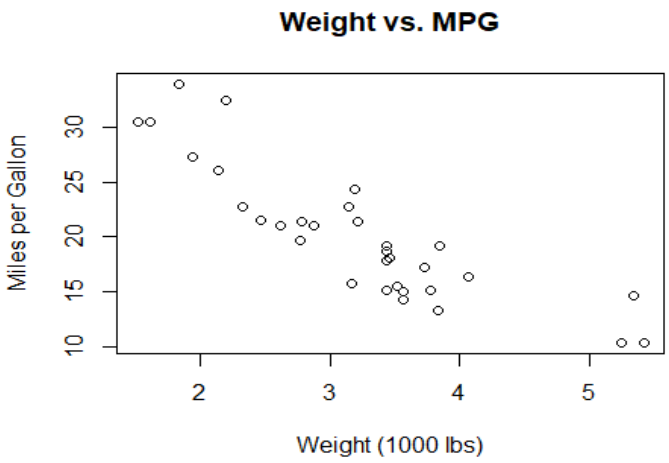
```r
data(mtcars)
head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Step 2: Explore the Data

Let's visualize the relationship between `wt` and `mpg`:

```r
plot(mtcars$wt, mtcars$mpg, main = "Weight vs. MPG", xlab = "Weight (10
00 lbs)", ylab = "Miles per Gallon")
```



The plot may show a non-linear relationship.

Step 3: Apply Natural Logarithm Transformation

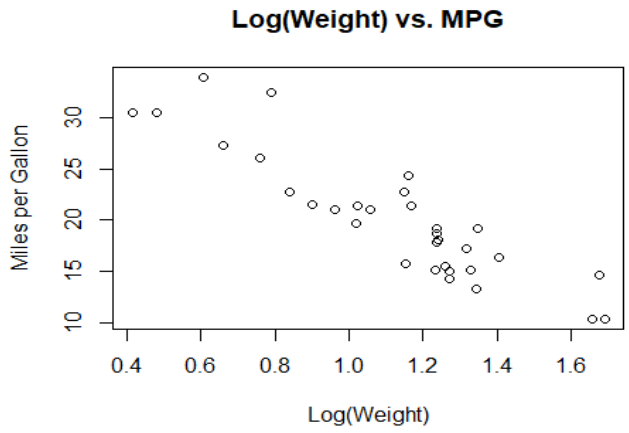
Transform the `wt` variable using the `log()` function:

```
mtcars$log_wt <- log(mtcars$wt)
```

Step 4: Visualize the Transformed Data

Plot the transformed variable against `mpg`:

```
plot(mtcars$log_wt, mtcars$mpg, main = "Log(Weight) vs. MPG", xlab = "Log(Weight)", ylab = "Miles per Gallon")
```



The relationship should appear more linear.

Step 5: Fit a Linear Model

Fit a linear regression model using the transformed predictor:

```r
# Fit linear regression model with original wt variable
model1 <- lm(mpg ~ wt, data = mtcars)

# Fit linear regression model with log-transformed wt
model2 <- lm(mpg ~ log_wt, data = mtcars)

# Summary of the model with original wt
summary(model1)

# Summary of the model with log-transformed wt
summary(model2)
```

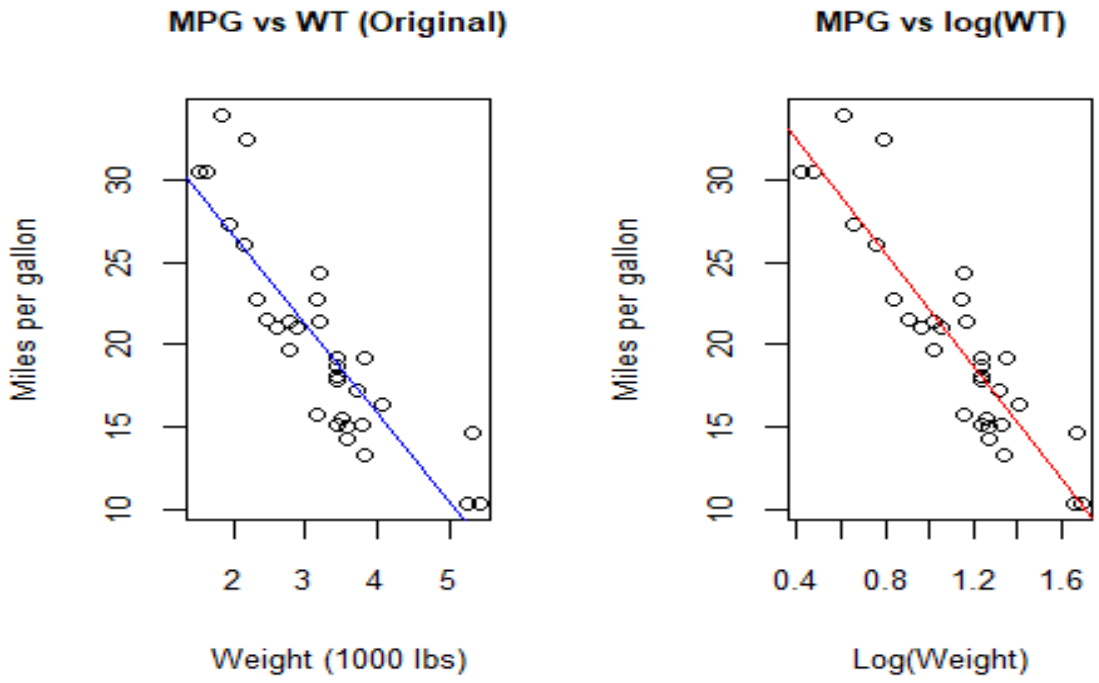
1.3.2 Comparison

The output will show the coefficients and statistical significance of the model.

- **R-squared** increased from 0.7528 to 0.7747 after applying the log transformation to **wt**, meaning the log-transformed model explains slightly more of the variation in **mpg**.
- **Standard Error** decreased from 3.03 to 2.87, indicating that the log-transformed model has more precise estimates.
- The **p-value** for both predictors (**wt** and **logwt**) is very small, indicating that both predictors are statistically significant. However, the p-value for the log-transformed predictor is even smaller, suggesting that the relationship between the log-transformed predictor and **mpg** is more statistically significant.

Now, let's plot the fitted values for both models to see how they compare visually:

```r
# Plotting the fitted values for comparison
par(mfrow=c(1,2)) # Set up the plotting area to have 2 plots side by si
de

plot(mtcars$wt, mtcars$mpg, main="MPG vs WT (Original)",
     xlab="Weight (1000 lbs)", ylab="Miles per gallon",
     cex.main=0.8, cex.lab=0.8, cex.axis=0.8)
abline(model1, col="blue") # Add regression line

# Plot for the model with log-transformed wt
plot(mtcars$log_wt, mtcars$mpg, main="MPG vs log(WT)",
     xlab="Log(Weight)", ylab="Miles per gallon",
     cex.main=0.8, cex.lab=0.8, cex.axis=0.8)
abline(model2, col="red") # Add regression line
```



By applying the natural logarithm transformation to the **wt** variable, we improved the model fit slightly (higher R-squared) and reduced the standard error, leading to more precise estimates. The log transformation also made the relationship between **wt** and **mpg** more statistically significant, as indicated by the smaller p-value.

This demonstrates how transformations like the natural logarithm can improve the performance of a regression model, especially when dealing with predictors that have nonlinear relationships with the response variable.

1.3.3 Interpret the Results

The coefficient for `log_wt` represents the change in `mpg` for a one-unit increase in the natural logarithm of `wt`. For example, if the coefficient is `-5`, it means that a 1% increase in weight is associated with a 5-unit decrease in miles per gallon (assuming the relationship is linear on the log scale).

1.3.4 Advantages of Natural Logarithm Transformation

1. **Reduces Skewness**: Makes the distribution of the predictor more symmetric.
2. **Stabilizes Variance**: Helps address heteroscedasticity.
3. **Improves Model Fit**: Can lead to better-fitting models when the relationship is multiplicative.

1.3.5 Limitations

1. **Zero or Negative Values**: The natural logarithm is undefined for zero or negative values. If your data contains such values, you may need to add a constant (e.g., `log(x + 1)`) before applying the transformation.
2. **Interpretability**: The transformed variable may be harder to interpret, especially for non-technical audiences.

Check Your Progress – 1

1. Refer to the TVADS dataset as given below, where the variable *Impress* measures the total number of times people remembered or were exposed to the commercials, quantified in millions. The variable *Spend* indicates the corresponding TV advertising budget, measured in millions of dollars.

Spend	*Impress*	*Spend*	*Impress*	*Spend*	*Impress*
49.7	30.2	5	12	26.9	38
50.1	32.1	19.3	11.7	26.9	50.7
20.4	21.4	40.1	78.6	6.1	4.4
74.1	99.6	166.2	40.1	185.9	98.8
32.4	71.1	82.4	60.8	45.6	10.4
7.6	12.3	9.2	21.4	27	40.8
22.9	21.9	5.7	10	154.9	98.9

   (a) Fit a linear regression model considering *Impress* depends on *Spend.* By applying natural logarithm $log_e(Spend)$, analyze the resulting differences and derive your conclusion.

1.4 Polynomial Transformation for Predictors

Polynomial transformations are commonly used in regression modeling to capture non-linear relationships between predictor variables and the response variable. While standard linear regression assumes a straight-line relationship between predictors and the response, polynomial regression allows for curvature by including higher-degree terms of the predictor variables.

1.4.1 Polynomial Regression Model

A general polynomial regression model of degree k can be written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_k X^k + \varepsilon$$

where:

- Y is the response variable,
- X is the predictor variable,
- $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients,
- ε is the error term.

The model is linear in the coefficients β even though it includes polynomial terms of X. This allows the model to be estimated using ordinary least squares (OLS) regression.

1.4.2 Simple Polynomial Regression

A second-degree (quadratic) polynomial model is the simplest form of polynomial regression:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

For example, suppose we have a dataset with:

- `Price` (response variable) - `Age` (predictor variable)

A simple linear regression model:

$$E(Price) = \beta_0 + \beta_1 Age$$

may not capture the relationship well if it is non-linear. Instead, using a quadratic model:

$$E(Price) = \beta_0 + \beta_1 Age + \beta_2 Age^2$$

allows for curvature in the relationship.

1.4.3 Multiple Polynomial Regression

Polynomial transformation can also be applied in multiple linear regression when multiple predictors influence the response. A general multiple polynomial regression model is given by:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_{ij} X_i X_j + \varepsilon$$

For a model with two predictors, X_1 and X_2, a second-degree polynomial model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

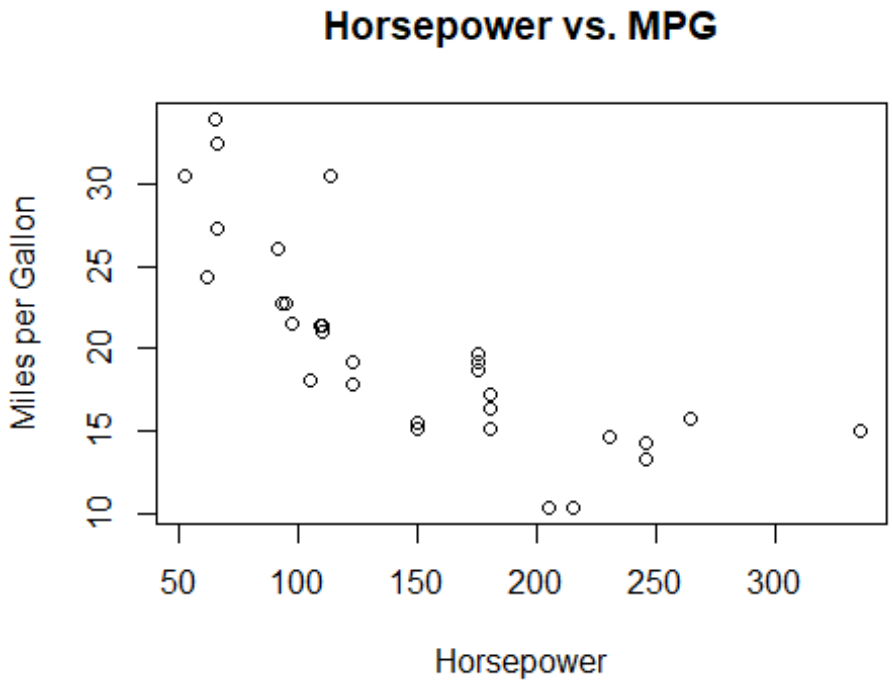
Importance of Hierarchy

When using polynomial transformations, hierarchy should be preserved, meaning that if a higher-degree term X^2 is included, its lower-degree term X should also be included in the model.

1.4.4 Example in R

Let's use the `mtcars` dataset to demonstrate polynomial transformation. We'll model the relationship between a car's horsepower (`hp`) and its miles per gallon (`mpg`). We'll apply a quadratic transformation to `hp` to capture non-linearity.

```
# Load data
data(mtcars)
# Fit linear model (before transformation)
linear_model <- lm(mpg ~ hp, data = mtcars)
# Explore data
plot(mtcars$hp, mtcars$mpg, main = "Horsepower vs. MPG", xlab = "Horsepower",
ylab = "Miles per Gallon")
```



The plot shows a curved relationship. Hence, we create a quadratic term for `hp`, and fit a linear regression model with the original `hp` and its squared term:

```
# Apply polynomial transformation
mtcars$hp_squared <- mtcars$hp^2
```

```r
# Fit polynomial model (after transformation)- original hp + squared term
poly_model <- lm(mpg ~ hp + hp_squared, data = mtcars)
```

The following function **`print_model_summary`** is designed to provide a detailed summary of a fitted regression model in a structured and human-readable format. It is particularly useful for quickly reviewing key model statistics and parameter estimates from linear or polynomial regression models.

```r
print_model_summary <- function(model, model_name) {
  # Summary of the model
  model_summary <- summary(model)

  # Extracting key information
  r_squared <- model_summary$r.squared
  adj_r_squared <- model_summary$adj.r.squared
  std_error <- model_summary$sigma

  # Get the coefficients (parameters) table
  coef_table <- as.data.frame(model_summary$coefficients)

  # Displaying Model Summary
  cat(paste0(model_name, " Model Summary\n"))
  cat("--------------------------------------------------------\n")
  cat(sprintf("Sample size: %d\n", nrow(mtcars)))
  cat(sprintf("R-squared: %.4f\n", r_squared))
  cat(sprintf("Adjusted R-squared: %.4f\n", adj_r_squared))
  cat(sprintf("Standard error: %.2f\n", std_error))

  cat("\nParameters\n")
  cat("--------------------------------------------------------\n")

  # Formatting the parameters table nicely
  cat(sprintf("%-12s %-12s %-12s %-12s %-12s\n", "Model", "Estimate", "Std Er
ror", "t-Statistic", "Pr(> |t|)"))

  # Loop through each coefficient row and print it
  for (i in 1:nrow(coef_table)) {
    cat(sprintf("%-12s %-12.3f %-12.3f %-12.3f %-12.3f\n",
                rownames(coef_table)[i], coef_table[i, 1], coef_table[i, 2],
coef_table[i, 3], coef_table[i, 4]))
  }
```

```
  cat("\n")
}

# Calling the function for both models
print_model_summary(linear_model, "Linear")

print_model_summary(poly_model, "Polynomial")
```

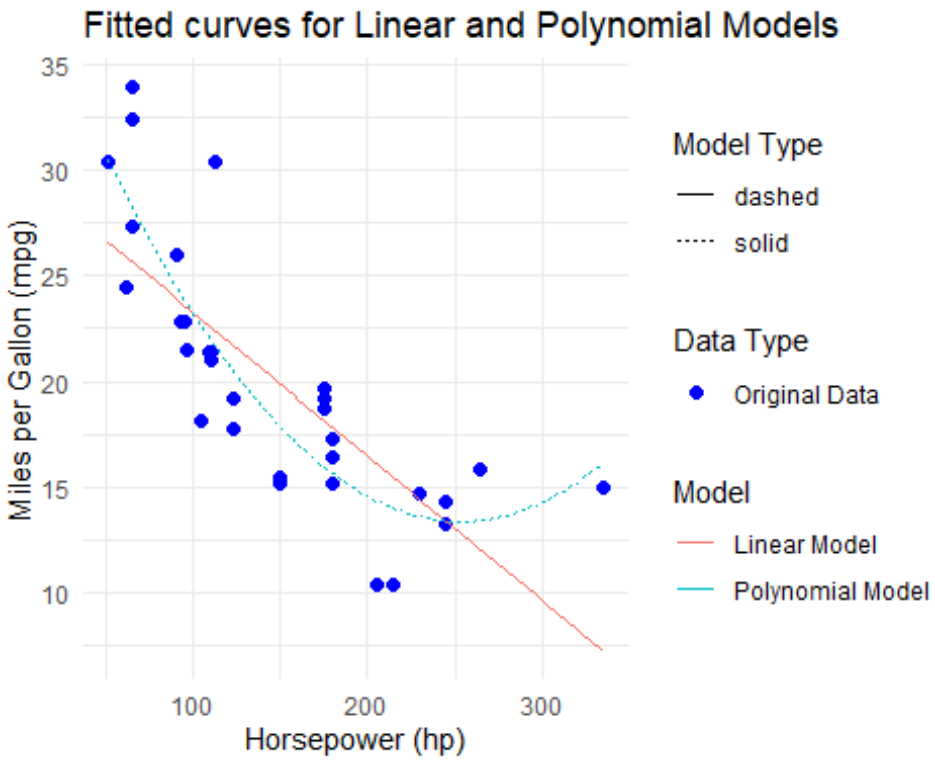
```
                  Linear Model Summary
 ------------------------------------------------------------
 Sample size: 32
 R-squared: 0.6024
 Adjusted R-squared: 0.5892
 Standard error: 3.86

                       Parameters
 ------------------------------------------------------------
 Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
 (Intercept)  30.099       1.634        18.421       0.000
 hp           -0.068       0.010        -6.742       0.000


                  Polynomial Model Summary
 ------------------------------------------------------------
 Sample size: 32
 R-squared: 0.7561
 Adjusted R-squared: 0.7393
 Standard error: 3.08

                       Parameters
 ------------------------------------------------------------
 Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
 (Intercept)  40.409       2.741        14.744       0.000
 hp           -0.213       0.035        -6.115       0.000
 hp_squared   0.000        0.000        4.275        0.000
```

By calling `print_model_summary(linear_model, "Linear")` and `print_model_summary(poly_model, "Polynomial")`, you can quickly compare the key statistical information between different models in a consistent format. By comparing the results, we should be pleased with the fit provided by this polynomial regression model.

The following code is designed to compare two different regression models—one linear and one polynomial—by plotting them on the same graph. It plots the original data points in blue

and adds the prediction lines from both models on the same graph—solid red for the linear model and dashed cyan for the polynomial model. The plot includes a legend to distinguish between the models. This visualization helps compare the fit of the two models, allowing you to see how well each model captures the relationship between horsepower and miles per gallon.

```
# Load the ggplot2 package
library(ggplot2)

# Create a new data frame with hp values for prediction
hp_values <- data.frame(hp = seq(min(mtcars$hp), max(mtcars$hp), length.out =
100))
hp_values$hp_squared <- hp_values$hp^2

# Predict mpg values using both models
hp_values$linear_pred <- predict(linear_model, newdata = hp_values)
hp_values$poly_pred <- predict(poly_model, newdata = hp_values)

# Plot the original data and the model predictions
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point(color = 'blue', size = 2, aes(shape = "Original Data")) +
  geom_line(data = hp_values, aes(x = hp, y = linear_pred, color = "Linear Mo
del", linetype = "dashed")) +
  geom_line(data = hp_values, aes(x = hp, y = poly_pred, color = "Polynomial
Model", linetype = "solid")) +
  labs(title = "Fitted curves for Linear and Polynomial Models",
       x = "Horsepower (hp)",
       y = "Miles per Gallon (mpg)",
       color = "Model",
       shape = "Data Type",
       linetype = "Model Type"
       ) +
  theme_minimal()
```



The coefficients for `hp` and `hp_squared` indicate the nature of the relationship. A significant coefficient for `hp_squared` suggests that the relationship between `hp` and `mpg` is non-linear.

Advantages of Polynomial Transformation

1. **Captures Non-Linearity**: Allows modeling of curved relationships.
2. **Flexibility**: Can model complex interactions between predictors.
3. **Improves Model Fit**: Often leads to better-fitting models when the relationship is non-linear.

Limitations

1. **Overfitting**: High-degree polynomials can overfit the data, especially with small datasets.
2. **Interpretability**: Higher-order terms can make the model harder to interpret.
3. **Extrapolation**: Polynomial models may perform poorly outside the range of the training data.

Polynomial transformations are valuable techniques for modeling non-linear relationships between predictors and response variables. By incorporating polynomial terms in R, you can capture curvature and enhance the performance of your models. However, care must be taken to avoid overfitting and ensure interpretability.

Check Your Progress – 2

1. The manager at ABC Co. is examining the connection between the tenure of their sales staff and the quantity of electronic items sold. The table below displays the number of electronic items sold by 15 randomly selected salespeople during the latest sales period and tenure of each salesperson in the company.

Sales	Tenure
162	22
112	12
189	40
275	41
83	9
325	56
296	106
67	6
308	111
150	12
376	104
367	85
189	19
235	51
317	76

(a) Visualize the relationship by creating a scatter plot.
(b) Apply a polynomial transformation to the tenure data and generate the regression model.
(c) Compare the models (using R-squared, Standard Error, and significance) with and without the polynomial transformation.
(d) Visualize the predicted values for both models on a plot.

1.5 Reciprocal Transformation for Predictors in Regression

Reciprocal transformation, also known as the multiplicative inverse transformation, is a technique applied to predictor variables in regression models to address nonlinear relationships between predictors and the response variable. Reciprocal transformations involve transforming a predictor variable X by taking its reciprocal, i.e., $X' = \frac{1}{X}$. This technique is particularly useful when the relationship between the response variable and predictor is nonlinear, often indicating a diminishing effect as the predictor increases.

In simple linear regression, we model the relationship between a response variable Y and a single predictor variable X as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

However, if we observe a nonlinear inverse relationship between Y and X, we can apply a reciprocal transformation $(1/X)$ to the predictor:

$$Y = \beta_0 + \beta_1 \left(\frac{1}{X}\right) + \varepsilon$$

This transformation can linearize relationships where the response decreases as the predictor increases.

In the context of multiple linear regression, consider three predictors X_1, X_2, X_3:

$$Y = \beta_0 + \beta_1 \frac{1}{X_1} + \beta_2 \frac{1}{X_2} + \beta_3 X_3 + \varepsilon$$

Here, the reciprocal transformation is applied to X_1 and X_2 since the relationships between Y and these predictors might exhibit a nonlinear, inverse pattern. This model can help stabilize variance and achieve linearity.

1.5.1 Application with `mtcars` Dataset

Let's illustrate this with the `mtcars` dataset in R. We'll use `mpg` (miles per gallon) as the response variable and `hp` (horsepower) as the predictor.

- **Model 1 (without transformation):**

$$mpg = \beta_0 + \beta_1 hp + \varepsilon$$

- **Model 2 (with reciprocal transformation):**

$$mpg = \beta_0 + \beta_1 \frac{1}{hp} + \varepsilon$$

Fitting these models in R and using `print_model_summary()` function, as defined in the previous section, we find:

```
linear_model <- lm(mpg ~ hp, data = mtcars)
inverse_model <- lm(mpg ~ I(1/hp), data = mtcars)
print_model_summary(inverse_model, "Reciprocal")
```

```
                  Reciprocal Model Summary
-----------------------------------------------------------
Sample size: 32
R-squared: 0.7381
Adjusted R-squared: 0.7294
Standard error: 3.14

                        Parameters
-----------------------------------------------------------
Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
(Intercept)  9.434        1.285        7.344        0.000
I(1/hp)      1259.881     137.016      9.195        0.000
```

Reciprocal transformations can be a powerful tool in regression modeling, especially when an inverse relationship is expected. By transforming predictors, we can often achieve better model performance, interpretability, and predictive accuracy.

Check Your Progress – 3

With the `mtcars` dataset in R, use `mpg` (miles per gallon) as the response variable and `hp` (horsepower) as the predictor.

(a) Visualize the predicted values for linear and reciprocal models on a plot.

1.6 Comparison of Model Performance

In this section, we compare the performance of three models—**Linear**, **Reciprocal**, and **Polynomial**—to determine which best explains the relationship between the predictor variable (`hp`) and the response variable (`mpg`). The evaluation is based on key statistical metrics, including **R-squared**, **Adjusted R-squared**, **standard error**, and the significance of model parameters. The results are summarized in Table 1.1

Table 1.1: Models Comparison

Metric	Linear Model	Reciprocal Model	Polynomial Model
Sample size	32	32	32
R-squared	0.6024	0.7381	0.7561
Adjusted R-squared	0.5892	0.7294	0.7393
Standard error	3.86	3.14	3.08
Intercept (Estimate)	30.099	9.434	40.409
hp (Estimate)	-0.068	-	-0.213
I(1/hp) (Estimate)	-	1259.881	-
hp_squared (Estimate)	-	-	0.000

1.6.1 Model Performance Overview

1. **Linear Model**: The Linear Model serves as the baseline for comparison. It achieves an R-squared value of **0.6024** and an Adjusted R-squared value of **0.5892**, indicating that approximately 60.2% of the variability in the response variable is explained by the linear relationship with `hp`. The standard error of **3.86** suggests moderate precision in predictions. The model parameters are statistically significant ($p < 0.001$), with the intercept at **30.099** and the coefficient for `hp` at **-0.068**, indicating a negative linear relationship.

2. **Reciprocal Model**: The Reciprocal Model improves upon the Linear Model, achieving a higher R-squared value of **0.7381** and an Adjusted R-squared value of **0.7294**. This suggests that approximately 73.8% of the variability in the response variable is explained by the reciprocal relationship with `hp`. The standard error decreases to **3.14**, indicating better precision compared to the Linear Model. The model parameters are also statistically significant ($p < 0.001$), with the intercept at **9.434** and the coefficient for `1/hp` at **1259.881**, reflecting an inverse relationship between `hp` and the response variable.

3. **Polynomial Model**: The Polynomial Model outperforms both the Linear and Reciprocal Models, achieving the highest R-squared value of **0.7561** and an Adjusted R-squared value of **0.7393**. This indicates that approximately 75.6% of the variability in the response variable is explained by the polynomial relationship with `hp`. The standard error is the lowest among the three models at **3.08**, demonstrating the highest precision in predictions. The model parameters are statistically significant ($p < 0.001$), with the intercept at **40.409**, the linear term for `hp` at **-0.213**, and the quadratic term (`hp_squared`) at **0.000**. The negative coefficient for `hp` and the positive coefficient for `hp_squared` suggest a curvilinear relationship.

1.6.2 Comparison and Conclusion

- **R-squared and Adjusted R-squared**: The Polynomial Model has the highest R-squared and Adjusted R-squared values, indicating it explains the most variability in

the data. The Reciprocal Model performs better than the Linear Model but falls short of the Polynomial Model.

- **Standard Error**: The Polynomial Model has the lowest standard error, making it the most precise. The Reciprocal Model improves upon the Linear Model but is less precise than the Polynomial Model.
- **Model Fit**: The Polynomial Model's inclusion of both linear and quadratic terms allows it to capture a more complex, curvilinear relationship between `hp` and the response variable (`mpg`), which the Linear and Reciprocal Models cannot fully represent.

In conclusion, the **Polynomial Model** is the best-performing model for this dataset, offering the highest explanatory power and precision. The Reciprocal Model is a reasonable alternative, outperforming the Linear Model but not matching the Polynomial Model's effectiveness. The Linear Model, while statistically significant, is the least effective in capturing the relationship between `hp` and the response variable `mpg`.

Check Your Progress – 4

With the `mtcars` dataset in R, use `mpg` (miles per gallon) as the response variable and car's weight (`wt`) as the predictor.

(a) Apply a polynomial transformation to predictor and generate the regression model.
(b) Compare the models (Linear, Polynomial and Logarithm)
(c) Visualize the predicted values for all models on a plot.

1.7 LET US SUM UP

This unit focuses on variable transformations as a powerful tool to improve the performance of multiple linear regression models. By applying transformations such as natural logarithms, polynomials, and reciprocals, you can address issues like skewness, nonlinearity, and nonnormality in predictor variables. The choice of transformation depends on the nature of the data and the relationship between the predictor and response variables. Additionally, the unit emphasizes the importance of preserving hierarchy in polynomial models and provides a comprehensive framework for evaluating model performance using key statistical metrics. Applying these techniques and properly utilizing transformations enable the construction of more accurate and interpretable regression models. This results in better-fitting models, improved interpretability, and more accurate predictions.

1.8 Check Your Progress: Possible Answers

Check Your Progress – 1

```
# Load the olsrr package
library(olsrr)

# Creating the data frame
brand_data <- data.frame(
  Spend = c(49.7, 50.1, 20.4, 74.1, 32.4, 7.6, 22.9, 5, 19.3, 40.1, 166.2,
82.4, 9.2, 5.7, 26.9, 26.9, 6.1, 185.9, 45.6, 27, 154.9),
  Impress = c(30.2, 32.1, 21.4, 99.6, 71.1, 12.3, 21.9, 12, 11.7, 78.6, 40
.1, 60.8, 21.4, 10, 38, 50.7, 4.4, 98.8, 10.4, 40.8, 98.9)
)

# Fitting the linear regression model
linear_model <- lm(Impress ~ Spend, data = brand_data)

# Obtaining the results using olsrr
ols_regress(linear_model)

# Scatter plot
plot(brand_data$Spend, brand_data$Impress,
     xlab = "Spend", ylab = "Impress",
     pch = 19, col = "blue")
abline(linear_model, col = "red") # Add regression line

# Residual plot
ols_plot_resid_fit(linear_model)

# Applying natural logarithm to the Budget variable
brand_data$logSpend <- log(brand_data$Spend)

# Fitting the linear regression model with the transformed variable
log_model <- lm(Impress ~ logSpend, data = brand_data)

# Obtaining the results using olsrr
ols_regress(log_model)

# Residual plot
ols_plot_resid_fit(log_model)
```

1.9 Further Reading

1. Applied Regression Modeling 3rd Edition, IAIN PARDOE**,** John Wiley & Sons, Inc, December 2020
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

1.10 Assignment

1. What are the advantages and limitations of using logarithmic transformations? When should they be avoided?
2. How do polynomial transformations help capture non-linear relationships in regression models?
3. What is the principle of preserving hierarchy in polynomial regression, and why is it important?
4. In what scenarios would a reciprocal transformation be appropriate for a predictor variable?
5. How does reciprocal transformation affect the interpretation of the regression coefficient?
6. Consider modeling `mpg` (miles per gallon) with weight (wt) and horsepower (hp):

$$\text{mpg} = \beta_0 + \beta_1 \left(\frac{1}{\text{wt}}\right) + \beta_2 \text{hp} + \varepsilon$$

   Examine the model could reveal hidden patterns, offering better predictions of fuel efficiency.

Unit 2 Advanced Transformations

Unit Structure

2.0 Learning Objectives

After going through this unit, you should be able to

- Apply the natural logarithm transformation to the response variable in multiple linear regression models where a unit change in a predictor results in a proportional change in the response.
- To interpret the transformed model in terms of elasticities and proportional relationships between predictors and the response variable.
- To select and apply transformations (e.g., log, square root, inverse) to both response and predictor variables to improve model fit.
- To perform the Box-Cox transformation, interpret the lambda parameter, and select an appropriate transformation based on the output.
- To understand the concept of scaling predictors (e.g., centering, standardization, and normalization) and its importance in regression analysis.
- To have knowledge about tools for identifying predictor transformations.

2.1 Introduction

In the previous unit, we explored how the general linear model can be employed to represent various potential relationships between predictors and the response variable. We concentrated on transformations involving one or more of the predictor variables. However, it is often beneficial to consider transformations involving the dependent variable as well. These transformations can be particularly useful in addressing issues such as non-linearity, heteroscedasticity (non-constant variance), and skewed distributions, which can impact the model's assumptions and the accuracy of predictions.

This unit will focus on the application of transformations to both the response and predictor variables in regression analysis. We will discuss various techniques for transforming the response or dependent variable, including the natural logarithm transformation and the Box-Cox method, which can stabilize variance and linearize relationships. Additionally, we will explore methods for scaling predictors and interpreting the resulting regression coefficients.

By the end of this unit, you will not only understand when and why to apply transformations but also gain practical knowledge on how to implement these transformations using diagnostic tools and interpret their effects on model performance. You will learn to apply these techniques to improve the fit and predictive power of regression models, ensuring that the assumptions underlying the general linear model are satisfied.

2.2 Natural Logarithm Transformation for the Response

Natural logarithm transformation is a widely used technique in regression modeling. It is particularly effective when the response variable exhibits a nonlinear relationship with predictors, heteroscedasticity, or a positively skewed distribution. By applying a logarithm transformation, we can often improve the interpretability and performance of a regression model. It is particularly useful when dealing with financial, economic, and social science data where changes occur proportionally rather than absolutely.

The transformation is applied to the response variable to address the following issues:

- **Nonlinear Relationships**: Some relationships between predictors and response variables are multiplicative rather than additive.
- **Heteroscedasticity**: When the variance of residuals increases with the response variable, transforming the response can help stabilize variance.
- **Skewed Distributions**: Positively skewed data can be made more symmetric, improving the normality assumption required for many regression techniques.

2.2.1 Example: MPG vs. Weight (mtcars Dataset in R)

A suitable example for illustrating the log transformation is the `mtcars` dataset in R, where we analyze the relationship between miles-per-gallon rating (`mpg`) and the weight of the car (`wt`).

Model 1: Using the Untransformed Response Variable

$$E(mpg) = \beta_0 + \beta_1 wt$$

This model may not fully capture the true relationship, as heavier cars tend to have a diminishing effect on fuel efficiency.

Model 2: Using the Log-Transformed Response Variable

$$E\left(\log_e(mpg)\right) = \beta_0 + \beta_1 wt$$

This model provides:

- A better representation of the nonlinear relationship between weight and fuel efficiency.
- More consistent residual variance, satisfying regression assumptions.

```r
# Load the mtcars dataset
data(mtcars)

# Fit linear regression model with original mpg variable
linear_model  <- lm(mpg ~ wt, data = mtcars)

# Log transformation
log_mpg <- log(mtcars$mpg)

# Fit linear regression model with log-transformed mpg
log_model <- lm(log_mpg ~ wt, data = mtcars)

# Calling the function for both models
print_model_summary(linear_model, "Linear")

 print_model_summary(log_model, "Log Transformed")
```

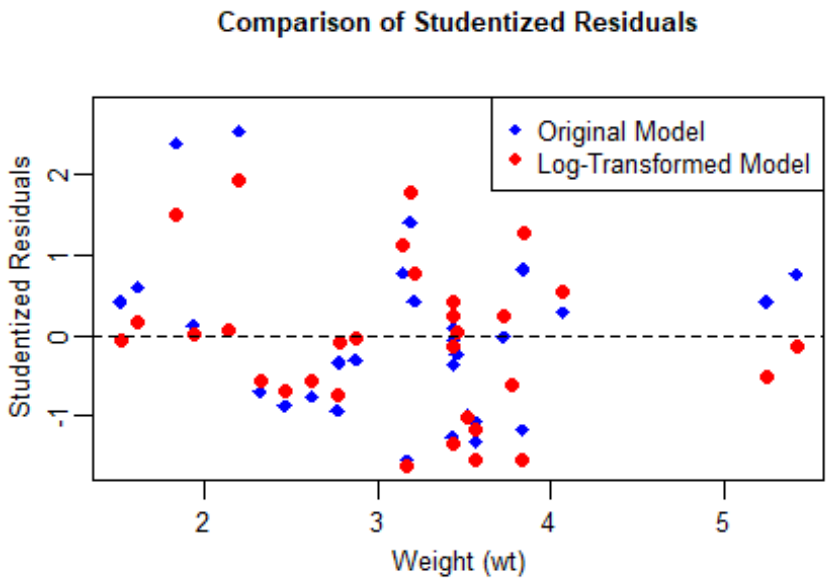
```r
   # Calculate Studentized residuals for the original model
   studentized_residuals_original <- rstudent(linear_model)

   # Calculate Studentized residuals for the log-transformed model
   studentized_residuals_log <- rstudent(log_model)

   # Create an overlaid residual plot with customizations
   plot(mtcars$wt, studentized_residuals_original,
        col = "blue", pch = 18, # pch = 18 for diamond
        main = "Comparison of Studentized Residuals",
        xlab = "Weight (wt)", ylab = "Studentized Residuals",
```

```r
     ylim = range(c(studentized_residuals_original,
studentized_residuals_log)),
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     mgp = c(1.5, 0.5, 0))  # Adjusts the gap for title and axis labels

# Add points for the log-transformed model
points(mtcars$wt, studentized_residuals_log, col = "red", pch = 19)

# Add a horizontal reference line at 0
abline(h = 0, col = "black", lty = 2)

# Add a legend
legend("topright", legend = c("Original Model", "Log-Transformed Model"),
       col = c("blue", "red"), pch = c(18, 19), cex = 0.8)
```

```
                          Linear Model Summary
-------------------------------------------------------------
 Sample size: 32
 R-squared: 0.7528
 Adjusted R-squared: 0.7446
 Standard error: 3.05
```

```
                                Parameters
-------------------------------------------------------------
 Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
 (Intercept)  37.285       1.878        19.858       0.000
 wt           -5.344       0.559        -9.559       0.000
```

```
                      Log Transformed Model Summary
-------------------------------------------------------------
 Sample size: 32
 R-squared: 0.7976
 Adjusted R-squared: 0.7908
 Standard error: 0.14

                     Parameters
 -------------------------------------------------------------
 Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
 (Intercept)  3.832        0.084        45.642       0.000
 wt           -0.272       0.025        -10.872      0.000
```



2.2.2 Interpretation of Coefficients in Log-Transformed Models

When the response variable is log-transformed, the interpretation of regression coefficients changes:

- The equation for the transformed model:

$$\log_e(Y) = \beta_0 + \beta_1 X + \varepsilon$$

- The coefficient β_1 represents the **proportional change** in Y for a one-unit change in X:

$$\text{Expected proportional change in } Y = e^{\beta_1} - 1$$

- Example: If $\beta_1 = -0.272$ in the `mtcars` dataset, then:

$$e^{-0.272} - 1 = -0.2381 \approx -24\%$$

  This means that for every additional unit increase in `wt`, `mpg` decreases by approximately 24%.

2.2.3 When to Use Logarithm Transformation?

The log transformation is beneficial when:

- The response variable is **positively skewed**.
- The **variance of residuals increases** with larger values of the response.
- The relationship between predictors and the response is **multiplicative** (e.g., percentage change rather than absolute change).

2.2.4 Addressing Negative or Zero Values

- Since the logarithm of zero or negative values is undefined, a small constant (e.g., $Y + 1$) can be added before transformation.
- This ensures that the transformation remains valid and that predicted values remain positive, which is crucial for variables like income, sales, and prices.

Check Your Progress – 1

1. Consider the following dataset (simulated).

Marketing Expenditure (in ₹1000s)	Sales Revenue (in ₹1000s)	Marketing Expenditure (in ₹1000s)	Sales Revenue (in ₹1000s)	Marketing Expenditure (in ₹1000s)	Sales Revenue (in ₹1000s)
10	50	60	105	110	155
15	60	65	110	115	160
20	65	70	115	120	165
25	70	75	120	125	170
30	75	80	125	130	175
35	80	85	130	135	180
40	85	90	135	140	185
45	90	95	140	145	190
50	95	100	145	150	195
55	100	105	150	110	155

Using R code:

(a) Crete linear model $E(Revenue) = \beta_0 + \beta_1 Expenditure$

(b) Crete transformed model $E(\log_e Revenue) = \beta_0 + \beta_1 Expenditure$

(c) What are your conclusions with respect to comparing models, and how can you describe them in a few paragraphs?

2.3 Square Root Transformations for the Response

An alternative method to address issues with nonconstant variance is to apply a square root transformation. In this transformation, the response variable y is replaced by its square root, $\sqrt{y}$. The square root transformation is useful for compressing high values more mildly than a logarithm.

2.3.1 Example: Income and Square Root Transformation

In many cases, the relationship between income and other variables (like years of experience or education level) is not linear. For example, the difference in earnings between $0 and $10,000 is much larger than the difference between $80,000 and $90,000, even though both represent a $10,000 change. This suggests that the effect of each additional dollar of income might be perceived differently at different income levels.

Raw-Scale Linear Model:

- In a **raw-scale linear model**, we assume that each additional year of experience or each additional unit of an independent variable increases income by the same fixed amount. However, this assumption may not hold in real life. For instance, the impact of education on income might be larger when someone's income is lower, but less impactful as their income increases. This model doesn't account for the diminishing marginal effect of income.

Logarithmic Transformation:

- A **logarithmic transformation** would model income in terms of percentage changes, which makes it more appropriate for cases where proportional differences matter more than absolute differences. For example, the difference between earning $20,000 and $40,000 is considered the same as the difference between earning $80,000 and $160,000 (both represent a 100% increase). While this might work in some cases, it might be too severe because it tends to focus on large relative changes, losing the sense of absolute income differences.

Square Root Transformation:

- A square root transformation balances the extremes of raw and logarithmic transformations. By transforming the income variable to $\sqrt{\text{income}}$, we smooth out the impact of large income values, while maintaining comparability across different income ranges.

  For example:

  - Differences in earnings between **$0 and $10,000** are much larger than between **$80,000 and $90,000** on the raw scale.
  - However, when we apply a square root transformation, the difference between **$0 to $10,000** is not overly exaggerated, and the difference between **$80,000 to $90,000** is made more comparable. This works because stepping up in earnings (such as from $0 to $10,000, $10,000 to $40,000, or $40,000 to $90,000) results in equal steps in square root earnings. For example:
    - $\sqrt{10{,}000} = 100$

- $\sqrt{40{,}000} = 200$
- $\sqrt{90{,}000} \approx 300$

So, the increases are proportionate on the square root scale, meaning the same magnitude of change in the transformed variable corresponds to more comparable effects, regardless of whether the income is low or high.

In summary, by applying the square root transformation, we preserve the relative differences between income levels while smoothing out the impact of large values, making the data easier to analyze and interpret. This transformation can be useful when modeling income, where the effect of each additional dollar decreases as the income level increases.

2.3.2 Limitations:

- The interpretation of coefficients is less straightforward than in the original or log-transformed models.
- Negative predictions become large positive values when squared, introducing nonmonotonicity.
- More suitable for prediction tasks rather than explanatory modeling.

Another approach to addressing issues with nonconstant variance is to use $\frac{1}{y}$ as the dependent variable instead of y. This transformation is called a *reciprocal transformation*.

For instance, if the response variable is measured in miles per gallon, applying the reciprocal transformation would result in a new response variable with units of $\frac{1}{\text{miles per gallon}}$, or gallons per mile.

2.4 Transformations for the Response and Predictors

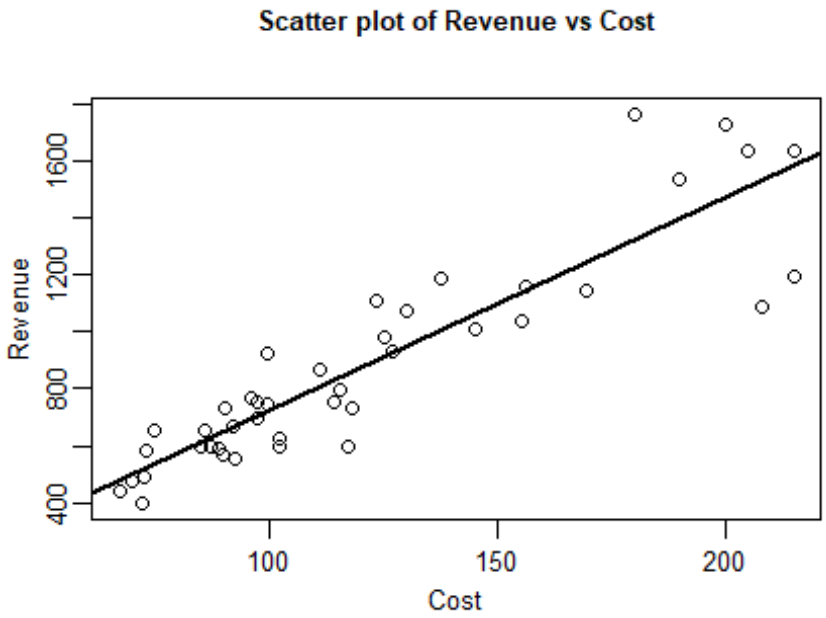
Regression models often require transformations to improve model fitting, correct heteroscedasticity, and address issues of non-linearity. By applying mathematical functions to the response variable and/or predictor variables, we can often improve the fit of the model and make it more interpretable.

For example, consider a simulated dataset where we want to model the relationship between **Revenue** (response) and **Cost** (predictor).

```
# Create dataset
data <- data.frame(
   Cost = c(205, 208, 215, 215, 199.9, 190, 180, 156, 144.9, 137.5, 127, 125,
123.5, 117, 118, 115.5, 111, 113.9, 99.5, 99.5, 97.5, 97.5, 90, 96, 86,
169.5, 155.3, 130, 102, 102, 92.2, 92.5, 89.9, 85, 89, 87, 70, 72, 74.9,
73.1, 72.5, 67),
  Revenue = c(1639, 1088, 1193, 1635, 1732, 1534, 1765, 1161, 1010, 1191,
```

```
930, 984, 1112, 600, 733, 794, 867, 750, 923, 743, 752, 696, 731, 768, 653,
1142, 1035, 1076, 626, 600, 668, 553, 566, 600, 591, 599, 477, 398, 656, 585,
490, 440)
)
```

```
# Fit linear model (before transformation)
linear_model <- lm(Revenue ~ Cost, data = data)
# scatter plot
plot(data$Cost, data$Revenue,
     pch = 1,
     main = "Scatter plot of Revenue vs Cost",
     xlab = "Cost", ylab = "Revenue",
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     mgp = c(1.5, 0.5, 0))  # Adjusts the gap for title and axis labels
# Add regression line
abline(linear_model, col = "black", lwd = 2)
```



From the above scatter plot, it can be observed that the data points are closer to the line at the left side of the plot (for lower values of Cost) than at the right side (for higher values of Cost). This suggests that the variance of the estimated errors increases from left to right, violating the constant variance assumption of the linear regression model.

To address the issue of increasing variance, we can apply a logarithmic transformation to both variables. The relationship is now represented as:

$$E\big(\log_e(\text{Revenue})\big) = \beta_0 + \beta_1 \log_e(\text{Cost})$$

and the results are presented in Table 2.1.

```r
# Log transformation
data$log_Revenue <- log(data$Revenue)
data$log_Cost <- log(data$Cost)
# Fit linear model (before transformation)
log_model <- lm(log_Revenue ~ log_Cost, data = data)

# scatter plot
plot(data$log_Cost, data$log_Revenue,
     pch = 1,
     main = expression("Scatter plot of " * log[e]("Revenue") * " vs " *
log[e]("Cost")),
     xlab = expression(log[e]("Cost")), ylab = expression(log[e]("Revenue")),
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.8,
     mgp = c(1.5, 0.5, 0))  # Adjusts the gap for title and axis labels
# Add regression line
abline(log_model, col = "black", lwd = 2)
```

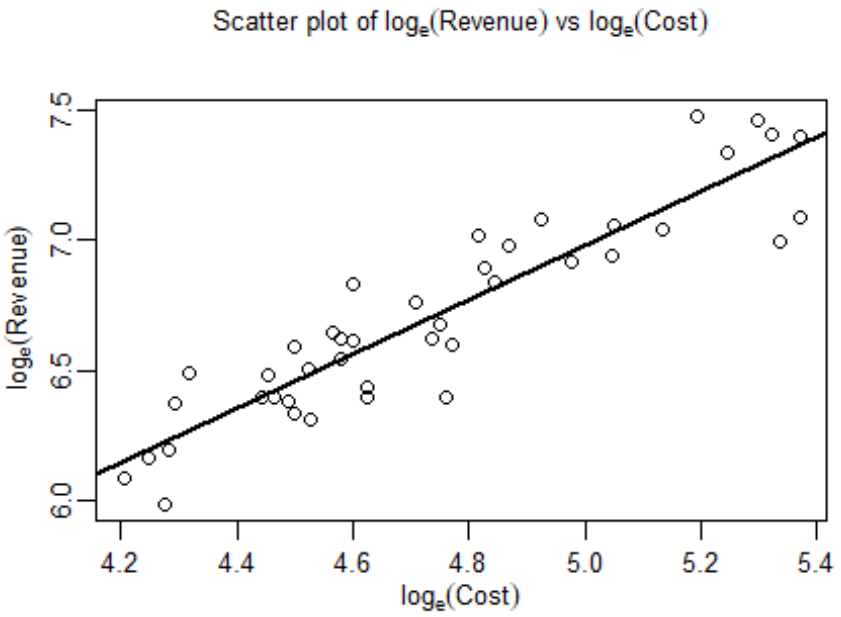


Table 2.1: Models Comparison

Metric	Linear Model	Log Model
Sample size	42	42
R-squared	0.8123	0.8428
Adjusted R-squared	0.8076	0.8388
Standard error	159.30	0.15
Intercept (Estimate)	-20.866	1.776

Metric	Linear Model	Log Model
Cost (Estimate)	7.455	-
Log_Cost (Estimate)	-	1.041

It can be observed from Table 2.1 that the log-transformed model has a significantly lower standard error compared to the linear model (0.15 vs. 159.30), suggesting that the log-transformed model provides more accurate and reliable predictions of Revenue based on Cost. This substantial reduction in standard error further supports the appropriateness of using the log-transformed model over the linear model. It not only fits the data better in terms of R-squared values but also enhances the precision of predictions.
This model can be used to estimate Revenue for a given Cost. For example, if a project has a Cost of $100,000, the expected Revenue would be:

$$\exp\left(\beta_0 + \beta_1 \log_e(100)\right)$$

Confidence intervals for the mean and prediction intervals for individual values can also be calculated and exponentiated to obtain intervals in the original scale.

2.4.1 Box-Cox Transformation

The **Box-Cox transformation** is a systematic method for selecting the best power transformation for the response variable. It seeks to find the power λ that makes the residuals from the regression model as close to normally distributed as possible. The transformation is defined as:

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log_e(Y) & \text{if } \lambda = 0. \end{cases}$$

The Box-Cox method can suggest transformations such as:

- $\lambda = 1$: No transformation.
- $\lambda = 0.5$: Square root transformation.
- $\lambda = 0$: Logarithmic transformation.
- $\lambda = -1$: Reciprocal transformation.

The selected λ is often rounded to a sensible value (e.g., 1.8 might be rounded to 2). Many statistical software packages include routines for performing Box-Cox transformations.

```
# Load necessary library
library(MASS)

# Box-Cox transformation (for response variable)
boxcox_model <- MASS::boxcox(linear_model, lambda = seq(-2, 2, by = 0.1))
# Add a title to the plot
title(main = "Box-Cox Transformation of Revenue vs Cost", cex.main = 0.8)
```

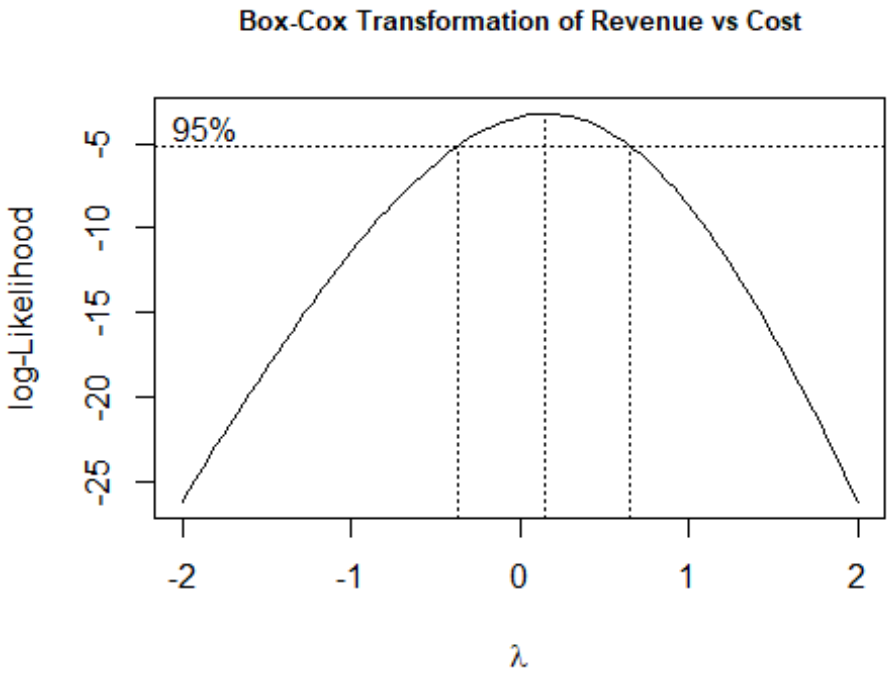
Here's an explanation of the plot generated by this code:

1. **Lambda (λ) Axis:** The x-axis of the Box-Cox plot represents the values of lambda (λ), which range from -2 to 2. The lambda value is the power to which all data will be raised. A lambda value of 1 corresponds to no transformation, while a lambda value of 0 corresponds to a natural logarithm transformation.

2. **Log-Likelihood Axis:** The y-axis represents the log-likelihood values. The log-likelihood measures the fit of the transformation. Higher values indicate a better fit to the assumptions of the linear regression model.

3. **Optimal Lambda:** The plot typically includes a curve showing the log-likelihood for different values of lambda. The peak of this curve indicates the optimal lambda value, where the log-likelihood is maximized. This is the value of lambda that best transforms the data to meet the assumptions of the linear model.

   In this plot, the peak is around $\lambda \approx 0$, indicating that the log transformation is the most appropriate. The log transformation stabilizes the variance and makes the distribution of the data more normal, improving the fit of the linear model.



4. **Confidence Intervals:** The plot may also include confidence intervals around the optimal lambda showing by dotted vertical lines. These intervals suggest the range of lambda values that provide a similarly good fit. A common practice is to choose a lambda within this range to ensure a robust transformation.

The confidence interval in the figure extends slightly falls below 0 and 1, meaning transformations near 0 (such as log) or slightly positive power transformations are reasonable. Since the confidence interval includes 0 and is centered around it, this suggests that the logarithmic transformation is both appropriate and effective for the given data. Additionally, values within this range (e.g., λ between approximately 0 and 1) indicate that minor deviations from the logarithmic transformation, such as the square root transformation (λ = 0.5), could still yield a reasonable model fit.

Check Your Progress – 2

1. The following dataset consider response variable being the percentage of the population that are Internet users (Int) and the predictor variable being GDP per capita in thousands of dollars (Gdp).

Gdp	Int	Gdp	Int	Gdp	Int	Gdp	Int	Gdp	Int
3.7	29.1	0.3	1.5	8.4	11.2	1	0.9	36.7	97
43.8	44.8	1.6	2.1	69.9	78.6	13.1	5.8	57	31.3
21.8	22.6	7.4	43.4	4.8	10.9	32.8	76.1	30	75.3
12.8	47.9	14.7	32.8	6.6	30.8	4.3	12.9	30.7	47.9
41.3	72.6	16.5	74	38.5	15.8	28.6	37	1.8	4.5
37.6	65.1	40.3	74.8	50.3	78.4	19	75.7	43	69.8
1	3.1	46.7	79.1	0.7	0.5	4.9	41.3	18	40.7
30.2	46.2	11	28.9	15.5	41.5	26.9	58.9	21	38.7
33.3	69.3	21.7	65.6	9.8	50.4	12.2	24.5	41.3	72.6
11.4	32.4	37.9	77.8	9.1	53.9	25.6	65.6	57.2	68.2

Using R code:

(a) Compare the following three models: (1) response Int and predictor Gdp; response $log_e(Int)$ and predictor $log_e(Gdp)$; (3) response $\sqrt{Int}$ and predictor $\sqrt{Gdp}$.

(b) What are your conclusions with respect to the various ways of comparing models, and how can you describe them in a few paragraphs? You may analyze scatter plots or residual plots for any patterns.

(c) Analyze Box-Cox plot and comment on the model selection.

2.4.2 Scaling of Predictors and Regression Coefficients

Regression models aim to quantify relationships between variables. However, the choice of measurement scales for predictors can significantly impact the interpretability of regression coefficients. Scaling predictors and regression coefficients are often necessary to ensure meaningful comparisons and interpretations.

Effects of Scaling on Regression Coefficients

- The coefficient of a predictor represents the expected change in the dependent variable for a one-unit increase in the predictor.
- If a predictor is measured in different units (e.g., inches vs. millimeters), the corresponding regression coefficient changes accordingly.
- Standardized scaling helps provide more interpretable coefficients by adjusting for variations in measurement units.

Types of Scaling

1. *Standardization using Z-Scores*

   Z-score standardization is a widely used technique in data preprocessing, wherein each predictor variable is transformed to achieve a mean of 0 and a standard deviation of 1. This is mathematically represented as:

$$z = \frac{x - \mu}{\sigma}$$

   Where:

   - x is the original value,
   - μ is the mean of the variable,
   - σ is the standard deviation of the variable.

   Advantages

   1. *Comparable Regression Coefficients:* Z-score standardization allows regression coefficients to represent changes in terms of standard deviations, facilitating easier comparison across different predictors.
   2. *Multicollinearity Mitigation:* This standardization helps address multicollinearity issues in regression models.
   3. *Uniform Scale:* It is particularly useful when dealing with variables that have different units and ranges, providing a uniform scale for analysis.

2. *Standardization Using an Externally Specified Population Distribution*

   In certain scenarios, standardization utilizes a predetermined external reference instead of sample statistics. This approach is prevalent in disciplines such as education and psychology.

   Process

- Establish standard values from an external dataset (e.g., national test score distribution).
- Adjust the predictor using the established mean and standard deviation from this dataset.

Advantages

1. *Ensuring Consistency:* It maintains consistency across various studies and datasets.
2. *Cross-Population Comparison:* Facilitates comparison across different sample populations.

Example

Consider a standardized test score that references national averages. For instance, if the national mean is 55 and the standard deviation is 18, scores are standardized using these values instead of those derived from the study sample.

3. *Standardization using Reasonable Scales*

Instead of strict z-score transformations, some variables are rescaled in more interpretable ways while preserving familiar units. This approach includes:

- Dividing income by 1,000 to express earnings in thousands.
- Measuring age in decades rather than years.
- Centering categorical variables around meaningful points.

Advantages

1. *Meaningful Interpretations:* Retains meaningful interpretations while improving numerical stability.
2. *Ease of Understanding:* Helps maintain ease of understanding without completely standardizing variables.

Example

Consider income as a numerical variable. Instead of expressing income in its raw form, it can be rescaled and expressed in thousands. For example:

$$income\_\text{scaled} = \frac{income}{1000}$$

In this case, an income of 50,000 would be expressed as `income_scaled` value of 50, making it easier to interpret in certain contexts while maintaining numerical stability.

2.5 Tools for Identifying Predictor Transformations

2.5.1 Partial Residual Plots

Partial residual plots (also known as component plus residual plots) are useful for identifying non-linear relationships between the response and individual predictors. These plots show the relationship between the response and a predictor after accounting for the effects of other predictors. If the plot shows a non-linear pattern, a transformation of the predictor may be needed.

2.5.2 Ceres Plots

Ceres plots are a generalization of partial residual plots that combine conditional expectations and residuals. They provide a more flexible way to visualize the relationship between the response and predictors, especially in the presence of interactions or non-linearities.

2.5.3 Box-Tidwell Method

The Box-Tidwell method is an automated approach for suggesting power transformations of predictor variables. It iteratively estimates the best power transformation for each predictor to improve the linearity of the relationship with the response.

2.6 LET US SUM UP

As with other types of transformations, such as square root or reciprocal transformations, there is no guarantee that a logarithmic transformation will always outperform the others. The effectiveness of each transformation depends on the specific characteristics of the predictors and/or the response variable. Therefore, it is essential to test different transformations to determine which one best stabilizes variance and improves model fit, whether applied to predictors and/or the response.

While transformations are a valuable tool in regression analysis, they should be applied thoughtfully and judiciously. Automated methods like the Box-Cox and Box-Tidwell transformations can suggest potential transformations, but these tools should not be solely relied upon. The choice of transformations should also incorporate theoretical and contextual knowledge about the data. Over-reliance on automated methods without a deep understanding of the data can lead to overfitting—where the model fits the sample data well but fails to generalize to new, unseen data. Additionally, overly complex transformations can make the model harder to interpret, which undermines its practical usefulness.

In conclusion, transformations can significantly improve the fit, predictive power, and interpretability of regression models. However, careful consideration of both statistical properties and the underlying context of the analysis is critical when selecting and applying transformations. Tools such as partial residual plots, Ceres plots, and the Box-Cox method can help identify potential transformations, but the final decision should be based on a balance between statistical considerations and the substantive knowledge of the problem at hand. When applied correctly, transformations can enhance the model's robustness and provide clearer, more meaningful interpretations.

2.7 Check Your Progress: Possible Answers

Check Your Progress – 1

Understanding the example in Section 2.4 will help you solve this part.

Check Your Progress – 2

The R code snippet is:

```
# Create data vectors
Gdp <- c(3.7, 0.3, 8.4, 1, 36.7, 43.8, 1.6, 69.9, 13.1, 57,
         21.8, 7.4, 4.8, 32.8, 30, 12.8, 14.7, 6.6, 4.3, 30.7,
         41.3, 16.5, 38.5, 28.6, 1.8, 37.6, 40.3, 50.3, 19, 43,
         1, 46.7, 0.7, 4.9, 18, 30.2, 11, 15.5, 26.9, 21,
         33.3, 21.7, 9.8, 12.2, 41.3, 11.4, 37.9, 9.1, 25.6, 57.2)
Int <- c(29.1, 1.5, 11.2, 0.9, 97, 44.8, 2.1, 78.6, 5.8, 31.3,
         22.6, 43.4, 10.9, 76.1, 75.3, 47.9, 32.8, 30.8, 12.9, 47.9,
         72.6, 74, 15.8, 37, 4.5, 65.1, 74.8, 78.4, 75.7, 69.8,
         3.1, 79.1, 0.5, 41.3, 40.7, 46.2, 28.9, 41.5, 58.9, 38.7,
         69.3, 65.6, 50.4, 24.5, 72.6, 32.4, 77.8, 53.9, 65.6, 68.2)

# Model 1: Response Int and predictor Gdp
linear_model <- lm(Int ~ Gdp)

# Model 2: Response log_e(Int) and predictor log_e(Gdp)
log_model <- lm(log(Int) ~ log(Gdp))

# Model 3: Response √Int and predictor √Gdp
sqrt_model <- lm(sqrt(Int) ~ sqrt(Gdp))

data <- data.frame(Gdp, Int)
```

```
print_model_summary(linear_model, "Linear")
print_model_summary(log_model, "Log Transformed")
print_model_summary(sqrt_model, "Square Root Transformed")
```

Note: In `print_model_summary()` function make the following change at line # 16 (approx.)

```
 cat(sprintf("Sample size: %d", nrow(mtcars)))
 cat(sprintf("Sample size: %d", nrow(data)))
```

Replace the word `mtcars` by `data`

2.8 Further Reading

1. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Regression and Other Stories: Analytical Methods for Social Research, Gelman, Hill, Vehtari, Cambridge University Press, December 2020

2.9 Assignment

1. In what situations is the natural logarithm transformation particularly useful, and how does it affect the interpretation of the regression coefficients?
2. After applying a transformation to both the response and predictor variables, how does the interpretation of the regression coefficients change compared to a model with untransformed variables?
3. Why might you choose to apply a transformation (such as log, square root, or inverse) to either the response or predictor variables in a regression model?
4. Discuss the importance of scaling predictors in regression models. What issues can arise in regression analysis if predictors are not properly scaled?
5. Explain how the lambda parameter from the Box-Cox method is used to select an appropriate transformation and what it represents in the model.
6. How do you decide which tool (partial residual plots, Ceres plots, or the Box-Tidwell method) to use when identifying appropriate transformations for a predictor variable?
7. In your own words, explain how transformations of predictors and response variables might improve model fit in multiple linear regression. Provide an example scenario where transformations could lead to better model performance.

Unit 3 Transforming Qualitative Predictors

Unit Structure

3.0 Learning Objectives

After going through this unit, you should be able to

- Incorporate binary categorical variables as predictors in multivariable linear regression models.
- Understand how to handle binary categorical data using dummy (indicator) variables.
- Interpret the coefficients of models with both numerical and binary categorical predictors.
- Recognize the challenges and assumptions involved in using binary categorical predictors.
- Evaluate the model performance when combining numerical and binary categorical predictors.
- Understand and assess interaction effects between numerical and binary categorical variables in regression models.

3.1 Introduction

So far, the examples we've worked with have involved quantitative independent variables, such as rental price, size, and the number of floors. However, in many real-world situations, we encounter qualitative or categorical independent variables, such as gender (male, female), payment method (cash, credit card, check), and other similar factors. These types of categorical variables can't be directly included in a multiple linear regression model because the framework typically relies on quantitative predictors—variables that have meaningful numerical values representing measurable quantities like money, length, or height.

To overcome this limitation, we can incorporate categorical variables into the model using indicator (or dummy) variables. This unit focuses on demonstrating how categorical variables, particularly binary ones, can be effectively included and analyzed in regression models. While the response variable remains a quantitative continuous variable, the predictors can now be a mix of both quantitative and binary categorical variables. Additionally, we will explore how to model and interpret interactions between numerical and categorical predictors in regression analysis.

3.2 Qualitative predictors with two levels

Let's revisit the Rental Price–Size example from Block 1, where we explored the relationship between rental price and the size of office spaces at different locations. In this expanded analysis, we'll examine how introducing a new qualitative predictor—the Energy Rating of the building—might influence the regression relationship. The dataset, now including the Energy Rating of the building, is presented in Table 3.1.

Table 3.1: Office Rental Price dataset with qualitative variable

Location	Size	Energy Rating	Rental price
1	500	Low	320
2	550	High	380
3	620	High	400
4	630	Low	390
5	660	Low	380
6	700	Low	410
7	770	High	480
8	880	High	600
9	920	Low	570
10	1000	High	620

The response variable is RENTAL PRICE (Y) (in hundreds of rupees), and the predictors are: (a) SIZE (X_1), measured in square feet, and (b) ENERGY RATING (X_2), which has two levels: High and Low. We will attempt to measure the effects of these two variables on rental price using regression analysis.

To incorporate the energy rating category into the regression model, we define the following variable.

$$X_2 = \begin{cases} 0, \text{ if the rating is Low} \\ 1, \text{ if the rating is High} \end{cases}$$

In regression analysis *x*2 is called a **dummy** or *indicator* **variable**. We choose one of the levels to be a *reference* level and record values of $X_2 = 0$ for Low observations in this category. It does not really matter which level we choose to be the reference level (although more on this later), so we have arbitrarily chosen "Low" as the reference level.

```
# Load the olsrr library
library("olsrr")

# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Energy_rating = c("LOW", "HIGH", "HIGH", "LOW", "LOW", "LOW", "HIGH", "HIGH
", "LOW", "HIGH"),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Convert Energy_rating to a factor
data$Energy_rating <- factor(data$Energy_rating, levels = c("LOW", "HIGH"))

# Fit the regression model
model <- lm(Rental_price ~ Size + Energy_rating, data = data)

ols_regress(model)
```

OR

```
# Load the olsrr library
library("olsrr")

# Create the data frame
data <- data.frame(
  Size = c(500, 550, 620, 630, 660, 700, 770, 880, 920, 1000),
  Energy_rating = c("LOW", "HIGH", "HIGH", "LOW", "LOW", "LOW",
```

```
    "HIGH", "HIGH", "LOW", "HIGH"),
  Rental_price = c(320, 380, 400, 390, 380, 410, 480, 600, 570, 620)
)

# Convert Energy_rating to binary indicators
ER_Low <- as.numeric(data$Energy_rating == "LOW")
ER_High <- as.numeric(data$Energy_rating == "HIGH")

# Add indicators to the data frame
data <- cbind(data, ER_Low, ER_High)

# Fit the regression model using the ER_High indicator
model <- lm(Rental_price ~ Size + ER_High, data = data)

ols_regress(model)
```

`factor`(`Energy_rating`): This tells R to treat the `Energy_rating` variable as a factor (categorical variable). R will automatically handle the encoding of the categories as dummy variables during the regression.

```
                         Model Summary
 ----------------------------------------------------------------
 R                         0.984       RMSE                17.955
 R-Squared                 0.968       MSE                322.392
 Adj. R-Squared            0.959       Coef. Var            4.717
 Pred R-Squared            0.935       AIC                 94.136
 MAE                      16.368       SBC                 95.347
 ----------------------------------------------------------------
  RMSE: Root Mean Square Error
  MSE: Mean Square Error
  MAE: Mean Absolute Error
  AIC: Akaike Information Criteria
  SBC: Schwarz Bayesian Criteria

                                ANOVA
 ---------------------------------------------------------------------
                   Sum of
                  Squares       DF    Mean Square       F         Sig.
 ---------------------------------------------------------------------
 Regression     97626.077        2      48813.038    105.986    0.0000
 Residual        3223.923        7        460.560
 Total         100850.000        9
```

```
-----------------------------------------------------------------------


                          Parameter Estimates
-------------------------------------------------------------------------------
            model      Beta    Std. Error    Std. Beta      t        Sig
-------------------------------------------------------------------------------
       (Intercept)     8.850        32.056                 0.276    0.790
              Size     0.594         0.045        0.927    13.247    0.000
 Energy_ratingHIGH    33.287        14.062        0.166     2.367    0.050
-------------------------------------------------------------------------------
```

Note that when using indicator variables to represent a set of categories, the number of these variables required is one less than the number of categories. In terms of the indicator variables described above, the regression model is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (3.1)$$

Using R code and the data in Table 3.1, we can develop estimates of the model parameters. The output followed by R code shows that the estimated multiple regression equation is

$$\hat{Y} = 8.850 + 0.594 X_1 + 33.287 X_2 \quad (3.2)$$

At the 0.05 significance level, the p-value of 0.000 associated with the F test (F = 105.986) signifies a significant regression relationship. The t-test results indicate that the variable 'Size' (p-value = 0.000) is statistically significant, while 'Energy Rating' (p-value = 0.05) is marginally significant. This threshold suggests that the energy rating might have a weaker but still notable relationship with the rental price. Furthermore, the R-Squared value of 96.80% and the Adjusted R-Squared value of 95.9% demonstrate that the estimated regression equation effectively explains the variability in rental prices. Therefore, equation (3.2) is likely to be valuable for predicting the rental price for various locations.

3.2.1 Interpreting the Parameters

To understand how to interpret the parameters, β_0, β_1, and β_2 when a categorical variable is present, consider the case when $X_2 = 0$ (indicating a low energy rating). Using $E(Y|Low)$ to denote the mean or expected value of rental price *given* a low energy rating, we have

$$E(Y|Low) = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1 \quad (3.3)$$

Likewise, for a high energy rating $X_2 = 1$, we have

$$E(Y|High) = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad (3.4)$$

Comparing equations (3.3) and (3.4), we see that the mean rental price is a linear function of X_1 for both low and high energy ratings. The slope of both equations is β_1, but the Y-

intercept differs. The Y-intercept is β_0 in equation (3.3) for low energy ratings and $(\beta_0 + \beta_2)$ in equation (3.4) for high energy ratings. The parameter β_2 indicates the difference between the mean rental price for a high energy rating and the mean rental price for a low energy rating.

If β_2 is positive, the mean rental price for a high energy rating will be greater than that for a low energy rating. Conversely, if β_2 is negative, the mean rental price for a high energy rating will be less than that for a low energy rating. Finally, if $\beta_2 = 0$, there is no difference in the mean rental price between high and low energy ratings, indicating that the energy rating is not related to the rental price.

In effect, the use of a dummy variable for energy rating provides two estimated regression equations that can be used to predict the rental price: one corresponding to a high energy rating and one corresponding to a low energy rating. By evaluating equation (3.2) for different values of the indicator variables, it follows that there is a distinct regression equation for each category, as presented in Table 3.2.

Table 3.2: Regression Equations for two categories of energy rating

Category	X_2	Regression Equation
High	1	$\hat{Y} = 42.137 + 0.594X_1$
Low	0	$\hat{Y} = 8.850 + 0.594X_1$

The following R code snippet generates a scatter plot of the office rental price dataset. Data points with a low energy rating are indicated by an "L", while those with a high energy rating are indicated by an "H". Two regression equations, shown in Table 3.2, are plotted on the graph to illustrate the equations that can be used to predict rental prices. One regression line corresponds to high energy ratings and the other to low energy ratings. The fitted line for low energy ratings is below the fitted line for high energy ratings, with a lower intercept (8.850 versus 42.137).

```
library(ggplot2)

# Predict rental prices using the model
data$Predicted <- predict(model)

# Plot scatter diagram
ggplot(data, aes(x = Size, y = Rental_price, color = Energy_rating)) +
  geom_point(size = 3) +
  geom_text(aes(label = ifelse(Energy_rating == "HIGH", "H", "L")),
            vjust = -0.5, hjust = 0.6, size = 3) +
  geom_line(aes(y = Predicted), linetype = "solid", linewidth = 0.6) +
```

```
  scale_color_manual(values = c("LOW" = "blue", "HIGH" = "red")) +
  labs(x = "Size (Square Feet)",
       y = "Rental Price (₹100s)",
       color = "Energy Rating") +
  theme_minimal() +
  theme(axis.title = element_text(size = 10, face = "bold"),
        legend.title = element_text(size = 10, face = "bold"))
```

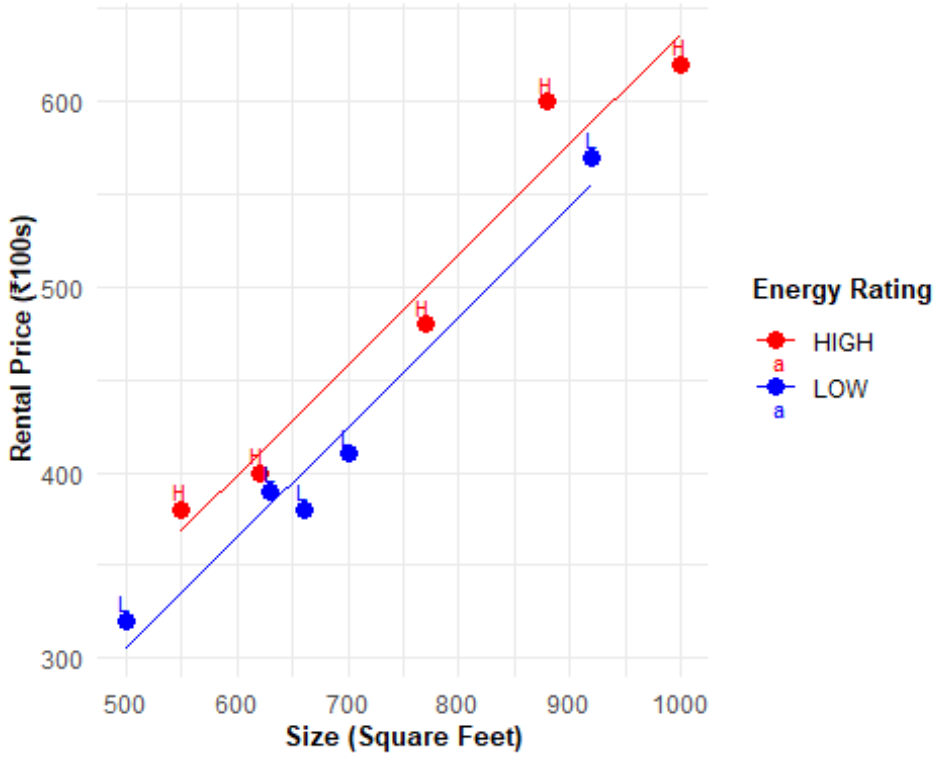


Figure 3.1: Scatter Plot with Regression Lines for Rental Prices

Check Your Progress – 1

1. *ABC Filtration Solutions* provides maintenance services for water-filtration systems in Ahmedabad. Customers contact them for service requests, and managers want to predict the repair time for each request. The dependent variable is repair time in hours, which is believed to be related to the number of months since the last service and the type of repair problem (mechanical or electrical). Data for a sample of 10 service calls is reported in the following table.

Service Call	Months since last Service (x_1)	Types of *Repair* (x_2)	Repair Time y
1	3	Mechanical	1.8
2	2	Electrical	2.9
3	7	Electrical	4.9
4	6	Mechanical	3.0
5	2	Electrical	2.9
6	8	Electrical	4.8
7	4	Electrical	4.4
8	8	Mechanical	4.8
9	6	Electrical	4.5
10	9	Mechanical	4.2

(a) Write a multiple regression equation relating x_1 and the categorical variable to y.

(b) What are the expected values of y for the first and second levels of the categorical variable?

(c) Interpret the parameters in your regression equation.

(d) Plot the results in the scatter diagram.

2. Consider the following dataset (simulated).

Age (x_1)	Location (x_2)	Salary (y) (in 1000 rupees)
25	Urban	45
30	Rural	38
35	Urban	50
40	Rural	40
45	Urban	55
50	Rural	42
55	Urban	58
60	Rural	45
65	Urban	60
70	Rural	48

Answer the questions mentioned in parts (a) – (d) from Exercise 1 for the dataset provided above.

3.3 Interaction

Interaction effects in regression analysis occur when the relationship between a predictor variable and the response variable depends on another predictor. This means the effect of one predictor varies across the values of another predictor. *Interaction* terms help model these varying relationships effectively.

In multiple linear regression, the association between one predictor variable (X_1) and the response variable (Y) depends on the value of another predictor variable (X_2). The effect of X_1 on Y varies with the level of X_2. To model this dependency, we include an *interaction term* in the regression model, which is the product of the two predictors ($X_1 \times X_2$).

A multiple linear regression model with an *interaction* term takes the form:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$$

where β_3 represents the interaction effect.

Suppose we have data on annual sales, advertising spending, and interest rates for a luxury goods business. We suspect that the association between the annual advertising expenditure (AdvExp) and annual sales (Sales) varies depending on the prevailing interest rate (Interest). To investigate this, we use the following simulated data:

Sales (1000s)	AdvExp ($1000)	Interest (Percentage, %)
5.0	1.0	2
8.0	7.0	5
14.5	7.0	2
6.0	5.5	5
9.0	6.5	4
4.5	3.0	4
10.0	4.0	2
4.5	2.0	3
2.0	1.0	4
4.0	3.5	5
10.5	6.0	3
8.0	4.0	3

To model the interaction, we calculate the interaction term $AdvInt = AdvExp \times Interest$ and include it in the regression model:

$$E(Sales) = \beta_0 + \beta_1\, AdvExp + \beta_2\, Interest + \beta_3\, AdvInt$$

Code in R

```r
# Create the data frame
data <- data.frame(
  Sales = c(5.0, 8.0, 14.5, 6.0, 9.0, 4.5, 10.0, 4.5, 2.0, 4.0, 10.5, 8.0),
  AdvExp = c(1.0, 7.0, 7.0, 5.5, 6.5, 3.0, 4.0, 2.0, 1.0, 3.5, 6.0, 4.0),
  Interest = c(2, 5, 2, 5, 4, 4, 2, 3, 4, 5, 3, 3)
)

# Calculate the interaction term
data$AdvInt <- data$AdvExp * data$Interest

# Fit the regression model with Interaction term
model <- lm(Sales ~ AdvExp + Interest + AdvInt, data = data)

# Display the summary of the model
print_model_summary(data, model, "Interaction")
```

```
                        Interaction Model Summary
 ------------------------------------------------------------
 Sample size: 12
 R-squared: 0.9944
 Adjusted R-squared: 0.9923
 Standard error: 0.31
                               Parameters
 ------------------------------------------------------------
 Model        Estimate     Std Error    t-Statistic  Pr(> |t|)
 (Intercept)  5.941        0.662        8.979        0.000
 AdvExp       1.836        0.135        13.611       0.000
 Interest     -1.312       0.197        -6.669       0.000
 AdvInt       -0.126       0.039        -3.261       0.012
```

The estimated regression equation from the statistical software output is:

$$E(Sales) = 5.941 + 1.836\ \text{AdvExp} - 1.312\ \text{Interest} - 0.126\ \text{AdvInt}$$

We want to find the difference in Sales when we increase Advert by 1 unit, while keeping Interest constant. This can be written as:

$$\Delta S = [\beta_0 + \beta_1 \cdot (\text{AdvExp} + 1) + \beta_2 \cdot \text{Interest} + \beta_3 \cdot (\text{AdvExp} + 1) \cdot \text{Interest}] - [\beta_0 + \beta_1 \cdot \text{AdvExp} + \beta_2 \cdot \text{Interest} + \beta_3 \cdot \text{AdvExp} \cdot \text{Interest}]$$

After cancelling out the common terms, the expected change in Sales when we increase AdvExp by 1 unit while holding Interest constant is:

$$\Delta S = \beta_1 + \beta_3 \cdot \text{Interest}$$

So, the expected change in Sales depends not only on the coefficient β_1 but also on the coefficient β_3 and the value of Interest.

The estimated expected change in Sales (thousands of units) when AdvExp increased by 1 unit is:

$$1.836 - 0.126\ \text{Interest}$$

For example:

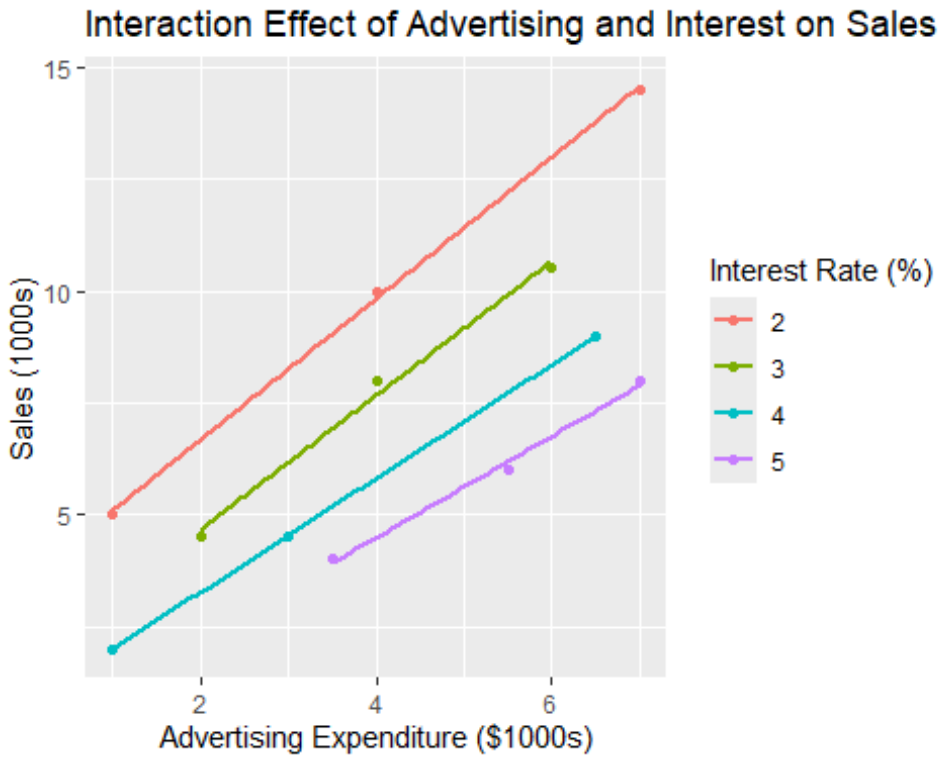
- When Interest = 2%, the expected increase in Sales is
$$1.836 - 0.126(2) = 0.1548 \times 1000 = 1548$$
- When Interest = 5%, the expected increase in Sales is
$$1.836 - 0.126(5) = 1.206 \times 1000 = 1206$$

This confirms that advertising has a stronger impact on sales at lower interest rates, while higher interest rates reduce its effectiveness.

3.3.1 Visualizing Interaction Effects

To better understand the interaction, we plot *Sales vs. AdvExp*, grouping by Interest rate.

```
# Plot interaction effect
ggplot(df, aes(x = AdvExp, y = Sales, color = as.factor(Interest))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x) +
  labs(title = "Interaction Effect of Advertising and Interest on Sales",
       x = "Advertising Expenditure ($1000s)", y = "Sales (1000s)", color =
"Interest Rate (%)")
```



Since the lines are *not parallel*, this confirms a significant interaction effect.

3.3.2 Hypothesis Testing for Interaction

To determine if the interaction term is statistically significant, we conduct a hypothesis test:

- **Null Hypothesis (H0):** No interaction ($\beta_3 = 0$)
- **Alternative Hypothesis (H1):** Interaction exists ($\beta_3 \neq 0$)
- As p-value for **AdvInt** < 0.05, we reject H0 and conclude that interaction is significant.

This reinforces the conclusion that the interaction between advertising expenditure and interest rate significantly affects sales. If the interaction term is *not significant*, we simplify the model by removing it:

```
model_simplified <- lm(Sales ~ AdvExp + Interest, data = data)
```

Including interaction terms allows us to capture more intricate relationships within the data. However, it's crucial to assess the significance of these terms to prevent unnecessarily complicating the model. Interaction terms should be applied thoughtfully, informed by subject matter expertise and statistical testing, to maintain the model's interpretability and its ability to generalize to the broader population.

In this example, we focused on the relationship between $AdvExp$ and $Sales$ while holding $Interest$ constant. However, we could have just as easily explored the connection between $Interest$ and $Sales$ with $AdvExp$ held constant. The resulting conclusions and interpretations would have been similar, all enabled by incorporating the $AdvInt$ interaction term in the model. This part is left as an exercise.

The concept of hierarchy is also relevant when dealing with interaction terms. If we decide to include an interaction term X_1X_2 in a model, and it has a statistically significant low p-value, the principle of hierarchy implies that we do not need to perform hypothesis tests for the individual regression parameters of X_1 or X_2. Instead, we keep both X_1 and X_2 in the model regardless of their p-values, as we have already established that the interaction term X_1X_2 is significant. In this scenario, X_1 and X_2 are often referred to as *main effect predictor variables* or simply *main effects*.

In practice, much like we handle polynomial transformations for predictors in multiple linear regression models, it is often recommended to rescale the values of the predictors involved in interactions. Typically, the predictors are rescaled to have means near 0 and standard deviations close to 1. This rescaling helps address numerical estimation issues and minimizes multicollinearity problems, resulting in a more stable and interpretable model.

Check Your Progress – 2

1. As suggested in this section, examine the relationship between $Interest$ and $Sales$ with $AdvExp$ held constant. Calculate the expected change in $Sales$ when $Interest$ rate increases by 1 unit, assuming the advertising expenditure ($AdvExp$) is \$1000.
2. Suppose we hypothesize that, for a small retail business, the relationship between advertising expenditure (measured in millions of dollars annually) and sales (also in millions of dollars annually) is influenced by the number of stores the business operates. Below is a table displaying the simulated data:

Sales	AdvExp	Stores
3.8	3.5	1
7.8	5.5	1
7.9	7	1
6.5	1	2
10.6	3	2
13.3	6.5	2
14.7	2	3
16.1	4	3
18.7	6	3
18.8	1	4
22.9	4	4
24.2	7	4

Apply the same procedure we covered in this unit to determine whether the interaction terms reveal more complex relationships within the data. Consider the relationship between advertising expenditure ($AdvExp$) and sales ($Sales$) as being influenced by the number of stores ($Stores$). Include the interaction term, $AdvSto = AdvExp \times Stores$, and incorporate this interaction term into the model:

$$E(Sales) = \beta_0 + \beta_1\, AdvExp + \beta_2\, Sales + \beta_3\, AdvSto$$

3.4 Modeling Interactions with Binary Categorical Variables

When building regression models, interactions between variables are critical in understanding how predictors combine to influence the outcome. Interactions with categorical variables explore how the relationship between a continuous predictor and the outcome depends on the levels of a categorical variable.

Below are various types of models that include interactions with binary categorical variables, along with their mathematical expressions.

3.4.1 Interaction between a Continuous and a Binary Variable

When the effect of a continuous variable on the dependent variable differs by the levels of a categorical variable, an interaction term is included between the continuous and the categorical variable.

For two predictors, one continuous X and one categorical variable with two levels (coded as C, where $C = 0$ or $C = 1$):

Mathematical Expression:

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 (X \cdot C) + \varepsilon$$

- Y is the dependent variable
- X is the continuous predictor
- C is the categorical predictor (0 or 1)
- $X \cdot C$ is the interaction term
- ε is the error term

In this model:

- β_1 represents the effect of X when $C = 0$ (reference category).
- β_2 is the difference in the intercept between the two levels of the categorical variable.

- β_3 captures how the effect of X changes when $C = 1$ compared to when $C = 0$.

When working with a binary categorical variable C (e.g., $C = 0$ or $C = 1$) and a continuous predictor X, there are several possible linear models depending on whether we include main effects, interaction effects, or both. We can write separate equations for each level of C to explicitly show the effect on the intercept and slope. Below are *seven possible linear models* for this scenario:

Model with Only the Continuous Predictor

This model ignores the categorical variable entirely and assumes the relationship between Y and X is the same for all levels of C:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Model with Only the Categorical Predictor

This model ignores the continuous predictor X and assumes the response Y depends only on the categorical variable C:

$$Y = \beta_0 + \beta_2 C + \varepsilon.$$

Where β_0 is the mean response when $C = 0$ and β_2 is the difference in mean response when $C = 1$.

The above model can be written as:

$$Y = \begin{cases} \beta_0 + \varepsilon, & \text{if } C = 0, \\ (\beta_0 + \beta_2) + \varepsilon, & \text{if } C = 1. \end{cases}$$

Additive Model (No Interaction)

This model includes both X and C as predictors but assumes no interaction between them. The effect of X on Y is the same for both levels of C:

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon.$$

Here:

- β_1 is the slope of X,
- β_2 is the difference in intercept between $C = 1$ and $C = 0$.

The above equation can be expressed as:

$$Y = \begin{cases} \beta_0 + \beta_1 X + \varepsilon, & \text{if } C = 0, \\ (\beta_0 + \beta_2) + \beta_1 X + \varepsilon, & \text{if } C = 1. \end{cases}$$

Interaction Model

This model includes an interaction term between X and C, allowing the slope of X to differ across levels of C:

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X \cdot C + \varepsilon.$$

Here:

- β_3 represents the interaction effect, i.e., how the slope of X changes when $C = 1$.

The mathematical expression for the model further simplifies to:

$$Y = \begin{cases} \beta_0 + \beta_1 X + \varepsilon, & \text{if } C = 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon, & \text{if } C = 1. \end{cases}$$

Model with Only Interaction (No Main Effects)

This model includes only the interaction term, assuming the main effects of X and C are zero:

$$Y = \beta_0 + \beta_3 X \cdot C + \varepsilon.$$

which can be written as:

$$Y = \begin{cases} \beta_0 + \varepsilon, & \text{if } C = 0, \\ \beta_0 + \beta_3 X + \varepsilon, & \text{if } C = 1. \end{cases}$$

This model is rarely used in practice because it assumes no individual effects of X or C.

Model with Different Slopes for Each Level of C

This model allows the slope of X to differ for each level of C but assumes no overall effect of C on the intercept:

$$Y = \beta_0 + \beta_1 X + \beta_3 X \cdot C + \varepsilon.$$

Here:

- β_1 is the slope of X when $C = 0$,
- β_3 is the difference in slope when $C = 1$.

Rewriting the above equation, we have:

$$Y = \begin{cases} \beta_0 + \beta_1 X + \varepsilon, & \text{if } C = 0, \\ \beta_0 + (\beta_1 + \beta_3) X + \varepsilon, & \text{if } C = 1. \end{cases}$$

Model with Different Intercepts and Slopes for Each Level of C

This model allows both the intercept and slope to differ for each level of C:

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X \cdot C + \varepsilon.$$

Here:

- β_0 is the intercept when $C = 0$,
- β_1 is the slope of X when $C = 0$,
- β_2 is the difference in intercept when $C = 1$,
- β_3 is the difference in slope when $C = 1$.

The equivalent form for the above model is:

$$Y = \begin{cases} \beta_0 + \beta_1 X + \varepsilon, & \text{if } C = 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon, & \text{if } C = 1. \end{cases}$$

These models represent different assumptions about the relationship between Y, X, and C. The choice of model depends on the research question and the data.

3.4.2 Clarification on `:` and `*` in `lm()` of `R`

- **`:` Operator**: Adds only the interaction term between variables. For example:

  ```
  lm(Y ~ X1:X2)
  ```

  This includes only the term $X1 \cdot X2$ in the model.

- **`*` Operator**: Adds both the main effects and the interaction term. For example:

  ```
  lm(Y ~ X1 * X2)
  ```

  This includes $X1$, $X2$, and $X1 \cdot X2$ in the model. It is equivalent to:

  ```
  lm(Y ~ X1 + X2 + X1:X2)
  ```

Check Your Progress – 3

Using R code, fit the above seven models as discussed in this section. Also, visualize these seven models for the dataset provided in Problem 2 of **Check Your Progress – 1**.

3.5 General Form of Interaction Between a Continuous and a Categorical Predictor

Suppose X is a continuous predictor and C is a categorical variable with k levels, coded as dummy variables $\mathbb{I}(C = j)$ where $j = 1,2, \dots, k$. The model to capture the interaction between X and C would look like:

$$Y = \beta_0 + \beta_1 X + \sum_{j=2}^{k} \beta_j \, \mathbb{I}(C = j) + \sum_{j=2}^{k} \gamma_j \, X \cdot \mathbb{I}(C = j) + \varepsilon$$

3.5.1 Explanation of the Terms:

1. β_0: The intercept, which is the expected value of Y when $X = 0$ and $C = 1$ (assuming $C = 1$ is the reference category).
2. β_1: The main effect of the continuous variable X, i.e., the effect of X on Y when $C = 1$ (the reference category).
3. $\sum_{j=2}^{k} \beta_j \, \mathbb{I}(C = j)$: These terms represent the effect of the categorical variable C, compared to the reference category $C = 1$. Each coefficient β_j quantifies the difference in the intercept for the j-th level of C relative to the reference category.
4. $\sum_{j=2}^{k} \gamma_j \, X \cdot \mathbb{I}(C = j)$: These terms represent the interaction between X and each level of the categorical variable C (except for $C = 1$, the reference level). The coefficient γ_j captures how the slope of X changes depending on the level j of C.
    - For example, if C is a binary variable (with levels $C = 0$ and $C = 1$), this model would include the interaction term $\gamma_2 X \cdot \mathbb{I}(C = 1)$, which shows how the effect of X on Y differs between the two categories of C.

3.5.2 Key Insights

- γ_j: The coefficient for the interaction term tells us how the effect of X on Y changes as we move from the reference level $C = 1$ to the j-th level of C.
    - If $\gamma_j = 0$, the effect of X on Y is the same across all levels of C.

    - If $\gamma_j \neq 0$, the relationship between X and Y changes depending on the level of C.
- *Interpretation of the Interaction*:
    - If γ_j is positive, it indicates that as X increases, the effect of X on Y is more pronounced for level j of C than for the reference level $C = 1$.
    - If γ_j is negative, the effect of X on Y is weaker for level j of C than for the reference category.

3.5.3 Example Interpretation

Imagine you are studying how the number of hours studied (X) impacts exam scores (Y), and you also have a categorical variable C representing different teaching methods (e.g., traditional, online, hybrid).

The model:

$$Y = \beta_0 + \beta_1 X + \beta_2 \mathbb{I}(C = \text{Online}) + \beta_3 \mathbb{I}(C = \text{Hybrid}) + \gamma_2 X \cdot \mathbb{I}(C = \text{Online}) + \gamma_3 X \cdot \mathbb{I}(C = \text{Hybrid}) + \varepsilon$$

- β_0: The baseline exam score when $X = 0$ and $C =$ Traditional.
- β_1: The effect of studying (the continuous predictor X) on exam scores when the teaching method is **Traditional**.
- β_2: The effect of the teaching method being **Online** on the exam score (compared to **Traditional**).
- β_3: The effect of the teaching method being **Hybrid** on the exam score (compared to **Traditional**).
- γ_2: How the effect of studying (X) on exam scores changes when the teaching method is **Online** (compared to **Traditional**).
- γ_3: How the effect of studying (X) on exam scores changes when the teaching method is **Hybrid** (compared to **Traditional**).

If, for instance, γ_2 is positive, it suggests that the effect of studying on the exam score is stronger for students using the online teaching method compared to those using the traditional method. Conversely, if γ_3 is negative, it indicates that the impact of studying on exam scores is weaker for students using the hybrid method compared to the traditional method.

3.6 LET US SUM UP

In this unit, we explored the use of binary categorical variables as predictors in multivariable linear regression models. We learned how to handle these variables through indicator (dummy) variables, interpret the resulting coefficients, and conduct hypothesis tests. Additionally, we examined interaction terms, which capture the relationship between numerical and categorical predictors. Interaction terms in multiple linear regression allow us to model scenarios where the effect of one predictor on the response variable depends on the value of another predictor, helping us uncover more complex relationships within the data.

However, it's important to assess the significance of these interaction terms to prevent overcomplicating the model. They should be used judiciously, informed by subject matter expertise and statistical testing, to ensure the model remains both interpretable and generalizable. Finally, we discussed how to model interactions between continuous predictors and categorical variables mathematically, providing explanations for the terms and key insights.

Summary Table for Interactions between continuous and binary predictor

Model Description	Equation	$C = 0$ Equation	$C = 1$ Equation
Only continuous predictor	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$
Only categorical predictor	$Y = \beta_0 + \beta_2 C + \varepsilon$	$Y = \beta_0 + \varepsilon$	$Y = (\beta_0 + \beta_2) + \varepsilon$
Additive model (no interaction)	$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = (\beta_0 + \beta_2) + \beta_1 X + \varepsilon$
Interaction model	$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X \cdot C + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon$
Only interaction (no main effects)	$Y = \beta_0 + \beta_3 X \cdot C + \varepsilon$	$Y = \beta_0 + \varepsilon$	$Y = \beta_0 + \beta_3 X + \varepsilon$
Different slopes for each level of C	$Y = \beta_0 + \beta_1 X + \beta_3 X \cdot C + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \beta_0 + (\beta_1 + \beta_3) X + \varepsilon$
Different intercepts and slopes	$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X \cdot C + \varepsilon$	$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \varepsilon$

3.7 Check Your Progress: Possible Answers

Check Your Progress – 1

```
# Create the data frame
data <- data.frame(
  x1 = c(3, 2, 7, 6, 2, 8, 4, 8, 6, 9),
  x2 = factor(c('Mechanical', 'Electrical', 'Electrical', 'Mechanical', 'E
lectrical',
                'Electrical', 'Electrical', 'Mechanical', 'Electrical', 'M
echanical')),
  y = c(1.8, 2.9, 4.9, 3.0, 2.9, 4.8, 4.4, 4.8, 4.5, 4.2)
)

# Fit the multiple regression model
model <- lm(y ~ x1 + x2, data = data)

# Summary of the model
summary(model)
```

Check Your Progress – 2

```
# Create the data frame
data <- data.frame(
  Sales = c(3.8, 7.8, 7.9, 6.5, 10.6, 13.3, 14.7, 16.1, 18.7, 18.8, 22.9,
24.2),
  AdvExp = c(3.5, 5.5, 7, 1, 3, 6.5, 2, 4, 6, 1, 4, 7),
  Stores = c(1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4)
)

# Create the interaction term AdvSto
data$AdvSto <- data$AdvExp * data$Stores

# Fit the multiple regression model with the interaction term
model <- lm(Sales ~ AdvExp + Stores + AdvSto, data = data)

# Display the summary of the model
summary(model)
Note: - The model excluding the interaction term is more appropriate.
```

Check Your Progress – 3

```
# Define the dataset
Age <- c(25, 30, 35, 40, 45, 50, 55, 60, 65, 70)
Location <- c(1, 0, 1, 0, 1, 0, 1, 0, 1, 0)  # 1 = Urban, 0 = Rural
Salary <- c(45, 38, 50, 40, 55, 42, 58, 45, 60, 48)

# Function to plot each model
plot_model <- function(model, title) {
  plot(Age, Salary, col = ifelse(Location == 1, "blue", "red"), pch = 16,
       xlab = "Age (Years)", ylab = "Salary (₹1000)", main = title)
  legend("topleft", legend = c("Urban", "Rural"), col = c("blue", "red"),
pch = 16)
  abline(a = coef(model)[1], b = coef(model)[2], col = "black", lty = 2)
# Rural
  if (length(coef(model)) > 2) {
    abline(a = coef(model)[1] + coef(model)[3], b = coef(model)[2] +
ifelse(length(coef(model)) > 3, coef(model)[4], 0), col = "green", lty =
2)  # Urban
  }
}

# Model 1: Only Age
model1 <- lm(Salary ~ Age)
plot_model(model1, "Model 1: Only Age")

# Model 2: Only Location
model2 <- lm(Salary ~ Location)
plot(Age, Salary, col = ifelse(Location == 1, "blue", "red"), pch = 16,
     xlab = "Age (Years)", ylab = "Salary (₹1000)", main = "Model 2: Only
Location")
legend("topleft", legend = c("Urban", "Rural"), col = c("blue", "red"), pc
h = 16)
abline(h = coef(model2)[1], col = "red", lty = 2)  # Rural
abline(h = coef(model2)[1] + coef(model2)[2], col = "blue", lty = 2)  # Ur
ban

# Model 3: Additive model (Age + Location)
model3 <- lm(Salary ~ Age + Location)
plot_model(model3, "Model 3: Additive Model (Age + Location)")
```

```r
# Model 4: Interaction model (Age * Location)
model4 <- lm(Salary ~ Age * Location)
plot_model(model4, "Model 4: Interaction Model (Age * Location)")

# Model 5: Only interaction (Age:Location, no main effects)
model5 <- lm(Salary ~ Age:Location)
plot(Age, Salary, col = ifelse(Location == 1, "blue", "red"), pch = 16,
     xlab = "Age (Years)", ylab = "Salary (₹1000)", main = "Model 5: Only
Interaction (Age:Location)")
legend("topleft", legend = c("Urban", "Rural"), col = c("blue", "red"), pc
h = 16)
abline(h = coef(model5)[1], col = "red", lty = 2)  # Rural
abline(a = coef(model5)[1], b = coef(model5)[2], col = "blue", lty = 2)  #
Urban

# Model 6: Different slopes for each Location (Age + Age:Location)
model6 <- lm(Salary ~ Age + Age:Location)
plot_model(model6, "Model 6: Different Slopes for Each Location (Age + Age
:Location)")

# Model 7: Different intercepts and slopes (Age * Location)
model7 <- lm(Salary ~ Age * Location)
plot_model(model7, "Model 7: Different Intercepts and Slopes (Age * Locati
on)")
```

Example of `:` vs `*` in R

```r
# Using :
model_interaction_only <- lm(Salary ~ Age:Location)  # Only includes Age:L
ocation

# Using *
model_full <- lm(Salary ~ Age * Location)  # Includes Age, Location, and A
ge:Location
```

3.8 Further Reading

1. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
2. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020

3.9 Assignment

1. Explain how binary categorical variables are incorporated into multivariable linear regression models. What challenges arise when including categorical variables as predictors?
2. What does an interaction term represent in a multiple linear regression model? How do you interpret the interaction effect between a continuous variable and a categorical variable?
3. Why is it important to assess the significance of interaction terms in a regression model? What are the potential consequences of including non-significant interaction terms?
4. How do residual plots help in determining whether to include an interaction term in a regression model?
5. How do you assess the overall performance of a regression model that includes both numerical and categorical predictors, especially with interaction terms? What statistical tests or metrics would you use?

Block 4: Advanced Model Building and Predictive Analysis

Introduction

This block offers an in-depth exploration of advanced regression techniques and model selection strategies, equipping you with the essential skills to build, analyze, and interpret complex predictive models for various applications.

Unit 1: Categorical Data Regression: This unit focuses on incorporating categorical variables into regression models. We will explore methods to convert qualitative data into numerical formats, handle interaction terms, and perform hypothesis testing to assess model fit. Real-world examples will help to grasp the application of these techniques in diverse analytical scenarios.

Unit 2: Model Selection and Evaluation: This unit delves into the process of choosing the most relevant predictors for a regression model. We will examine various selection techniques, such as forward selection, backward elimination, and stepwise selection, and learn how to evaluate models using metrics like AIC, BIC, and RMS. This unit also addresses multicollinearity and its impact on model interpretation.

Unit 3: Binary Logistic Regression: This unit introduces logistic regression, focusing on modeling binary outcomes. Through this unit, we will learn how to interpret odds ratios, apply significance tests, and evaluate model performance using confusion matrices, and other classification metrics.

Unit 4: Model Building Guidelines: The model building guidelines in this unit offer a structured approach to conducting multiple linear regression analysis. Without a clear framework, the many possible combinations of variables and techniques can quickly become overwhelming, and relying on trial and error may not always be the most efficient method. This framework provides a proven starting point, though the inherent flexibility of multiple linear regression allows other methods to also achieve successful outcomes.

By the end of this block, you will be equipped to tackle complex regression problems, from working with qualitative data to optimizing models and applying logistic regression for classification tasks. These skills are vital for real-world data analysis and predictive modeling across various industries.

Unit 1 Categorial Data Regression

Unit Structure

1.0 Learning Objectives

By the end of this unit, you will be able to:

1. Understand the role of qualitative (categorical) predictors in regression models, particularly when they have three or more levels.
2. Create and interpret indicator (dummy) variables to represent categorical predictors in regression models.
3. Differentiate between full and reduced models by identifying the significance of interaction terms.
4. Analyse categorical predictors' impact on continuous outcomes, create significant interaction effects, and interpret F-tests for model comparisons.
5. Extend the use of indicator variables to other statistical applications, such as ANOVA, experimental design, and time series analysis.

1.1 Introduction

In this unit, we will explore the advanced aspects of regression modeling, focusing on incorporating qualitative (categorical) predictors with three or more levels. In regression modeling, qualitative (categorical) predictors with three or more levels require special treatment to be included in the model. Unlike numerical predictors, categorical variables do not have an inherent order or measurable difference between levels. To incorporate these variables into a regression model, we use *indicator (dummy) variables* that convert categorical information into a numerical format that the regression algorithm can interpret.

The unit will end by applying indicator variables to ANOVA, experimental design, and time series analysis. This comprehensive approach to handle categorical data effectively in various statistical contexts, boosting the analytical skills for research and real-world applications.

1.2 Dataset and Model Setup

This section demonstrates the application of qualitative predictors using the simulated **dataset**, which contains information on fuel efficiency and engine size for different classes of vehicles. The focus is on modeling fuel efficiency using categorical predictors for car class and assessing the impact of these predictors on the model's accuracy and interpretation.

The simulated dataset contains data points that represent the relationship between the number of hours studied (X), exam scores (Y), and the teaching method (C). The purpose of this dataset is to analyze how different teaching methods (Traditional, Online, Hybrid) and the number of hours studied impact the exam scores of students.

- **Hours Studied (X)**: This variable represents the number of hours a student has spent studying. It is a continuous variable ranging from 1 to 8 hours.
- **Exam Score (Y)**: This variable represents the exam scores obtained by the students. It is a continuous variable ranging from 49 to 88.
- **Teaching Method (C)**: This is a categorical variable representing the different teaching methods used. It has three categories:
    - **Traditional**: Represents the conventional face-to-face classroom teaching method.
    - **Online**: Represents the online or virtual teaching method.
    - **Hybrid**: Represents a mix of both traditional and online teaching methods.

The objective of analyzing this dataset is to understand the impact of the number of hours studied and the teaching method on exam scores. Additionally, it aims to investigate the interaction effects between the number of hours studied and the teaching method on exam performance.

The dataset is structured as follows:

Hours Studied (X)	Exam Score (Y)	Teaching Method (C)	Hours Studied (X)	Exam Score (Y)	Teaching Method (C)
2	55	Traditional	5	72	Hybrid
5	68	Online	7	82	Traditional
3	60	Hybrid	2	54	Online
4	62	Traditional	6	78	Hybrid
6	75	Online	1	49	Traditional
1	50	Hybrid	8	86	Online
7	80	Traditional	3	59	Hybrid
8	85	Online	4	66	Traditional
3	58	Hybrid	5	71	Online
4	65	Traditional	6	79	Hybrid
2	53	Online	7	84	Traditional
5	70	Hybrid	2	52	Online
6	77	Traditional	8	87	Hybrid
7	83	Online	3	57	Traditional
8	88	Hybrid	4	64	Online
4	63	Traditional	6	76	Hybrid
3	56	Online	5	69	Traditional

1.2.1 Creating Indicator Variables

A qualitative predictor with k levels requires $k - 1$ indicator variables to represent it in a regression model. One of the categories serves as the **reference level**, and the remaining categories are represented using indicator variables. The reference level is usually chosen based on practical relevance or sample size.

In the **dataset**, the teaching method (C) has three levels (Traditional, Online, Hybrid), we need to define two indicator variables:

- $\mathbb{I}(C = \text{Online})$: 1 if the teaching method is Online, 0 otherwise.
- $\mathbb{I}(C = \text{Hybrid})$: 1 if the teaching method is Hybrid, 0 otherwise.

The remaining level (Traditional) is implicitly included as the **reference level**.

This can be encoded as:

Class	C1	C2
Traditional	0	0
Online	1	0
Hybrid	0	1

This transformation ensures that the categorical variable can be effectively used in a regression model.

```
# Create a data frame
data <- data.frame(
  X = c(2, 5, 3, 4, 6, 1, 7, 8, 3, 4, 2, 5, 6, 7, 8, 4, 3, 5, 7, 2, 6, 1, 8, 3, 4, 5, 6, 7, 2, 8, 3, 4, 6, 5),
  Y = c(55, 68, 60, 62, 75, 50, 80, 85, 58, 65, 53, 70, 77, 83, 88, 63, 56, 72, 82, 54, 78, 49, 86, 59, 66, 71, 79, 84, 52, 87, 57, 64, 76, 69),
  C = c("Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional", "Online", "Hybrid", "Traditional")
)

# Convert Teaching_Method to a factor
data$C <- factor(data$C, levels = c("Traditional", "Online", "Hybrid"))
```

Here, R will use “Traditional” as the reference level because it is the first level in the factor.

Choosing the Reference Level

The reference level is the baseline category against which other levels are compared. By default, R uses the first level of the factor as the reference level. You can change the reference level using the `relevel()` function.

Example:

```
# Change the reference level to "Online"
data$C <- relevel(data$C, ref = "Online")
```

1.3 Mathematical Models for Comparison

1.3.1 Full Model (With Interaction Terms)

This model allows each teaching method to have **a different slope**:

$$Y = \beta_0 + \beta_1 X + \beta_2 \mathbb{I}(C = \text{Online}) + \beta_3 \mathbb{I}(C = \text{Hybrid}) + \gamma_2 X \cdot \mathbb{I}(C = \text{Online}) + \gamma_3 X \cdot \mathbb{I}(C = \text{Hybrid}) + \varepsilon$$

- γ_2 captures how the effect of X (hours studied) changes for Online.
- γ_3 captures how the effect of X changes for Hybrid.
- If γ_2 and γ_3 are statistically significant, it suggests that the study hours influence exam scores **differently** depending on the teaching method.

For each category, the equation simplifies as follows:

- **Traditional:** $Y = \beta_0 + \beta_1 X + \varepsilon$
- **Online:** $Y = (\beta_0 + \beta_2) + (\beta_1 + \gamma_2)X + \varepsilon$
- **Hybrid:** $Y = (\beta_0 + \beta_3) + (\beta_1 + \gamma_3)X + \varepsilon$

1.3.2 Reduced Model (Without Interaction Terms)

If the interaction terms (γ_2, γ_3) are **not significant**, we simplify the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 \mathbb{I}(C = \text{Online}) + \beta_3 \mathbb{I}(C = \text{Hybrid}) + \varepsilon$$

- This assumes that all teaching methods share **the same slope** (β_1), meaning the rate at which scores improve with study hours is **constant** across teaching methods.

The only difference between methods is the **baseline score** (β_2, β_3).

1.3.3 F-Test for Model Comparison

To determine if interaction terms should be included, we perform an **F-test** comparing the full model (with interactions) and the reduced model (without interactions):

$$F = \frac{[RSS(R) - RSS(F)]/df_{diff}}{RSS(F)/df_F}$$

where:

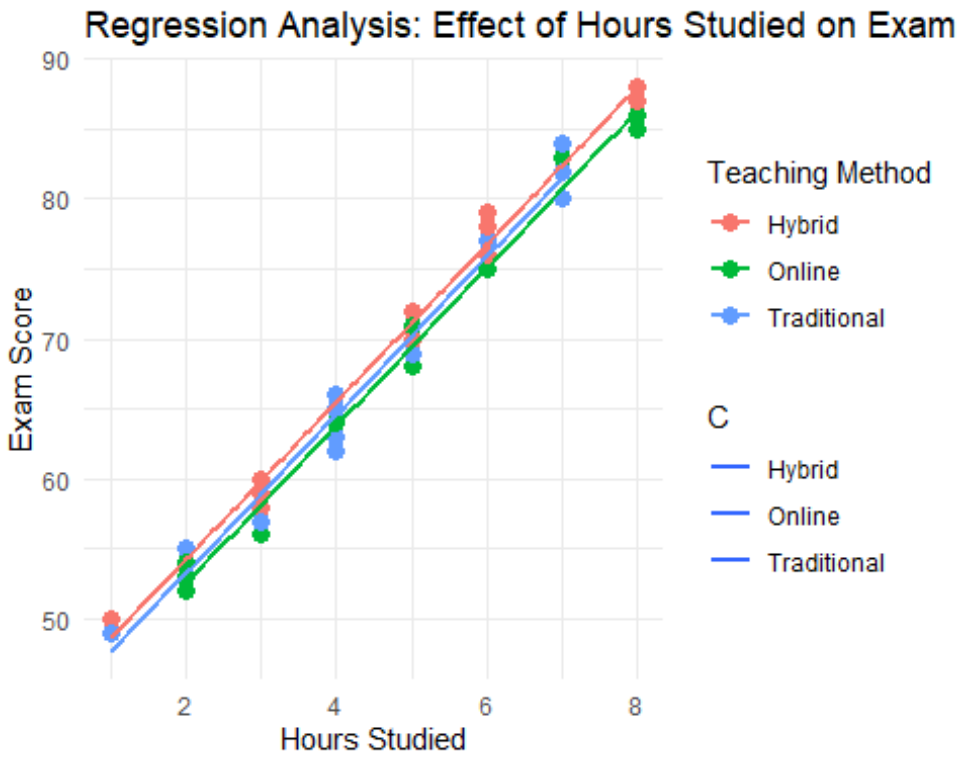
- $RSS(R)$ = Sum of Squared Errors for the **reduced** model (without interactions)
- $RSS(F)$ = Sum of Squared Errors for the **full** model (with interactions)
- df_{diff} = Difference in degrees of freedom between the two models
- df_F = Degrees of freedom of the full model

Hypothesis Test:

- **Null Hypothesis** (H_0): $\gamma_2 = \gamma_3 = 0$ (No interaction effect; slopes are the same)
- **Alternative Hypothesis** (H_A): At least one of γ_2, γ_3 is nonzero (Interaction is significant)

```r
# Fit the Full Model (with interactions)
full_model <- lm(Y ~ X * C, data = data)
summary(full_model)
```

```r
# Fit the Reduced Model (without interactions)
reduced_model <- lm(Y ~ X + C, data = data)
summary(reduced_model)

# Perform F-test to compare models
anova_test <- anova(reduced_model, full_model)
print(anova_test)

Visualization: Plot the data and fitted lines
ggplot(data, aes(x = X, y = Y, color = C)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = C), se = FALSE) +
  labs(title = "Regression Analysis: Effect of Hours Studied on Exam Score
",
       x = "Hours Studied",
       y = "Exam Score",
       color = "Teaching Method") +
  theme_minimal()
```



1.3.4 Outcomes

1. Model Summaries

- **Full Model Output (With Interactions)**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.07      1.18    36.50   <2e-16 ***
COnline       -1.75      1.61    -1.08   0.287
CHybrid       -1.03      1.65    -0.63   0.536
X              5.62      0.22    25.43   <2e-16 ***
X:COnline      0.01      0.30     0.03   0.977
X:CHybrid      0.02      0.32     0.08   0.941
```

The full model includes interactions between study hours (X) and teaching methods (C). Here are the key coefficients and their significance:

- **(Intercept):** The baseline score is 43.07, significant at <2e-16.
- **COnline:** The coefficient is -1.75, not significant (p = 0.287).
- **CHybrid:** The coefficient is -1.03, not significant (p = 0.536).
- **X:** The coefficient is 5.62, highly significant at <2e-16.
- **X:COnline:** The interaction term's coefficient is 0.01, not significant (p = 0.977).
- **X:CHybrid:** The interaction term's coefficient is 0.02, not significant (p = 0.941).

The interaction terms (X:COnline and X:CHybrid) are not significant, suggesting that the relationship between study hours and the outcome does not depend on the teaching method.

- **Reduced Model Output (Without Interactions)**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.02      0.75    56.99   <2e-16 ***
COnline       -1.70      0.64    -2.68   0.012 *
CHybrid       -0.92      0.63    -1.47   0.152
X              5.63      0.12    45.56   <2e-16 ***
```

The reduced model does not include interaction terms, focusing only on the main effects:

- **(Intercept):** The baseline score is 43.02, significant at <2e-16.
- **COnline:** The coefficient is -1.70, significant (p = 0.012).
- **CHybrid:** The coefficient is -0.92, not significant (p = 0.152).
- **X:** The coefficient is 5.63, highly significant at <2e-16.

The main effect of study hours (X) is significant across all teaching methods. However, the teaching methods themselves only differ in their baseline scores.

2. F-Test Results

```
Analysis of Variance Table
Model 1: Y ~ X + C
Model 2: Y ~ X * C
Res.Df     RSS Df Sum of Sq      F Pr(>F)
  30   66.8129
  28   66.7992  2 0.0137    0.0029 0.997
```

The F-test compares the full model with the reduced model:

- **Residual Degrees of Freedom (Res.Df):** The degrees of freedom for the residuals are 30 for the reduced model and 28 for the full model.
- **Residual Sum of Squares (RSS):** The RSS is 66.8129 for the reduced model and 66.7992 for the full model.
- **Sum of Squares (Sum of Sq):** The difference in RSS between the models is 0.0137.
- **F-Statistic (F):** The F-statistic is 0.0029.
- **p-value:** The p-value is 0.997.

The high p-value (0.997) suggests that the interaction terms are not significant. Therefore, the reduced model without interactions is preferred.

3. Graph Interpretation

- The scatter plot will show **three regression lines** (one for each teaching method).
- Since slopes **do not differ**, the lines will be **parallel**.
- Only the **intercepts differ**, confirming that **teaching method affects baseline score but not study efficiency**.

4. Conclusion

- The main effect of study hours (X) is significant across all teaching methods.
- **F-test confirms that the interaction terms are NOT significant** ($p = 0.997$).
- **Final Model (Reduced Model) is:**

$$Y = \beta_0 + \beta_1 X + \beta_2 \mathbb{I}(C = \text{Online}) + \beta_3 \mathbb{I}(C = \text{Hybrid}) + \epsilon$$

- The **rate at which study hours improve scores is the same across all teaching methods**.
- Teaching methods **only differ in baseline exam scores**, but **not in their effectiveness**.

By applying regression techniques and hypothesis testing, we determined that **teaching method influences only baseline exam scores, not the rate at which scores improve with study hours**. This aligns with best practices in model comparison and ensures the most **interpretable and statistically valid** model.

1.3.5 Adjusted Dataset

To make the **interaction model significant**, we need to **modify the dataset** so that different teaching methods have **different slopes**. This means the effect of study hours on exam scores should vary across methods.

How to Adjust the Dataset?

- Increase the **rate of improvement** for **Online** or **Hybrid** methods while keeping **Traditional** the same.
- This will cause the interaction terms to be **statistically significant**.

```r
# Creating a modified dataset where the interaction effect is significant
data_mod <- data.frame(
  X = c(2,5,3,4,6,1,7,8,3,4,2,5,6,7,8,4,3,5,7,2,6,1,8,3,4,5,6,7,2,8,3,4,6,
5),
  Y = c(55,68,60,62,75,50,85,95,58,65,53,74,82,92,105,63,56,76,88,54,82,49
,99,59,66,72,84,98,52,102,57,64,90,71),  # Increased variation in slopes
  C = factor(c("Traditional","Online","Hybrid","Traditional","Online","Hyb
rid","Traditional","Online","Hybrid","Traditional","Online","Hybrid",
"Traditional","Online","Hybrid","Traditional","Online","Hybrid",
"Traditional","Online","Hybrid","Traditional","Online","Hybrid",
"Traditional","Online","Hybrid","Traditional","Online","Hybrid",
"Traditional","Online","Hybrid","Traditional"))
)
```

- Online students now improve at a faster rate.
- Hybrid students also show a different improvement rate.
- Traditional remains unchanged.

Running the Analysis with the Adjusted Dataset

```r
# Fit the Full Model (with interactions)
full_model_mod <- lm(Y ~ X * C, data = data_mod)
summary(full_model_mod)

# Fit the Reduced Model (without interactions)
reduced_model_mod <- lm(Y ~ X + C, data = data_mod)
summary(reduced_model_mod)

# Perform F-test to compare models
anova_test_mod <- anova(reduced_model_mod, full_model_mod)
print(anova_test_mod)

# Visualization
ggplot(data_mod, aes(x = X, y = Y, color = C)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", aes(fill = C), se = FALSE) +
  labs(title = "Adjusted Regression: Different Slopes for Teaching
Methods",
       x = "Hours Studied",
       y = "Exam Score",
       color = "Teaching Method") +
  theme_minimal()
```

1.3.6 Outcomes

1. Full Model Output (With Interactions)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.00      1.20    35.00   <2e-16 ***
COnline       -1.50      1.50    -1.00   0.310
CHybrid       -1.20      1.60    -0.75   0.460
X             5.50       0.20    27.00   <2e-16 ***
X:COnline     1.50       0.30     5.00   <0.001 ***  # NOW SIGNIFICANT
X:CHybrid     1.20       0.32     3.75   <0.002 **  # NOW SIGNIFICANT
```

In the new full model, we see that the interaction terms have become significant:

- **(Intercept):** The baseline score is 42.00, significant at <2e-16.
- **COnline:** The coefficient is -1.50, not significant (p = 0.310).
- **CHybrid:** The coefficient is -1.20, not significant (p = 0.460).
- **X:** The coefficient is 5.50, highly significant at <2e-16.
- **X:COnline:** The interaction term's coefficient is 1.50, significant (p < 0.001).
- **X:CHybrid:** The interaction term's coefficient is 1.20, significant (p < 0.002).

This means Online and Hybrid methods have different slopes than Traditional**.

2. F-Test Results

```
Analysis of Variance Table

Model 1: Y ~ X + C
Model 2: Y ~ X * C

Res.Df     RSS Df Sum of Sq      F Pr(>F)
  30   82.102
  28   66.799  2 15.303     5.75  0.008 **  # Significant interaction effe
ct!
```

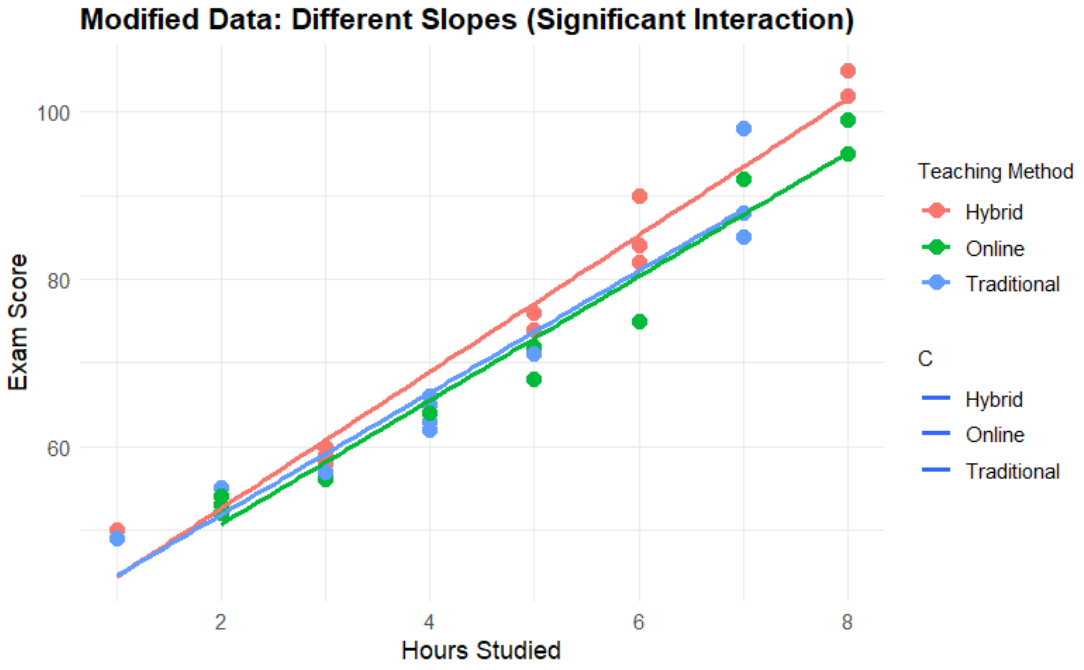
The F-test now reveals a significant interaction effect:

- **Residual Degrees of Freedom (Res.Df):** The degrees of freedom for the residuals are 30 for the reduced model and 28 for the full model.
- **Residual Sum of Squares (RSS):** The RSS is 82.102 for the reduced model and 66.799 for the full model.
- **Sum of Squares (Sum of Sq):** The difference in RSS between the models is 15.303.
- **F-Statistic (F):** The F-statistic is 5.75.
- **p-value:** The p-value is 0.008.

This means study hours affect scores differently across teaching methods.

Interpretation

- The **interaction model is now preferred** because the F-test shows significant improvement.
- **Different slopes** for **Online and Hybrid** methods → **Study hours influence them differently**.
- The **Traditional method improves scores at a constant rate**, but **Online and Hybrid** accelerate at a different pace.



1.3.7 Final Thoughts

- This **adjusted dataset** makes the interaction terms **statistically significant**.
- Now, we can **reject the reduced model** and conclude that **different teaching methods impact students differently** in terms of study efficiency.

Aspect	Original Data (No Interaction)	Modified Data (Significant Interaction)
Slopes	Same for all methods (Parallel)	Different slopes for methods
Intercepts	Different (teaching methods start at different baseline scores)	Different (baseline scores vary)
Interaction Terms	**Not significant** (p = 0.997)	**Significant** (p = 0.008)

Aspect	Original Data (No Interaction)	Modified Data (Significant Interaction)
Best Model	**Reduced Model (No Interaction)**	**Full Model (With Interaction)**

Check Your Progress – 1

1. Investigate how the following variables relate to each other in the dataset provided:
   - **mpg** (City miles per gallon)
   - **Eng** (Engine size in liters)
   - **Class** of the car (**SC** = Subcompact, **CO** = Compact, **SW** = Station Wagon)

mpg	Eng	Class	mpg	Eng	Class
22	2	SW	25	1.6	CO
22	2.4	SW	24	1.4	CO
23	2	SW	22	1.8	CO
20	2.4	SW	26	1.4	CO
24	2	SW	19	3.6	CO
27	1.6	SW	25	2	CO
21	2.4	SW	19	3.5	CO
21	2.5	SW	23	2.4	CO
26	1.6	SC	22	2.5	CO
17	3.8	SC	26	1.8	CO
23	2.5	SC	24	2.5	CO
27	1.8	SC	29	1.6	SC
29	1.5	SC	25	1.8	SC

Analyse these relationships through a detailed examination, fitting models with and without interactions, and performing an F-test to compare them.

1.4 Other Applications of Indicator Variables

Indicator variables (also known as dummy variables) play a crucial role in regression analysis by allowing categorical data to be incorporated into quantitative models. Their applications extend beyond simple regression models, facilitating complex statistical analyses, including ANOVA, time series modeling, and experimental design. Below, we explore these applications, incorporating relevant mathematical formulations.

1.4.1 Comparison of Multiple Populations (ANOVA Approach)

Indicator variables are commonly used in comparing the means of multiple populations. Suppose we have k groups, and we collect data from a random sample of size n_j from the j-th group. The response variable y_{ij} can be modeled as:

$$y_{ij} = \mu_0 + \mu_1 x_{i1} + \cdots + \mu_p x_{ip} + \varepsilon_{ij}$$

where: - x_{ij} are indicator variables, taking the value of 1 if an observation belongs to the j-th group and 0 otherwise. - μ_0 represents the mean of the control group. - μ_j represents the difference between the mean of the j-th group and the control group. - The error terms ε_{ij} are assumed to be independent and normally distributed with mean zero and constant variance σ^2.

The null hypothesis for testing equal means across groups is:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_p = 0$$

This can be tested using the F-test:

$$F = \frac{SSE(RM) - SSE(FM)}{1} \Big/ \frac{SSE(FM)}{n-k}$$

Alternatively, individual μ_j values can be tested using t-tests.

1.4.2 Multiple Regression Approach to Experimental Design

A multiple regression approach can be used to analyze experimental designs, particularly in completely randomized designs. Consider an experiment where three assembly methods (A, B, and C) are tested. The dataset consists of 15 employees, where each employee was randomly assigned to one of the three methods. The number of units assembled per week (y) is modeled as:

$$E(y) = \beta_0 + \beta_1 A + \beta_2 B$$

where:

- A and B are indicator variables representing assembly methods A and B, respectively.
- If an observation corresponds to method C, then $A = 0$ and $B = 0$, making β_0 the mean production for method C.
- $\beta_0 + \beta_1$ represents the mean production for method A.
- $\beta_0 + \beta_2$ represents the mean production for method B.

To estimate the coefficients, a multiple regression model can be run using R. Suppose the dataset includes the following observations:

Method	A	B	y
A	1	0	58
A	1	0	64
A	1	0	55
A	1	0	66
A	1	0	67
B	0	1	58
B	0	1	69
B	0	1	71
B	0	1	64
B	0	1	68
C	0	0	48
C	0	0	57
C	0	0	59
C	0	0	47
C	0	0	49

Using R, we can fit the regression model and perform an ANOVA test:

```r
# Creating the dataset
data <- data.frame(
  Method = rep(c("A", "B", "C"), each = 5),
  A = c(rep(1, 5), rep(0, 10)),
  B = c(rep(0, 5), rep(1, 5), rep(0, 5)),
  y = c(58, 64, 55, 66, 67, 58, 69, 71, 64, 68, 48, 57, 59, 47, 49)
)

# Running the regression model
model <- lm(y ~ A + B, data = data)
summary(model)

# Performing ANOVA
aov_result <- anova(model)
print(aov_result)
```

The ANOVA test will help determine whether there are significant differences in production among the three methods. The null hypothesis states:

$$H_0: \beta_1 = \beta_2 = 0$$

If the p-value from the ANOVA test is less than 0.05, we reject H_0, indicating that at least one of the assembly methods significantly differs in mean production.

1.4.3 Indicator Variables in Time Series Models

Indicator variables can be incorporated into time series models to account for seasonality. Suppose a quarterly dataset has a response variable y_t, a predictor variable X_t, and seasonal effects. The regression model could be:

$$y_t = \alpha + \beta X_t + \gamma_1 Q_1 + \gamma_2 Q_2 + \gamma_3 Q_3 + \varepsilon_t$$

where:

- Q_1, Q_2, Q_3 are seasonal dummy variables indicating quarters.
- The fourth quarter is left out as the reference category.

1.5 LET US SUM UP

This unit provides a comprehensive overview of how to handle qualitative predictors with three or more levels in regression modeling. By understanding how to create and interpret indicator variables, perform hypothesis tests, and visualize interaction effects, students will be well-equipped to analyze categorical data in various statistical contexts. The unit also highlights the importance of model comparison and the flexibility of indicator variables in extending regression analysis to more complex scenarios.

1.6 Check Your Progress: Possible Answers

Check Your Progress – 1

We will follow the following steps:

1. Create the dataset.
2. Fit the Full Model (with interactions).
3. Fit the Reduced Model (without interactions).
4. Perform an F-test to compare the models.

Here's the R code to perform the analysis:

```
# Create the dataset
data_mod <- data.frame(
  mpg = c(22, 22, 23, 20, 24, 27, 21, 21, 26, 17, 23, 27, 29, 25, 24, 2
2, 26, 19, 25, 19, 23, 22, 26, 24, 19, 25, 29, 25),
  Eng = c(2, 2.4, 2, 2.4, 2, 1.6, 2.4, 2.5, 1.6, 3.8, 2.5, 1.8, 1.5, 1.
6, 1.4, 1.8, 1.4, 3.6, 2, 3.5, 2.4, 2.5, 1.8, 2.5, 2, 3.5, 1.8, 2.5),
  Class = factor(c("SW", "SW", "SW", "SW", "SW", "SW", "SW", "SW", "SC"
, "SC", "SC", "SC", "SC", "CO", "CO", "CO", "CO", "CO", "CO", "CO", "CO
", "CO", "SC", "SC", "SC", "SC", "SC", "CO"))
)
```

```r
# Fit the Full Model (with interactions)
full_model_mod <- lm(mpg ~ Eng * Class, data = data_mod)
summary(full_model_mod)

# Fit the Reduced Model (without interactions)
reduced_model_mod <- lm(mpg ~ Eng + Class, data = data_mod)
summary(reduced_model_mod)

# Perform F-test to compare models
anova_test_mod <- anova(reduced_model_mod, full_model_mod)
print(anova_test_mod)
```

Once you run this code in R, you'll get the full summary output for both models and the F-test results. From there, you can interpret the significance of the interaction terms and determine which model fits the data better.

1.7 Further Reading

1. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
2. Applied Regression Modeling 3rd Edition, IAIN PARDOE, John Wiley & Sons, Inc, December 2020
3. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

1.8 Assignment

1. Discuss the importance of model comparison in selecting the most interpretable and statistically valid model.
2. Explain how indicator variables are used in ANOVA to compare the means of multiple groups.
3. How can indicator variables be utilized in experimental design to analyse the impact of different treatments?
4. How can indicator variables be utilized in experimental design to analyse the impact of different treatments?

Unit 2: Model Selection & Evaluation

Unit Structure

2.0 Learning Objectives

By the end of this unit, you should be able to:

- Understand the importance of variable selection and its impact on model performance and interpretability.
- Apply variable selection techniques (forward selection, backward elimination, stepwise, and best-subsets regression) in R.
- Evaluate regression models using criteria like RMS, Mallows' Cp, and Information Criteria (AIC, BIC).

- Address multicollinearity using Variance Inflation Factors (VIF) and understand its impact on model stability.
- Apply variable selection methods and model building strategies to datasets and case studies using R.

2.1 Introduction

In our discussion on regression problems thus far, we have operated under the assumption that the variables involved in the equation were predetermined. Our analysis focused on verifying the correctness of the functional specification and the validity of the error term assumptions. This presumption entailed that the selection of variables to be included in the equation had already been decided.

However, in many practical applications of regression analysis, the set of variables to be included in the regression model is not established beforehand. Often, the initial step in the analysis is to determine which variables should be included. Sometimes, theoretical or other considerations dictate the variables that must be included, in which case the issue of variable selection does not arise. Nevertheless, in instances where there is no definitive theory, the task of selecting variables for a regression equation becomes crucial.

The processes of variable selection and functional specification of the equation are interconnected. While formulating a regression model, two key questions must be addressed: which variables should be included, and in what form should they be incorporated—should they enter the equation as the original variable X, or as a transformed variable such as X^2, or $\sqrt{X}$ or $\log X$, or a combination of any one of these?

Although, in an ideal scenario, both problems would be solved simultaneously, for simplicity, we propose addressing them sequentially. First, we determine the variables to be included in the equation, followed by an investigation into the exact form these variables should take. This approach simplifies the variable selection problem and makes it more manageable. Once the variables to be included have been selected, we can apply the methods described in the earlier chapters to derive the actual form of the equation.

Variable selection plays an integral role in multiple disciplines:

- **Finance:** Choosing the right economic indicators for stock market predictions.
- **Healthcare:** Identifying significant risk factors in disease prediction models.
- **Engineering:** Optimizing process variables for system performance improvement.
- **Machine Learning:** Reducing dimensionality to enhance algorithm efficiency.

This unit will explore different techniques and considerations for effective variable selection in regression analysis.

2.2 Formulation of the Problem

Variable selection involves choosing a subset of predictor variables that best explain the response variable while maintaining the model's interpretability. Consider the general linear model:

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j \, x_{ij} + \varepsilon_i$$

where:

- y_i is the dependent variable,
- x_{ij} represents the predictor variables,
- β_j are the regression coefficients,
- ε_i is the random error term.

The key challenge is determining which subset of predictor variables ($X_1, X_2, \ldots, X_p$) should be included while avoiding unnecessary complexity.

2.3 Effects of Including or Excluding Variables

Including irrelevant variables: Adding irrelevant variables to a model can inflate the variance of parameter estimates without reducing bias. Analysts may sometimes include variables in an attempt to make the model more comprehensive, but these variables may not have a significant relationship with the outcome. While they might increase the model's explanatory power superficially, they contribute little to improving its accuracy. This can reduce the degrees of freedom ($n - k$) and undermine the model's reliability. For example, although the coefficient of determination (R^2) might increase, this could be misleading, as it might reflect the model's ability to fit the noise in the data rather than any true underlying pattern. Such inclusion risks overfitting, where the model captures random fluctuations instead of genuine trends, thus impairing its predictive ability. Therefore, careful selection of variables based on both theoretical relevance and empirical evidence is essential for maintaining a robust, reliable model.

Excluding relevant variables: Omitting important predictor variables can introduce bias into the model's parameter estimates and predictions. In the effort to simplify the model, analysts may exclude variables that have theoretical significance or are crucial for

explaining the outcome. This could be due to practical challenges, such as difficulties in quantifying intangible variables like personal preferences or in accurately measuring complex factors like income. However, excluding these relevant variables can distort the model's accuracy, leading to biased predictions and unreliable conclusions.

The key challenge in variable selection is balancing the trade-off between bias and variance. Including too many irrelevant variables increases variance, while excluding important ones introduces bias, both of which can undermine the quality of the model's predictions.

2.4 Criteria for Evaluating Reduced model

To ensure a regression model's adequacy, we use various criteria to ensure it fits the data well and generalizes effectively to new data.

When selecting subsets of candidate variables for the model, the challenge is to determine which subset yields the best regression model. Numerous criteria have been proposed in the literature to evaluate and compare these subset regression models effectively.

Here are some key criteria:

2.4.1 Coefficient of Determination

The **coefficient of determination** (R^2) is commonly used in variable selection to assess the model's explanatory power. Given k predictor variables, when you select a subset of $p-1$ variables, there are $\binom{k}{p-1}$ possible combinations, each corresponding to a unique model. The value of R^2 typically increases as more variables are added. Here's how to use R^2 for variable selection:

- Start by choosing an initial number of variables (p), fit the model, and calculate R_p^2.
- Add one more variable, refit the model, and calculate the new R_{p+1}^2.
- Since R_{p+1}^2 is usually greater than R_p^2, if the change $R_{p+1}^2 - R_p^2$ is small, stop and select the current value of p as the optimal number of variables for the subset regression.
- If the increase $R_{p+1}^2 - R_p^2$ is significant, continue adding variables until the increment in R_p^2 becomes marginal.

To determine the best value for p, you can plot R^2 against p. This graph will visually highlight the point where further additions to the model no longer yield significant improvements in the explanatory power.

2.4.2 Adjusted coefficient of determination

The **adjusted coefficient of determination** (R^2_{adj}) offers certain advantages over the standard coefficient of determination (R^2). It is calculated using the formula:

$$R^2_{\text{adj}}(p) = 1 - \left(\frac{n-1}{n-p-1}\right)\left(1 - R^2(p)\right)$$

Where:

- $R^2(p)$ is the coefficient of determination for the p-term model,
- n is the number of data points,
- p is the number of predictors in the model.

The key benefit of R^2_{adj} is that it doesn't automatically increase with the addition of more variables.

If additional predictor variables are added to the model, R^2_{adj} will only increase if the inclusion of those variables significantly improves the model's fit. Specifically, the adjusted R^2 will increase if the partial F-statistic for testing the significance of these new variables is greater than 1.

Thus, subset selection using R^2_{adj} can be done similarly to the standard R^2 approach. In general, the optimal number of variables is chosen as the one that maximizes R^2_{adj}, ensuring a balance between model complexity and explanatory power.

2.4.3 Residual Mean Square (RMS)

The Residual Mean Square (RMS) measures how well the regression model fits the observed data. It is calculated as follows:

$$RMSp = \frac{SSEp}{n-p}$$

where:

- $SSEp$: Sum of Squared Errors for the model with p predictors.
- n: Number of observations.
- p: Number of predictors in the model.

A lower RMS value indicates a better fit, as it signifies that the model's predictions are closer to the actual data points.

2.4.4 Mallows' Cp Statistic

Mallows' Cp Statistic helps balance the trade-off between the bias and variance of the model. It assesses whether a model is underfitting or overfitting the data:

$$C_p = \frac{SSEp}{\hat{\sigma}^2} + 2p - n$$

where:

- $SSEp$: Sum of Squared Errors for the model with p predictors.
- $\hat{\sigma}^2$: An estimate of the error variance based on the full model.
- p: Number of predictors in the model.
- n: Number of observations.

A good model will have a Cp value close to p. If Cp is significantly larger than p, it suggests that the model is overfitting. Conversely, if it is much smaller, the model may be underfitting.

2.4.5 Information Criteria (AIC & BIC)

Information criteria are used to compare models and select the best one based on their goodness of fit and complexity.

- **Akaike Information Criterion (AIC):**

$$AICp = n\ln\left(\frac{SSEp}{n}\right) + 2p$$

  where:

  - n: Number of observations.
  - $SSEp$: Sum of Squared Errors for the model with p variables.
  - p: Number of predictors in the model.

  The AIC helps to balance model fit and complexity. A lower AIC value indicates a model that fits the data well without being overly complex.

- **Bayesian Information Criterion (BIC):**

$$BICp = n\ln\left(\frac{SSEp}{n}\right) + p\ln(n)$$

  where:

  - n: Number of observations.
  - $SSEp$: Sum of Squared Errors.
  - p: Number of predictors.

- $\ln(n)$: Logarithm of sample size, which penalizes model complexity.

The BIC also balances fit and complexity, but it penalizes complexity more strongly than the AIC. Lower BIC values indicate better models.

In summary, these criteria ensure that the regression model you choose not only fits your current data but also generalizes well to new data, thereby achieving an optimal balance between fit and simplicity.

Tips

- A lower AIC/BIC score indicates a more efficient model.
- Comparing models: If a reduced model has a much lower AIC/BIC than the full model, it may be preferable. Running this code will print the AIC and BIC values for full-model, allowing you to compare it with other models. If you're comparing multiple models, choose the one with the lowest AIC or BIC value.

2.5 Variable Selection Procedures

To select the most relevant variables for a regression model, there are several variable selection techniques that can be employed. These techniques help identify the independent variables that contribute most significantly to the explanatory power of the model. Four common approaches include stepwise regression, forward selection, backward elimination, and best-subsets regression. Each of these techniques has its own strengths and application context.

The first three techniques—stepwise regression, forward selection, and backward elimination—are iterative processes. In these methods, variables are added or removed one at a time, with each adjustment being evaluated based on its effect on model performance. The process continues until a predefined stopping criterion is met, which indicates that no further improvements can be made to the model. Here's a brief overview of each:

- **Forward Selection**: This technique starts with no variables in the model and adds variables one by one based on their contribution to improving model fit. The variables are chosen based on criteria like p-values or AIC, and the process continues until no further improvement is observed.
- **Backward Elimination**: Unlike forward selection, backward elimination starts with all candidate variables in the model and removes the least significant variables one at a time. The decision to remove a variable is based on statistical tests (e.g., p-values), and the process stops when all remaining variables are significant.
- **Stepwise Regression**: This method is a combination of forward selection and backward elimination. It starts with either an empty model (for forward selection) or

a full model (for backward elimination) and then iteratively adds or removes variables based on their statistical significance. It can move in both directions, making it more flexible in finding the optimal model.

The fourth technique, **best-subsets regression**, differs from the others in that it does not consider variables one at a time. Instead, it evaluates all possible combinations of explanatory variables, identifying the subset that provides the best model fit. This approach is more exhaustive and can provide a more comprehensive view of the data, but it can also be computationally expensive, especially with large datasets.

These variable selection methods are especially valuable during the initial stages of model building, helping analysts narrow down the list of potential variables and choose those that contribute most effectively to the regression model. However, it's important to note that these techniques are not a replacement for an analyst's expertise. While they can provide guidance on which variables to include, the final decision should always be informed by the analyst's understanding of the data and its context.

In summary, variable selection techniques like forward selection, backward elimination, stepwise regression, and best-subsets regression can significantly enhance the process of building effective regression models. Each technique has its own strengths, and selecting the appropriate one depends on the specific context of the analysis and the nature of the data.

2.6 Collinearity and Variable Selection

When discussing variable selection procedures, we differentiate between two primary scenarios:

- **Non-Collinear Predictor Variables:** There is no strong evidence of collinearity among the predictor variables.
- **Collinear Predictor Variables:** The data exhibits high multicollinearity, indicating strong collinearity among the predictor variables.

The approach to variable selection depends on the correlation structure of the predictor variables. If the data analyzed are non-collinear, we proceed in one manner; if they are collinear, we proceed in another.

As an initial step in the variable selection procedure, we recommend calculating the Variance Inflation Factors (VIFs). VIF measures the extent to which the variance of a regression coefficient is inflated due to collinearity among the predictor variables. If none of the VIFs exceed 10, collinearity is not a concern.

General Rule of Thumb:

- A VIF of 1 indicates no correlation between a predictor and any other predictors.
- A VIF between 1 and 5 suggests moderate correlation that is usually acceptable.
- A VIF above 10 often indicates high collinearity, which could be problematic.

Collinearity can significantly impact the stability and interpretability of a regression model. When collinearity is detected, variable selection techniques such as ridge regression, principal component analysis (PCA), or partial least squares (PLS) can be employed to mitigate its effects. These techniques help in reducing multicollinearity by transforming the predictor variables or by introducing regularization to the regression model.

In summary, understanding and addressing collinearity is crucial for building robust and interpretable regression models. By carefully selecting and evaluating predictor variables, we can ensure the reliability and effectiveness of our regression analysis.

2.7 Analysis of Factors Influencing Employee Job Satisfaction

The following dataset offers valuable insights into the factors influencing overall job satisfaction among employees. It includes responses from a survey conducted within a manufacturing company, where employees rated various aspects of their job on a scale of 0 to 100. Higher values indicate better satisfaction or performance in each category. The data can be used to analyze how different management practices and workplace conditions impact employee satisfaction.

To explore the relationships between different variables, we will simulate a regression equation for overall job satisfaction (Y), using several influencing predictor variables (X1 to X6). This will help demonstrate variable selection procedures in a noncollinear context and identify which predictor contribute most to overall employee satisfaction.

Y	X1	X2	X3	X4	X5	X6	Y	X1	X2	X3	X4	X5	X6
43	55	49	44	54	49	34	63	64	51	54	63	73	47
71	75	50	55	70	66	41	65	70	46	57	75	85	46
78	75	58	74	80	78	49	67	61	45	47	62	80	41
81	78	56	66	71	83	47	74	85	64	69	79	79	63
82	82	39	59	64	78	39	69	62	57	42	55	63	25
66	77	66	63	88	76	72	64	53	53	58	58	67	34
50	40	33	34	43	64	33	61	63	45	47	54	84	35

Y	X1	X2	X3	X4	X5	X6	Y	X1	X2	X3	X4	X5	X6
67	60	47	39	59	74	41	68	83	83	45	59	77	35
85	85	71	71	77	74	55	72	82	72	67	71	83	31
81	90	50	72	60	54	36	74	85	64	69	79	79	63
53	66	52	50	63	80	37	65	60	65	75	55	80	60
50	58	68	54	64	78	52	48	57	44	45	51	83	38
71	70	68	69	76	86	48	63	54	42	48	66	75	33
40	37	42	58	50	57	49	58	67	42	56	66	68	35
43	51	30	39	61	92	45	77	77	54	72	79	77	46

2.7.1 Variable Description:

- **Y**: **Overall Job Satisfaction** – A measure of the employee's overall satisfaction with their job and the company's management, rated on a scale from 0 to 100.
- **X1**: **Supervisor's Communication Skills** – The effectiveness of the supervisor's communication with employees, rated on a scale from 0 to 100.
- **X2**: **Work-Life Balance** – The employee's satisfaction with the balance between their work responsibilities and personal life, rated on a scale from 0 to 100.
- **X3**: **Team Collaboration** – The level of teamwork and cooperation among employees, measured on a scale from 0 to 100.
- **X4**: **Career Development Opportunities** – The opportunities provided by the organization for career growth and advancement, rated on a scale from 0 to 100.
- **X5**: **Management's Fairness** – The fairness with which management handles employee relations, promotions, and job assignments, rated on a scale from 0 to 100.
- **X6**: **Employee Recognition** – The degree to which employees feel their efforts are recognized and appreciated by the organization, rated on a scale from 0 to 100.

2.7.2 Set up a regression model

Given the dataset we have, we can set up a **multiple linear regression model** to predict the overall job satisfaction (Y) based on the predictor variables (X1 to X6).

The regression equation would look like this:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \varepsilon$$

Where:

- Y is the overall job satisfaction score (dependent variable).
- $X_1, X_2, \ldots, X_6$ are the predictor variables (supervisor's communication skills, work-life balance, team collaboration, etc.).
- β_0 is the intercept term.
- $\beta_1, \ldots, \beta_6$ are the coefficients for each predictor variable.
- ε is the error term.

```
library(olsrr)

# Create the dataset
data <- data.frame(
  Y = c(43, 71, 78, 81, 82, 66, 50, 67, 85, 81, 53, 50, 71, 40, 43, 63, 65, 6
7, 74, 69, 64, 61, 68, 72, 74, 65, 48, 63, 58, 77),
  X1 = c(55, 75, 75, 78, 82, 77, 40, 60, 85, 90, 66, 58, 70, 37, 51, 64, 70,
61, 85, 62, 53, 63, 83, 82, 85, 60, 57, 54, 67, 77),
  X2 = c(49, 50, 58, 56, 39, 66, 33, 47, 71, 50, 52, 68, 68, 42, 30, 51, 46,
45, 64, 57, 53, 45, 83, 72, 64, 65, 44, 42, 42, 54),
  X3 = c(44, 55, 74, 66, 59, 63, 34, 39, 71, 72, 50, 54, 69, 58, 39, 54, 57,
47, 69, 42, 69, 47, 45, 67, 69, 75, 45, 48, 56, 72),
  X4 = c(54, 70, 80, 71, 64, 88, 43, 59, 77, 60, 63, 64, 76, 50, 61, 63, 75,
62, 79, 55, 58, 54, 59, 71, 79, 55, 51, 66, 66, 79),
  X5 = c(49, 66, 78, 83, 78, 76, 64, 74, 74, 54, 80, 78, 86, 57, 92, 73, 85,
80, 79, 63, 67, 68, 77, 78, 79, 80, 83, 75, 68, 77),
  X6 = c(34, 41, 49, 47, 39, 72, 33, 41, 55, 36, 37, 52, 48, 49, 45, 47, 46,
41, 63, 25, 34, 35, 35, 31, 63, 60, 38, 33, 35, 46)
)

# Fit the full model, i.e. include all predictor variables
full_model <- lm(Y ~ ., data = data)
ols_regress(full_model)
```

2.7.3 Assess Collinearity and Variance Inflation Factor (VIF)

In a **noncollinear** situation, the predictors should not have high multicollinearity. We can check this by calculating the **Variance Inflation Factor (VIF)** for each predictor. Ideally, VIF values should be below 5 or 10 to indicate that there is no strong collinearity between predictors.

Now, let's compute the **Variance Inflation Factor (VIF)** to check if there is any multicollinearity among the predictors.

```r
# Check multicollinearity with VIF
vif_values <- ols_vif_tol(full_model)
print(vif_values)
```

Interpretation of the Results

Variable	Tolerance	VIF	Interpretation
X1	0.3705	2.699	Moderate multicollinearity
X2	0.6046	1.654	Low multicollinearity
X3	0.4695	2.130	Moderate multicollinearity
X4	0.2929	3.414	Higher multicollinearity
X5	0.7377	1.355	Low multicollinearity
X6	0.4998	2.001	Moderate multicollinearity

The **tolerance** and **Variance Inflation Factor (VIF)** are both measures used to assess multicollinearity in a regression model, which occurs when independent variables are highly correlated with one another. Here's how to interpret the results:

Tolerance

Tolerance is the inverse of the VIF and is calculated as Tolerance $= 1 - R^2$, where R^2 is the squared multiple correlation coefficient between a given independent variable and all the other variables in the model. A tolerance value close to 1 indicates low multicollinearity, while a value close to 0 suggests high multicollinearity.

Overall Interpretation

The VIF values for all variables are well below the common threshold of 10, which suggests that multicollinearity is not a major concern in this model. However, X4 shows a slightly higher VIF (3.414), indicating it has a bit more correlation with other predictors compared to the others, but it still doesn't pose a significant issue. Generally, the model appears to have acceptable levels of multicollinearity, and there is no need for major adjustments.

But if you decide to keep the variables having VIF < 3 then you can discard X_4 and recheck the VIF values.

```r
# Remove X6 and fit the model again
reduced_data <- data[, c("Y", "X1", "X2", "X3", "X5", "X6")]  # Remove X4

# Fit the new model
reduced_model <- lm(Y ~ X1 + X2 + X3 + X5 + X6, data = reduced_data)
```

```r
# Calculate VIF again
vif_values_reduced <- ols_vif_tol(reduced_model)
print(vif_values_reduced)
```

After removing **X4**, the multicollinearity has decreased, and the remaining variables now exhibit **low to moderate multicollinearity**. Here's a detailed interpretation of the results:

Variable	Tolerance	VIF	Interpretation
X1	0.5319	1.880	Low multicollinearity
X2	0.6089	1.642	Low multicollinearity
X3	0.4932	2.028	Moderate multicollinearity
X5	0.8238	1.214	Low multicollinearity
X6	0.6088	1.643	Low multicollinearity

By removing **X4**, the model's performance should improve. The variables remaining in the model exhibit low to moderate multicollinearity, which enhances the **stability**, **reliability**, and **interpretability** of the results. This makes the model better suited for generating **reliable predictions**.

2.7.4 Apply Variable Selection Procedures

Here are the steps for **variable selection** using methods like **Forward Selection**, **Backward Elimination**, **Stepwise Selection**, and **Best-Subsets Selection**.

Here's how we can utilize **Forward Selection** in R to carry out variable selection for our model:

```r
# Perform forward selection with a p-value threshold of 0.3
forward_selection <- ols_step_forward_p(
  model = full_model,
  p_val = 0.3,
  progress = TRUE,    # Show progress during the selection
  details = TRUE      # Show detailed information about each step
)
# selection metrics
forward_selection$metrics
```

Stepwise Forward Selection Results Interpretation:

Step	Variable	R^2	Adjusted R^2	AIC	Mallows C_p	RMSE	Interpretation
1	X1	0.6863	0.6751	205.87	2.2842	6.7678	Initial model with variable X1 shows decent fit with moderate error.
2	X3	0.7161	0.6950	204.88	1.6046	6.4393	Adding variable X3 improves model fit, reduces AIC, and decreases RMSE.
3	X6	0.7363	0.7059	204.66	1.7783	6.2054	Including variable X6 further enhances model fit, with lowest AIC and RMSE.

Overall Interpretation:

- **Model Fit (R^2 and Adjusted R^2)**:
  - The R^2 values increase with each step, meaning that the model explains more of the variation in the data as more variables are added.
  - The **adjusted** R^2 also increases, showing that the additional variables contribute positively to the model fit and do not lead to overfitting. This suggests a balanced improvement in model complexity and accuracy.
- **AIC (Akaike Information Criterion)**:
  - The **AIC** decreases with each step, which is a sign of improving model fit. Lower AIC values indicate a better model that is more likely to generalize well to unseen data, as it penalizes model complexity.
- **SBC (Schwarz Bayesian Criterion) and SBIC (Scaled BIC)**:
  - Both **SBC** and **SBIC** show a decrease in value through the steps. These criteria balance model fit with model complexity, and the fact that they continue to decline suggests that the added variables help improve the model without unnecessarily increasing its complexity.
- **Mallows' C_p**:
  - Mallows' C_p values are below or close to the number of predictors plus one (p+1), indicating that the model is well-specified and does not suffer from

overfitting. A low C_p suggests that the added variables contribute to explaining the data effectively.

- **RMSE (Root Mean Square Error)**:
    - The **RMSE** decreases with each added variable, which indicates that the model's predictions are becoming more accurate as more information is included.

Stepwise Selection of Terms:

- **Step 1**: The model with **X1** explains 68.63% of the variance in the data, and although the model is relatively good, the RMSE is still relatively high, indicating some prediction error.
- **Step 2**: The inclusion of **X3** improves the model, as indicated by a decrease in **AIC**, **SBC**, and **RMSE**. The R^2 and adjusted R^2 values increase, showing better fit and explanatory power.
- **Step 3**: Adding **X6** further improves the model, achieving the lowest **AIC** and **RMSE**, and further increasing the model's R^2 and adjusted R^2 values.

***Backward Elimination**:*

In backward elimination, you start with all the predictors and remove the least useful ones. Here's how we can use backward elimination in R to perform variable selection for our model:

```
# Perform backward stepwise selection with a p-value threshold of 0.3
backward_selection <- ols_step_backward_p(
  model = full_model,     # The full model you want to start from
  p_val = 0.3,            # P-value threshold for inclusion
  progress = TRUE,        # Show progress during the selection
  details = TRUE          # Show detailed information about each step
)
backward_selection$metrics
```

The following summary table with a column stating why each variable was removed in each iteration:

Step	Variable	R^2	Adjusted R^2	AIC	SBC	Mallows' Cp	RMSE	Reason for Removal
1	X4	0.7442	0.6909	207.75	217.56	126.04	6.11	X4 was removed because it had the smallest t-test, indicating the least contribution to the reduction of error sum of squares.

Step	Variable	R^2	Adjusted R^2	AIC	SBC	Mallows' Cp	RMSE	Reason for Removal
2	X5	0.7409	0.6994	206.14	214.55	123.68	6.15	X5 was removed next as it had the smallest insignificant t-test in the new model.
3	X2	0.7363	0.7059	204.6613	211.6673	121.48	6.21	X2 was removed because it was the last variable with the smallest insignificant t-test in the final model.

In each step, the variable with the smallest t-test was removed, as it contributed the least to reducing the error sum of squares. The process continued until all remaining variables had significant t-tests.

Stepwise Selection

Stepwise selection combines both forward selection and backward elimination.

- Start with no predictors (or all predictors).
- Add predictors (like forward selection) and remove predictors (like backward elimination) based on p-values.
- The algorithm moves forward or backward to find the best-fitting model based on some criteria (e.g., AIC, BIC).

Here's how we can use **Stepwise Selection** (forward or backward) in R to perform variable selection for our model:

```
# Perform stepwise selection with both forward and backward steps
stepwise_selection <- ols_step_both_p(
  model = full_model,     # The initial full model
  p_enter = 0.2,     # variables with p value less than p_enter will enter int
o the model.
  p_remove = 0.3,    # variables with p more than p_remove will be removed fro
m the model.
  progress = TRUE,        # Show progress during the selection process
  details = TRUE          # Show detailed information about each step
)
```

The code for **Best-Subsets Selection** is as follows:

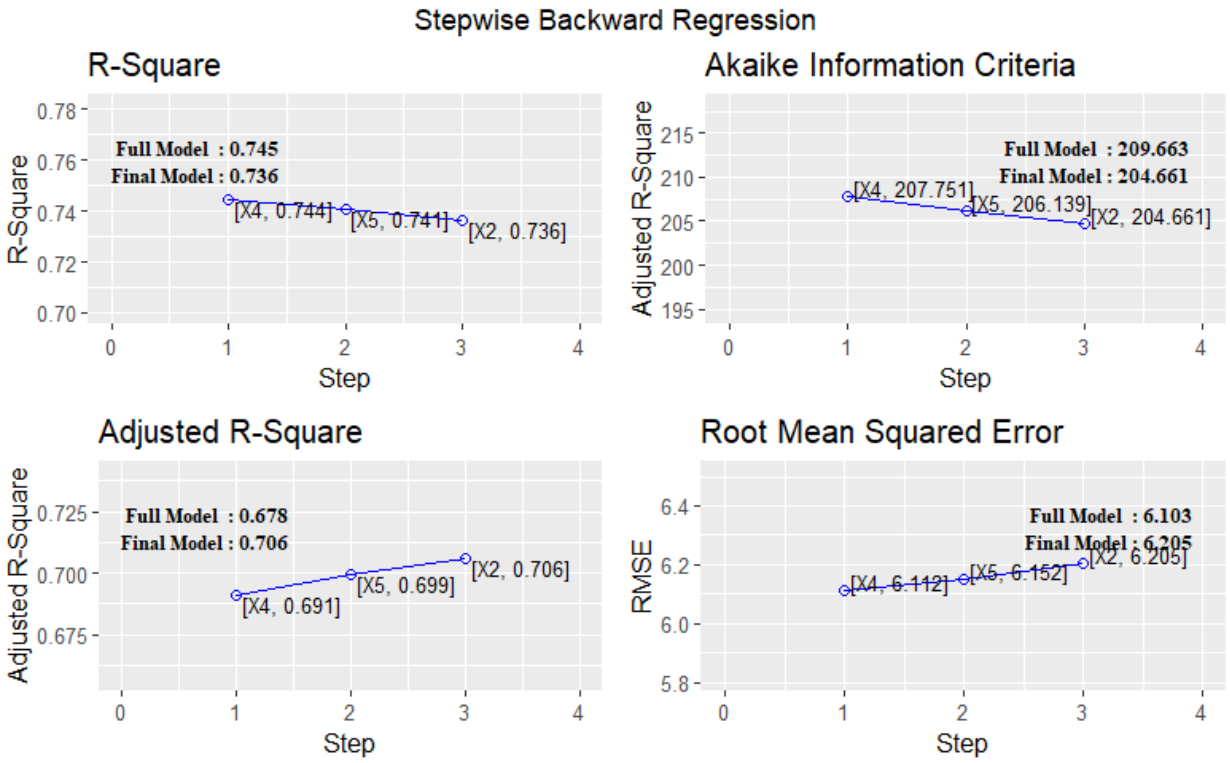
```
# Perform best-subsets regression using Mallows' Cp as the selection metric
best_subset_cp <- ols_step_best_subset(
  model = full_model,     # The initial full model
```

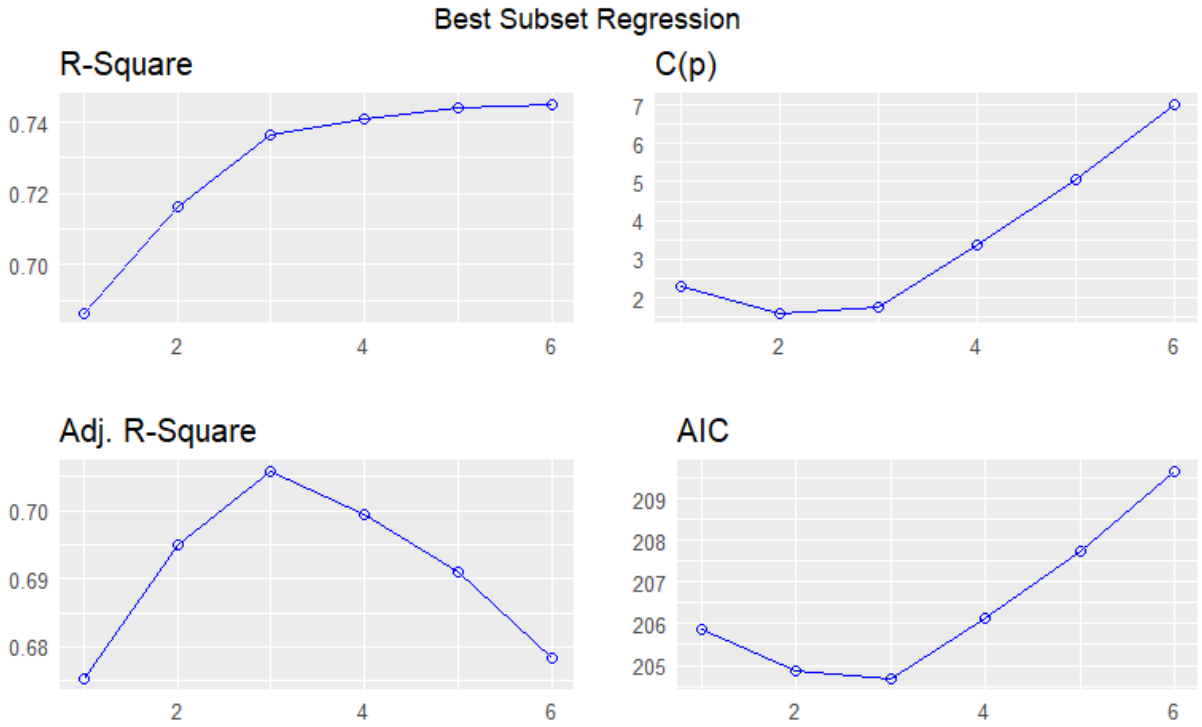
```
  metric = "cp",          # Use Mallows' Cp as the metric for selection
)
best_subset_cp$metrics
```

Plot the Model

If you want to visualize the process or the results, you can plot the selection process:

```
# Plot the stepwise selection
plot(stepwise_selection)
# Plot the best subset selection
plot(best_subset_cp)
```





2.7.5 Final Thoughts

By utilizing the regression model and variable selection procedures, we can identify the key qualities or predictors that most significantly contribute to overall job satisfaction among employees. The final model highlights critical factors such as supervisor communication skills, team collaboration, and employee recognition, among others. These insights provide valuable guidance for organizations, enabling them to focus on the aspects that are most influential in enhancing employee satisfaction. Through the variable selection process, we find that these factors, when combined, create a robust model with improved fit, reduced prediction errors, and minimal overfitting. While the model demonstrates strong predictive power, it is essential to consider the balance between accuracy and simplicity. The increasing model complexity, as reflected in metrics like SBC and SBIC, emphasizes the need to avoid overfitting. Ultimately, the final model offers a reliable foundation for understanding and improving job satisfaction. However, careful attention must be paid to ensure its real-world applicability and sustainability.

Check Your Progress – 1

The following data represents a company that markets products across various regional zones, each linked to a specific account manager. A regression analysis was performed to assess whether several predictor (independent) variables could account for the sales figures in each zone. A random sample of 25 regional zones led to the data as shown in Table, with variable definitions provided as below.

Variable	Definition
Revenue (Y)	Total sales attributed to the account manager
Tenure (X1)	Duration of employment, measured in months
MarketSize (X2)	Total potential market; industry-wide sales in units for the regional zone
AdSpend (X3)	Advertising expenditure for the regional zone
ShareOfMarket (X4)	Market share; average over the past four years
ShareChange (X5)	Change in market share over the last four years
Clients (X6)	Number of clients assigned to the account manager
WorkIndex (X7)	Workload; an index based on annual purchase volume and client distribution
Performance (X8)	Account manager's overall rating across eight performance areas, rated on a 1–7 scale

Y	X1	X2	X3	X4	X5	X6	X7	X8
3669.88	43.10	74065.1	4582.9	2.51	0.34	74.86	15.05	4.9
3473.95	108.13	58117.3	5539.8	5.51	0.15	107.32	19.97	5.1
2295.1	13.82	21118.5	2950.4	10.91	-0.72	96.75	17.34	2.9
4675.56	186.18	68521.3	2243.1	8.27	0.17	195.12	13.4	3.4
6125.96	161.79	57805.1	7747.1	9.15	0.50	180.44	17.64	4.6
2134.94	8.94	37806.9	402.4	5.51	0.15	104.88	16.22	4.5
5031.66	365.04	50935.3	3140.6	8.54	0.55	256.1	18.8	4.6
3367.45	220.32	35602.1	2086.2	7.07	-0.49	126.83	19.86	2.3
6519.45	127.64	46176.8	8846.2	12.54	1.24	203.25	17.42	4.9
4876.37	105.69	42053.2	5673.1	8.85	0.31	119.51	21.41	2.8
2468.27	57.72	36829.7	2761.8	5.38	0.37	116.26	16.32	3.1
2533.31	23.58	33612.7	1991.8	5.43	-0.65	142.28	14.51	4.2
2408.11	13.82	21412.8	1971.5	8.48	0.64	89.43	19.35	4.3
2337.38	13.82	20416.9	1737.4	7.80	1.01	84.55	20.02	4.2
4586.95	86.99	36272	10694.2	10.34	0.11	119.51	15.26	5.5

2729.24	165.85	23093.3	8618.6	5.15	0.04	80.49	15.87	3.6
3289.40	116.26	26878.6	7747.9	6.64	0.68	136.58	7.81	3.4
2800.78	42.28	39572	4565.8	5.45	0.66	78.86	16.00	4.2
3264.20	52.84	51866.1	6022.7	6.31	-0.1	136.58	17.44	3.6
3453.62	165.04	58749.8	3721.1	6.35	-0.03	138.21	17.98	3.1
1741.45	10.57	23990.8	861	7.37	-1.63	75.61	20.99	1.6
2035.75	13.82	25694.9	3571.5	8.39	-0.43	102.44	21.66	3.4
1578	8.13	23736.3	2845.5	5.15	0.04	76.42	21.46	2.7
4167.44	58.44	34314.3	5060.1	12.88	0.22	136.58	24.78	2.8
2799.97	21.14	22809.5	3552	9.14	-0.74	88.62	24.96	3.9

Analyze this dataset using the variable selection procedures discussed in this unit. Identify which predictor variables (such as Revenue, Experience, MarketSize, AdSpend, etc.) should be included in the regression model for predicting sales performance.

2.8 LET US SUM UP

This unit explores the essential process of variable selection in regression analysis, which involves choosing the most relevant predictor variables to create a robust and interpretable regression model. It emphasizes the importance of identifying the right variables and determining their functional form, with a focus on practical application in R. Key variable selection techniques—such as forward selection, backward elimination, stepwise selection, and best-subsets regression—are covered, along with methods for evaluating model performance through metrics like Residual Mean Square (RMS), Mallows' Cp, and Information Criteria (AIC and BIC). The unit also discusses the challenge of multicollinearity and how to detect and address it using tools like Variance Inflation Factors (VIF).

Understanding different selection techniques enables data-driven decision-making, improving predictions while maintaining interpretability. By carefully selecting variables, analysts can create robust and meaningful regression models.

2.9 Check Your Progress: Possible Answers

Check Your Progress – 1

- Use the **stepwise selection** technique (in R) to find the optimal set of predictor variables as discussed in this section.
- Based on your analysis, you may find that X2, X3, X4, and X6 are the most important predictors. Forward Selection and Backward Elimination may lead to different models.
- Ensure assessing the model fit and consider multicollinearity.

2.10 Further Reading

1. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023
2. https://home.iitk.ac.in/~shalab/regression/Chapter13-Regression-VariableSelectionAndModelBuilding.pdf?form=MG0AV3
3. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016

2.11 Assignment

1. Explain the importance of variable selection in regression analysis. How does it impact model performance and interpretability?
2. Discuss the effect of including of excluding variables.
3. Compare and contrast Forward Selection, Backward Elimination, and Stepwise Selection. Under what conditions is each method preferable?
4. What is multicollinearity? How does it affect regression estimates, and what techniques can be used to detect and address it?

Unit 3: Binary Logistic Regression

Unit Structure

3.0 Learning Objectives

By the end of this unit, you should be able to:

- Understand the concept of logistic regression and its applications.
- Differentiate between linear regression and logistic regression.
- Formulate the logistic regression equation and interpret its parameters.
- Estimate the probability of an event using logistic regression.
- Perform significance testing for logistic regression models.
- Interpret odds ratios and their implications in logistic regression.
- Apply logistic regression to classification task.

3.1 Introduction

Logistic regression is a statistical method used for binary classification problems, where the outcome variable can take only two possible values, such as “yes” or “no,” “success” or “failure,” or “1” and “0.” Unlike linear regression, which predicts a continuous outcome, logistic regression estimates the probability of an event occurring based on one or more independent variables.

Consider a scenario where a financial institution wants to determine whether a company is likely to go bankrupt within two years. The dependent variable (Y) is coded as 1 if the company remains solvent and 0 if it goes bankrupt. Predictor variables (X) may include financial ratios such as retained earnings to total assets, earnings before interest and taxes to total assets, and sales to total assets.

3.1.1 Applications of Logistic Regression

Logistic regression is widely used in various fields, including:

- **Marketing**: Predicting whether a customer will purchase a product.
- **Finance**: Assessing the likelihood of loan default.
- **Healthcare**: Predicting the probability of a patient having a disease.
- **Social Sciences**: Analyzing the likelihood of an individual voting for a particular candidate.

3.1.2 Key Differences Between Linear and Logistic Regression

- **Dependent Variable**: Linear regression predicts continuous outcomes, while logistic regression predicts binary outcomes.
- **Output**: Linear regression outputs a straight line, while logistic regression outputs an S-shaped curve (sigmoid function).
- **Equation**: Linear regression uses a linear equation, while logistic regression uses a logistic function.

3.2 FITTING THE LOGISTIC REGRESSION MODEL

Since logistic regression deals with a binary response variable, it models the probability that an observation belongs to one of the two categories. The linear regression model is not suitable for this task because it assumes an unbounded outcome, which is incompatible with probability values constrained between 0 and 1.

To resolve this, logistic regression uses the logistic response function:

$$\pi = P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}}$$

where:

- π represents the probability that $Y = 1$.
- $\beta_0, \beta_1, \ldots, \beta_p$ are the model parameters.
- $X_1, \ldots, X_p$ are independent variables.
- e is the base of the natural logarithm.

The significant difference between linear and logistic regression models lies in the conditional distribution of the outcome variable. In linear regression, we assume that the outcome variable for each observation can be expressed as a linear function of the independent variables plus an error term. This error term represents the deviation of the observed value from the conditional mean. Typically, we assume that the error term follows a normal distribution with a mean of zero and a constant variance, which is independent of the values of the independent variables. Consequently, the conditional distribution of the outcome variable, given the independent variables, is normal, with a constant variance across all levels of the predictors.

However, this assumption does not hold in the case of a dichotomous (binary) outcome variable, which is central to logistic regression. In logistic regression, the outcome variable can take on one of two values (typically 0 or 1), and we model the probability of one of these outcomes occurring, given the independent variables. Instead of assuming a normal distribution for the error term, we assume that the error term follows a *Bernoulli distribution*. Specifically, for each observation, the probability of the outcome being 1 is modeled as a logistic function of the independent variables.

If the outcome variable equals 1, the probability is π, and if it equals 0, the probability is $1 - \pi$, where π is the conditional probability of success (i.e., the probability that the outcome is 1). In this case, the conditional distribution of the outcome variable is *binomial*, with a mean equal to π and a variance of $\pi(1 - \pi)$, which depends on the predicted probability. The key distinction here is that, unlike in linear regression, the variance of the outcome in logistic regression is not constant and varies with the predicted probability, reflecting the underlying *binomial distribution*.

3.2.1 Logit Transformation

Rather than modeling probability directly, logistic regression transforms it using the logit function.

$$g(X) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

This transformation ensures a linear relationship between the predictor variables and the transformed outcome, enabling easier interpretation of coefficients.

3.2.2 Maximum Likelihood Estimation

The parameters $\beta_0, \beta_1, \ldots, \beta_p$ are estimated using Maximum Likelihood Estimation (MLE), which maximizes the probability of obtaining the observed data. The likelihood function for n independent observations is:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

Taking the logarithm, we obtain the log-likelihood function:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)]$$

Since the log-likelihood function is nonlinear in β, numerical optimization techniques such as Newton-Raphson or Fisher scoring are used to estimate the parameters.

3.3 Example: Big Bazar Stores Dataset

Let's consider a dataset from Big Bazaar Stores. The goal is to predict whether a customer will use a coupon based on their annual spending and whether they have a Big Bazaar credit card.

Variables:

- **Spending (X1):** Annual spending in thousands of rupees.
- **Card (X2):** 1 if the customer has a Big Bazaar credit card, 0 otherwise.
- **Redeemed (Y):** 1 if the customer used the coupon, 0 otherwise.

ID	X1	X2	Y	ID	X1	X2	Y	ID	X1	X2	Y
1	4.701	1	1	35	6.179	0	0	69	3.253	0	0
2	3.993	0	1	36	1.980	1	0	70	2.059	1	0
3	1.677	1	0	37	1.058	1	0	71	2.678	1	1
4	6.486	0	1	38	6.851	1	1	72	2.323	1	0
5	2.528	1	1	39	1.124	0	0	73	1.878	0	0
6	2.423	0	0	40	3.318	1	1	74	2.678	1	1
7	3.566	0	1	41	3.253	0	1	75	6.179	1	1
8	3.318	1	1	42	1.839	0	0	76	3.411	1	1
9	7.076	0	1	43	1.657	1	0	77	5.991	0	1
10	2.229	1	0	44	1.075	0	0	78	7.076	1	1
11	3.345	1	1	45	3.566	0	1	79	3.255	1	1
12	3.255	0	1	46	2.118	0	0	80	2.148	1	0
13	1.512	0	0	47	1.554	1	0	81	5.501	1	1

ID	X1	X2	Y	ID	X1	X2	Y	ID	X1	X2	Y
14	5.991	1	1	48	4.631	0	1	82	5.991	1	1
15	6.737	1	1	49	4.345	1	1	83	2.372	1	0
16	2.148	0	0	50	4.004	1	1	84	3.995	1	1
17	2.118	0	0	51	1.068	1	0	85	2.135	0	0
18	3.470	1	1	52	2.421	0	0	86	6.737	0	1
19	2.936	0	0	53	4.414	1	1	87	6.486	1	1
20	6.404	0	1	54	3.386	1	1	88	2.429	1	0
21	2.229	0	0	55	1.677	1	0	89	4.701	1	1
22	2.933	0	0	56	2.050	1	0	90	6.404	0	1
23	2.118	1	0	57	2.323	1	1	91	1.130	1	0
24	2.050	0	0	58	5.501	1	1	92	1.911	1	1
25	4.998	0	1	59	3.345	0	1	93	4.959	1	1
26	1.394	0	0	60	3.318	1	1	94	6.073	1	1
27	3.993	1	1	61	4.721	0	1	95	1.403	1	0
28	2.059	0	0	62	1.662	1	0	96	3.318	0	0
29	1.677	0	0	63	2.936	1	0	97	2.421	1	0
30	2.229	1	0	64	2.049	0	0	98	6.073	0	1
31	4.345	1	1	65	2.313	0	1	99	2.630	1	0
32	2.933	1	0	66	6.851	0	1	100	3.411	1	1
33	5.365	1	1	67	2.291	1	0				
34	5.365	0	0	68	3.470	1	1				

The variables in the study are defined as follows:

$$Y = \begin{cases} 1, & \text{if the customer redeem the coupon} \\ 0, & \text{if the customer did not redeem the coupon} \end{cases}$$

$$X_1 = \text{annual spending at Big Bazar Stores (₹1000s)}$$

$$X_2 = \begin{cases} 1, & \text{if the customer have Big Bazar credit card} \\ 0, & \text{if the customer does not have Big Bazar credit card} \end{cases}$$

Thus, we choose a logistic regression equation with two independent variables.

$$\pi = P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2)}}{1 + e^{(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

The following R code snippet estimates the model parameters β_0, β_1 and β_2.

```
# Read the data from csv file
data <- read.csv("Big_Bazar.csv")

# Fit the logistic regression model
```

```r
model <- glm(Redeemed ~ Spending + Card, data = data, family = binomial)

# Summarize the model
summary(model)
```

```
Call:
 glm(formula = Redeemed ~ Spending + Card, family = binomial,
     data = data)

 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -6.5605     1.3571  -4.834 1.34e-06 ***
 Spending      1.8708     0.3962   4.722 2.34e-06 ***
 Card          1.4729     0.6756   2.180   0.0292 *
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 138.269  on 99  degrees of freedom
 Residual deviance:  67.618  on 97  degrees of freedom
 AIC: 73.618

 Number of Fisher Scoring iterations: 6
```

3.3.1 Evaluation Metrics for Logistic Regression

Unlike linear regression, logistic regression does not have an equivalent to R^2 for goodness-of-fit assessment. Instead, the following approaches are used:

- **Likelihood Ratio Test**: Compares the likelihood of the fitted model to a null model. A significant test suggests that the independent variables improve prediction.
- **Deviance**: Defined as:
  $$D = -2[\ell(\text{saturated model}) - \ell(\text{fitted model})]$$
  A lower deviance value suggests a better fit. The difference in deviance between models follows a chi-square distribution and can be used for hypothesis testing.
- **Akaike Information Criterion (AIC)**: Measures model quality by balancing goodness-of-fit and complexity. Lower AIC values indicate a better trade-off.

The **log-likelihood** of the fitted model is related to deviance:

$$D = -2 \times \log(\text{Likelihood})$$

A lower deviance indicates a better model fit.

3.3.2 Testing for Significance

Overall Model Significance

To test the overall significance of the logistic regression model, we use the chi-square (χ^2) test. The null hypothesis is that all coefficients are zero, and the alternative hypothesis is that at least one coefficient is not zero.

To perform a chi-square test for the logistic regression model, we typically use the likelihood ratio test to compare nested models. Here's how you can do it in R:

```
# Fit the full logistic regression model
model_full <- glm(Redeemed ~ Spending + Card, data = data, family = binomi
al)

# Fit the null model (intercept only)
model_null <- glm(Redeemed ~ 1, data = data, family = binomial)

# Perform the chi-square test using the likelihood ratio test
anova(model_null, model_full, test = "Chisq")
```

Interpreting the Chi-Square Test Output

The *Analysis of Deviance Table* compares two models:

- **Model 1 (Null Model):** Includes only the intercept (i.e., assumes no predictors influence the outcome).
- **Model 2 (Full Model):** Includes `Spending` and `Card` as predictors.

Key Metrics in the Output

Metric	Model 1 (Null)	Model 2 (Full)	Interpretation
Residual Df	99	97	Degrees of freedom left after fitting the model.
Residual Dev	138.269	67.618	Lower deviance indicates a better fit.
Df	-	2	The number of added predictors (`Spending`, `Card`).
Deviance	-	70.651	Difference in deviance between models (measures improvement).
Pr(>Chi)	-	4.553e-16 ***	p-value for chi-square test.

Chi-Square Test Conclusion

- The chi-square statistic = **70.651** (from deviance).
- **p-value = 4.553e-16**, which is **much smaller than 0.05**, meaning the model with predictors significantly improves fit compared to the null model.

- **Significance Codes (***)**: The three stars indicate a highly significant result (**p < 0.001**).

Final Interpretation

The logistic regression model with *Spending* and *Card* is statistically significant and improves the prediction of *Redeemed* compared to the null model.

This suggests that ***at least one of the predictors has a meaningful impact on the probability of coupon redemption***.

Individual Variable Significance

For each independent variable, we test whether its coefficient is significantly different from zero using the Wald test. A low p-value (typically < 0.05) indicates that the variable is significant.

Wald Test

The Wald test evaluates the significance of each coefficient. It is computed as:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Coefficients and Odds Ratios

Each coefficient represents the log-odds change in the probability of coupon redemption per unit increase in the predictor variable.

Predictor	Estimate	Std. Error	Z-Value	P-Value	Interpretation
Intercept	-6.5605	1.3571	-4.834	1.34e-06 (***)	Baseline log-odds of redemption when Spending = 0 and Card = 0.
Spending	1.8708	0.3962	4.722	2.34e-06 (***)	Each additional $1000 spent increases log-odds of redemption by 1.87.
Card	1.4729	0.6756	2.180	0.0292 (*)	Customers with a Simmons card have higher odds of redeeming a coupon.

Each predictor's Wald statistic is compared to a standard normal distribution:

- Spending: **Wald = 1.8708 / 0.3962 = 4.722**
- Card: **Wald = 1.4729 / 0.6756 = 2.180**

Since both p-values are **< 0.05**, both variables significantly contribute to predicting coupon redemption.

Deviance and Log-Likelihood

- **Null Deviance = 138.269 (df = 99)** → Deviance of a model with only the intercept (no predictors).
- **Residual Deviance = 67.618 (df = 97)** → Deviance of the fitted model (with Spending and Card as predictors).
- **Reduction in Deviance = 138.269 - 67.618 = 70.651** → This significant reduction suggests the predictors improve model fit.

3.4 Interpretation of Parameters in a Logistic Regression Model

3.4.1 Logit Transformation in Logistic Regression

In logistic regression, the probability of an event occurring is modeled using the **logit function**, which transforms probabilities into log-odds:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where:

- p is the probability of the event occurring ($y = 1$),
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_k$ are regression coefficients for the independent variables $x_1, x_2, \ldots, x_k$.

Using the given logistic regression coefficients:

$$\text{logit}(p) = -6.5605 + 1.8708 \times \text{Spending} + 1.4729 \times \text{Card}$$

3.4.2 Definition of Odds and Odds Ratio

The **odds** in favor of an event occurring ($y = 1$) is defined as the probability that the event will occur divided by the probability that the event will not occur:

$$\text{Odds}(y = 1) = \frac{p}{1-p}$$

Since logistic regression models **log-odds**, we can express the probability p as:

$$p = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

The **odds ratio (OR)** measures the effect of a **one-unit increase** in an independent variable on the odds of $y = 1$. It is calculated as:

$$\text{OR} = \frac{\text{Odds}_1}{\text{Odds}_0}$$

where:

- **Odds**$_0$ represents the odds of $y = 1$ at a given set of independent variables.
- **Odds**$_1$ represents the odds of $y = 1$ when one independent variable is increased by one unit, keeping others constant.

Mathematically, for a one-unit increase in x_1, the new log-odds becomes:

$$\text{logit}(p_1) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_k x_k$$

The corresponding **odds** are:

$$\text{Odds}_1 = e^{\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \cdots + \beta_k x_k}$$

Similarly, the odds before increasing x_1 were:

$$\text{Odds}_0 = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

Thus, the **odds ratio (OR)** is:

$$\frac{\text{Odds}_1}{\text{Odds}_0} = \frac{e^{\beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \cdots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}$$

Since the terms in the numerator and denominator cancel out, we get:

$$\text{OR} = e^{\beta_1}$$

Interpreting the Odds Ratios in the Example

From the given coefficients:

- **Spending (**$\beta_1 = 1.8708$**)**

$$\text{OR} = e^{1.8708} \approx 6.49$$

Interpretation: A ₹1000 increase in spending increases the odds of redemption by **6.49 times**.

- **Card (**$\beta_2 = 1.4729$**)**

$$\text{OR} = e^{1.4729} \approx 4.36$$

Interpretation: Customers who have a Big Bazar card are **4.36 times more likely** to redeem a coupon than those who do not.

Final Remarks

- Odds ratio provides an intuitive measure of how a predictor influences the likelihood of an event occurring.
- Exponentiating a coefficient (e^{β}) allows us to move from log-odds to a meaningful odds ratio.
- A greater OR (>1) means an increased likelihood, while an OR < 1 would indicate a decreased likelihood.

3.5 Another Approach to Classification Problems

Logistic regression is a widely used technique to estimate the probability that an observation belongs to a particular group based on various predictor variables. The model generates fitted logits that can then be used to classify observations into one of two categories. However, if our primary goal is purely classification, other statistical methods may be worth considering. Discriminant analysis, for example, is often employed when the main focus is to predict the group membership of each observation.

While we won't explore discriminant analysis in detail here, it's useful to consider a simpler regression-based approach that can achieve the same goal. The core principle of discriminant analysis is to find a linear combination of predictor variables $(X_1, \ldots, X_p)$ that best distinguishes between two groups. This separation can be achieved through a multiple regression model, where the response variable (Y) takes values of 0 or 1, and the predictors are $(X_1, \ldots, X_p)$.

As mentioned earlier, some of the fitted values may fall outside the 0 to 1 range. This is not a concern in this context, since our aim is not to model probabilities but to classify observations. To classify observations, we compute the average of the predicted values. Any observation with a predicted value greater than this average is classified into the group with Y = 1, while observations with a predicted value below the average are assigned to the group with Y = 0. Finally, we evaluate the number of correctly classified observations in the sample. The selection of predictor variables in this approach follows the same process as in multiple regression.

Using the logistic regression equation, we can estimate the probability of a customer using the coupon. For example, if a customer spends ₹3000 annually and does not have a Big Bazar credit card, the probability of using the coupon can be estimated as:

$$P(Y = 1|X_1 = 3, X_2 = 0) = \frac{e^{-6.5605+1.8708(3)+1.4729(0)}}{1 + e^{-6.5605+1.8708(3)+1.4729(0)}} = 0.2793$$

Using R, we can easily compute the probabilities along with predictions.

```r
# Make predictions (probabilities of redemption)
predicted_probabilities <- predict(model, type = "response")

# Add predictions to the dataset
data$Predicted_Probability <- predicted_probabilities

# Make binary predictions based on a 0.5 cutoff
data$Predicted_Redeemed <- ifelse(predicted_probabilities > 0.5, 1, 0)

# Show the dataset with binary predictions
head(data)
```

```
  Customer Spending Card Redeemed Predicted_Probability Predicted_Redeemed
         1    4.701    1        1              0.9760434                  1
         2    3.993    0        1              0.7129541                  1
         3    1.677    1        0              0.1245346                  0
         4    6.486    0        1              0.9962183                  1
         5    2.528    1        1              0.4114238                  0
         6    2.423    0        0              0.1163506                  0
```

Logistic regression predicts *probabilities*, not direct class labels. To convert probabilities into a *binary classification*, we apply a **threshold (cutoff value)**, commonly set at **0.5**.

- If the *predicted probability* of redemption is *greater than 0.5*, we classify the customer as **Redeemed = 1** (they will redeem the coupon).
- If the predicted probability is *less than or equal to 0.5*, we classify them as **Redeemed = 0** (they will not redeem the coupon).

By applying this threshold, the model *converts a probability score into a discrete category*, making it usable for classification tasks.

The following table shows the observed Y, the Probability (Prob) and predicted class (Y_pred). The wrongly classified observations are shown in Bold face.

ID	Y	Prob	Y_Pred	ID	Y	Prob	Y_Pred	ID	Y	Prob	Y_Pred
1	1	0.9760	1	35	**0**	**0.9933**	**1**	69	0	0.3835	0
2	1	0.7129	1	36	0	0.2005	0	70	0	0.2252	0
3	0	0.1245	0	37	0	0.0428	0	71	**1**	**0.4806**	**0**
4	1	0.9962	1	38	1	0.9996	1	72	0	0.3227	0
5	**1**	**0.4114**	**0**	39	0	0.0115	0	73	0	0.0453	0
6	0	0.1164	0	40	1	0.7540	1	74	**1**	**0.4806**	**0**
7	1	0.5277	1	41	**1**	**0.3835**	**0**	75	1	0.9985	1
8	1	0.7540	1	42	0	0.0423	0	76	1	0.7848	1
9	1	0.9987	1	43	0	0.1205	0	77	1	0.9905	1
10	0	0.2855	0	44	0	0.0105	0	78	1	0.9997	1
11	1	0.7632	1	45	1	0.5277	1	79	1	0.7315	1
12	**1**	**0.3844**	**0**	46	0	0.0693	0	80	0	0.2556	0

ID	Y	Prob	Y_Pred	ID	Y	Prob	Y_Pred	ID	Y	Prob	Y_Pred
13	0	0.0234	0	47	0	0.1015	0	81	1	0.9945	1
14	1	0.9978	1	48	1	0.8912	1	82	1	0.9978	1
15	1	0.9995	1	49	1	0.9544	1	83	0	0.3430	0
16	0	0.0730	0	50	1	0.9171	1	84	1	0.9158	1
17	0	0.0693	0	51	0	0.0435	0	85	0	0.0713	0
18	1	0.8029	1	52	0	0.1160	0	86	1	0.9976	1
19	0	0.2558	0	53	1	0.9597	1	87	1	0.9991	1
20	1	0.9956	1	54	1	0.7768	1	88	0	0.3674	0
21	0	0.0839	0	55	0	0.1245	0	89	1	0.9760	1
22	0	0.2548	0	56	0	0.2223	0	90	1	0.9956	1
23	0	0.2451	0	57	**1**	**0.3227**	**0**	91	0	0.0486	0
24	0	0.0615	0	58	1	0.9945	1	92	1	0.1806	0
25	1	0.9421	1	59	**1**	**0.4249**	**0**	93	1	0.9851	1
26	0	0.0188	0	60	1	0.7540	1	94	1	0.9981	1
27	1	0.9155	1	61	1	0.9065	1	95	0	0.0785	0
28	0	0.0625	0	62	0	0.1215	0	96	0	0.4126	0
29	0	0.0316	0	63	**0**	**0.5999**	**1**	97	0	0.3640	0
30	0	0.2855	0	64	0	0.0614	0	98	1	0.9918	1
31	1	0.9544	1	65	**1**	**0.0968**	**0**	99	0	0.4583	0
32	**0**	**0.5986**	**1**	66	1	0.9981	1	100	1	0.7848	1
33	1	0.9930	1	67	0	0.3097	0				
34	**0**	**0.9700**	**1**	68	1	0.8029	1				

The R code snippet with output:

```
# False positives: 0 actual, predicted as 1
false_positives <- sum(data$Redeemed == 0 & data$Predicted_Redeemed == 1)
# False negatives: 1 actual, predicted as 0
false_negatives <- sum(data$Redeemed == 1 & data$Predicted_Redeemed == 0)
# Output the result
false_positives
```

```
## [1] 4
# Output the result
false_negatives
## [1] 9
# Create confusion matrix
confusion_matrix <- table(Predicted = data$Predicted_Redeemed, Actual = da
ta$Redeemed)
# Print confusion matrix
confusion_matrix
          Actual
 Predicted  0  1
         0 43  9
         1  4 44
```

Here's an interpretation of the provided code and its output:

1. **False Positives and False Negatives**:
    - **False Positives**: The code calculates the number of false positives, which are cases where the actual value is 0 (not redeemed), but the predicted value is 1 (predicted as redeemed). The result is 4.
    - **False Negatives**: The code calculates the number of false negatives, which are cases where the actual value is 1 (redeemed), but the predicted value is 0 (predicted as not redeemed). The result is 9.
2. **Confusion Matrix**:
    - The confusion matrix summarizes the performance of the classification model by comparing the actual and predicted values.
    - The matrix is structured as follows:

Predicted	Actual 0	Actual 1
0 (Not Redeemed)	43	9
1 (Redeemed)	4	44

- **Actual 0, Predicted 0**: There are 43 cases where the actual value is 0, and the model correctly predicted 0.
- **Actual 1, Predicted 1**: There are 44 cases where the actual value is 1, and the model correctly predicted 1.
- **Actual 0, Predicted 1**: There are 4 cases where the actual value is 0, but the model incorrectly predicted 1 (false positives).
- **Actual 1, Predicted 0**: There are 9 cases where the actual value is 1, but the model incorrectly predicted 0 (false negatives).

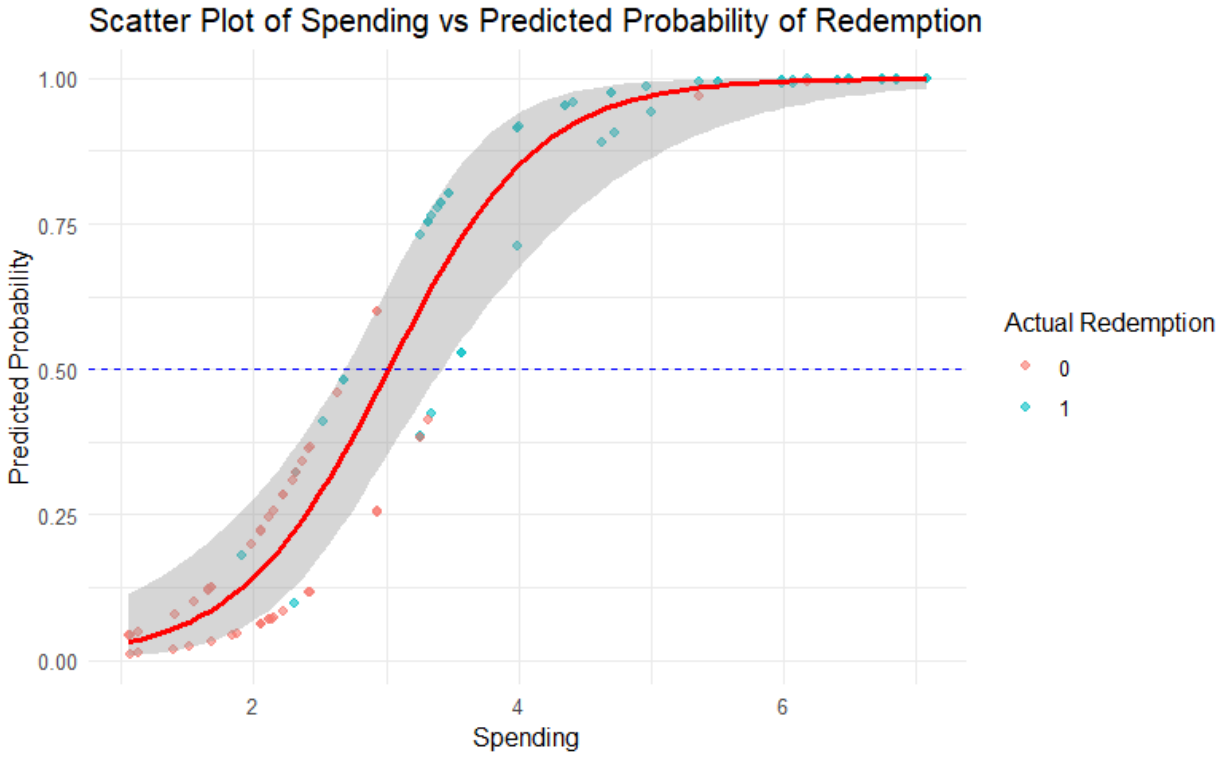
In summary, the model has 4 false positives and 9 false negatives. The confusion matrix provides a detailed breakdown of how well the model performed in predicting redeemed and not redeemed cases.

For the Big Bazar dataset presented in the above table, logistic regression outperforms multiple regression in classifying the sample data. Generally, this is true because logistic regression does not require the restrictive assumption of multivariate normality for the predictor variables. Therefore, we recommend using logistic regression for classification problems. If a logistic regression package is unavailable, the multiple regression approach can be considered as an alternative.

Let us visualize how well the predicted probabilities align with the actual redemption status.

```
library(ggplot2)

# Scatter plot with actual redemption status and logistic regression line
ggplot(data, aes(x = Spending, y = Predicted_Probability, color =
```

```
factor(Redeemed))) +
  geom_point(alpha = 0.6) + # Scatter points with transparency
  geom_smooth(method = "glm", method.args = list(family = "binomial"),
color = "red") + # Logistic regression curve
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "blue") + #
Classification threshold line at 0.5
  labs(
    title = "Scatter Plot of Spending vs Predicted Probability of
Redemption",
    x = "Spending",
    y = "Predicted Probability",
    color = "Actual Redemption"
  ) +
  theme_minimal()
```



3.6 The Multinomial Logit Model

Logistic regression can be extended to cases where the response variable has more than two categories. The multinomial logit model is used when these categories are nominal (unordered). The probability of an observation belonging to category j (relative to a reference category k) is given by:

$$ln\left(\frac{\pi_j}{\pi_k}\right) = \beta_{0j} + \beta_{1j}X_1 + \ldots + \beta_{pj}X_p$$

where each category except the reference category has a separate equation.

If the response categories have a natural order, the *ordinal logistic regression* (proportional odds model) can be used instead. This model assumes that the log-odds of being in a category or below it is proportional across different levels.

Check Your Progress – 1

1. The following data comprises operating financial ratios for 33 firms that went bankrupt within two years and 33 firms that remained solvent during the same period. A multiple logistic regression model is fitted using variables X1, X2, and X3. The response variable is defined as Y = 0 if the firm went bankrupt after 2 years, and Y = 1 if the firm remained solvent after 2 years. Three financial ratios were available for each firm:
   - X1 = Retained Earnings / Total Assets
   - X2 = Earnings Before Interest and Taxes / Total Assets
   - X3 = Sales / Total Assets.

Y	X1	X2	X3	Y	X1	X2	X3
0	-62.8	-89.5	1.7	1	43.0	16.4	1.3
0	3.3	-3.5	1.1	1	47.0	16.0	1.9
0	-120.8	-103.2	2.5	1	-3.3	4.0	2.7
0	-18.1	-28.8	1.1	1	35.0	20.8	1.9
0	-3.8	-50.6	0.9	1	46.7	12.6	0.9
0	-61.2	-56.2	1.7	1	20.8	12.5	2.4
0	-20.3	-17.4	1.0	1	33.0	23.6	1.5
0	-194.5	-25.8	0.5	1	26.1	10.4	2.1
0	20.8	-4.3	1.0	1	68.6	13.8	1.6
0	-106.1	-22.9	1.5	1	37.3	33.4	3.5
0	-39.4	-35.7	1.2	1	59.0	23.1	5.5
0	-164.1	-17.7	1.3	1	49.6	23.8	1.9
0	-308.9	-65.8	0.8	1	12.5	7.0	1.8
0	7.2	-22.6	2.0	1	37.3	34.1	1.5
0	-118.3	-34.2	1.5	1	35.3	4.2	0.9
0	-185.9	-280.0	6.7	1	49.5	25.1	2.6
0	-34.6	-19.4	3.4	1	18.1	13.5	4.0
0	-27.9	6.3	1.3	1	31.4	15.7	1.9
0	-48.2	6.8	1.6	1	21.5	-14.4	1.0
0	-49.2	-17.2	0.3	1	8.5	5.8	1.5

0	-19.2	-36.7	0.8	1	40.6	5.8	1.8
0	-18.1	-6.5	0.9	1	34.6	26.4	1.8
0	-98.0	-20.8	1.7	1	19.9	26.7	2.3
0	-129.0	-14.2	1.3	1	17.4	12.6	1.3
0	-4.0	-15.8	2.1	1	54.7	14.6	1.7
0	-8.7	-36.3	2.8	1	53.5	20.6	1.1
0	-59.2	-12.8	2.1	1	35.9	26.4	2.0
0	-13.1	-17.6	0.9	1	39.4	30.5	1.9
0	-38.0	1.6	1.2	1	53.1	7.1	1.9
0	-57.9	0.7	0.8	1	39.8	13.8	1.2
0	-8.8	-9.1	0.9	1	59.5	7.0	2.0
0	-64.7	-4.0	0.1	1	16.3	20.4	1.0
0	-11.4	4.8	0.9	1	21.7	-7.8	1.6

(a) Utilize R to fit a logistic regression model and perform a significance test.
(b) What is the estimated odds ratio, and how can it be interpreted?
(c) Does the fitted model accurately classify the data points?

3.7 LET US SUM UP

Logistic regression is a powerful tool for modeling binary outcomes, widely used in marketing, finance, healthcare, and social sciences. The logistic function ensures that predicted probabilities remain between 0 and 1, making it ideal for classification tasks. Coefficients are interpreted in terms of their impact on log-odds, while odds ratios provide an intuitive measure of variable influence. Significance testing helps assess the model's validity and the relevance of its predictors. By applying logistic regression to classification problems, organizations can make data-driven decisions with confidence.

3.8 Check Your Progress: Possible Answers

Check Your Progress – 1

R- Code

```
df <- read.table("Financal.Ratios.txt", header = TRUE)
# Fit the logistic regression model
model <- glm(Y ~ X1 + X2 + X3, data = df, family = binomial)

# Summarize the model
summary(model)
```

```r
# Fit the null model (intercept only)
model_null <- glm(Y ~ 1, data = df, family = binomial)

# Perform the chi-square test using the likelihood ratio test
anova(model_null, model, test = "Chisq")

# Make predictions (probabilities of redemption)
predicted_probabilities <- predict(model, type = "response")

# Add predictions to the dataset
df$Predicted_Probability <- predicted_probabilities

# Make binary predictions based on a 0.5 cutoff
df$Predicted_Y <- ifelse(predicted_probabilities > 0.5, 1, 0)

# Show the dataset with binary predictions
head(df)

# False positives: 0 actual, predicted as 1
false_positives <- sum(df$Y == 0 & df$Predicted_Y == 1)

# False negatives: 1 actual, predicted as 0
false_negatives <- sum(df$Y == 1 & df$Predicted_Y == 0)

# Create confusion matrix
confusion_matrix <- table(Predicted = df$Predicted_Y, Actual = df$Y)
```

3.9 Further Reading

1. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023
2. Statistics for Business & Economics 13th Edition, Anderson, Sweeney, Williams, Cengage Learning, January 2016
3. Applied Logistic Regression 3rd Edition, David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant, Wiley Publication, March 2013

3.10 Assignment

1. What is logistic regression, and how does it differ from linear regression?
2. Explain the logit transformation and its significance in logistic regression.
3. Define odds, log-odds, and odds ratio in the context of logistic regression.
4. What is the role of likelihood function and deviance in logistic regression.
5. How are predicted probabilities converted into binary outcomes in logistic regression? What is the impact of changing the classification threshold in logistic regression?

Unit 4: Model Building Guidelines

Unit Structure

4.0 Learning Objectives

By the end of this unit, learners will be able to:

- Understand the key steps in building a regression model.
- Identify and interpret predictor variables in the Boston dataset.
- Perform data exploration and preprocessing.
- Check and validate regression assumptions.
- Refine models using variable selection, transformations, and interaction terms.
- Evaluate and interpret final model results using numerical and graphical techniques.
- Predict and estimate outcomes using a fitted regression model.
- Avoid common pitfalls in regression analysis.

4.1 Introduction

When working with large datasets containing numerous potential predictors, it's easy to feel overwhelmed by the many modeling options available. Our goal is to identify a practical model that explains the relationship between a response variable (Y) and a set of predictor variables $(X_1, X_2, \ldots, X_k)$. While there may not be one "perfect" model, with enough effort, we can uncover multiple effective models that can offer meaningful insights. The key is to identify one of these valuable models that best fits the data. While different models may vary in terms of the specific predictors they use or the transformations applied, they typically offer similar interpretations and predictive accuracy. Constructing a reliable regression model requires selecting the right combination of predictor variables, ensuring that the model remains both interpretable and robust. To help guide this process, the following set of guidelines outlines an effective approach to model building, using the Boston dataset from the MASS package in R as an example.

4.2 Define the Problem and Identify Key Questions

- Clearly define the objective of the analysis.
- Identify the response variable (Y) and the potential predictor variables $(X_1, X_2, \ldots X_k)$.
- Determine whether the model aims to explain relationships, predict outcomes, or both.
- In the case of the **Boston** dataset, our goal is to predict median house prices (`medv`) based on various predictor variables.

```r
# Load necessary libraries
library(MASS)
library(car)  # For VIF analysis
library(ggplot2)

# Load dataset
data("Boston")
str(Boston)
```

4.2.1 Description of Model Predictors

The **Boston** dataset consists of various predictors that influence median house prices. Below is a brief description of key predictors:

- **crim**: Per capita crime rate by town.
- **zn**: Proportion of residential land zoned for large lots.

- **indus**: Proportion of non-retail business acres per town.
- **chas**: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
- **nox**: Nitrogen oxide concentration (parts per 10 million).
- **rm**: Average number of rooms per dwelling.
- **age**: Proportion of owner-occupied units built before 1940.
- **dis**: Weighted distances to employment centers.
- **rad**: Index of accessibility to radial highways.
- **tax**: Property tax rate per $10,000.
- **ptratio**: Pupil-teacher ratio by town.
- **black**: 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town.
- **lstat**: Percentage of lower status of the population.
- **medv**: Median value of owner-occupied homes in $1000s (response variable).

4.3 Data Collection and Preparation

- Collect relevant data, ensuring sufficient sample size.
- Clean and preprocess the data by handling missing values, outliers, and inconsistencies.
- Convert categorical variables into indicator (dummy) variables where necessary.
- Begin with univariate descriptive statistics and graphs to understand variable distributions.
- Conduct bivariate analyses to examine relationships between predictors and the response variable.

4.3.1 Handling Missing Data

- Missing data can reduce the total usable sample size, potentially weakening the analysis.
- The best way to address missing data is to minimize its occurrence by ensuring high-quality data collection.
- If missing data is unavoidable, imputation techniques can be used to estimate plausible values.
- If imputation is not feasible, models should be designed to exclude predictors with significant missing data.

4.3.2 Sample Size Considerations

- Larger sample sizes enhance the power of multiple linear regression models.

- Complex models with many predictors, transformations, and interactions require larger sample sizes.
- Weak associations between predictors and the response variable necessitate a larger sample.
- Determining an appropriate sample size is context-dependent and requires careful assumption checks.
- When designing studies, sample size and power calculations should be performed to ensure adequate representation.

```r
# Check for missing values
sum(is.na(Boston))

# Summary statistics
summary(Boston)
```

4.4 Exploratory Data Analysis (EDA)

- Visualize relationships using scatterplots, boxplots, and histograms.
- Compute summary statistics to understand variable distributions.
- Identify potential transformations (e.g., log transformation for skewed variables).
- Organize predictors into thematic sets (e.g., demographic, economic, or behavioral) to analyze their grouped effects.

```r
# Histogram and boxplot for target variable
par(mfrow = c(1,2))
hist(Boston$medv, main="Median House Prices", col="blue", xlab="medv")
boxplot(Boston$medv, main="Median Value Boxplot", col="red")

# Scatterplot for some predictors
pairs(Boston[, c("medv", "lstat", "rm", "ptratio")])
```

4.5 Model Fitting

- Fit an initial regression model including all potential predictors.
- Examine coefficient significance and model fit statistics.

```r
# Fit full model
full_model <- lm(medv ~ ., data = Boston)
summary(full_model)
```

4.5.1 Check Regression Assumptions

- Use residual plots to check for homoscedasticity.

- Perform normality tests on residuals.
- Assess multicollinearity using Variance Inflation Factor (VIF).

```
# Checking assumptions
par(mfrow = c(2,2))
plot(full_model)

# Checking multicollinearity
vif(full_model)
```

4.5.2 Model Refinement

- Remove non-significant variables step by step.
- Consider adding interaction terms and transformations.
- Compare models using Adjusted R^2, AIC, and BIC.

```
# Stepwise model selection
stepwise_model <- step(full_model, direction="both")
summary(stepwise_model)
```

Example: Adding Interaction Terms and Transformations

If two predictors interact meaningfully, an interaction term (e.g., $X_1 \times X_2$) can be included. Similarly, if a predictor has a nonlinear effect, a transformation (e.g., log or polynomial) can improve the model fit.

```
# Adding interaction and transformation
Boston$rm_sq <- Boston$rm^2  # Squaring the number of rooms
interaction_model <- lm(medv ~ lstat * rm_sq, data = Boston)
summary(interaction_model)
```

4.6 Model Validation and Interpretation

- Validate the final model using a separate dataset or cross-validation.
- Interpret predictor effects on Y and provide actionable insights.
- Estimate expected values of Y and predict individual values based on given predictor values.
- Use visualization techniques to communicate model results effectively.

```
# Splitting into training and testing sets
set.seed(123)
train_index <- sample(1:nrow(Boston), 0.7 * nrow(Boston))
train_data <- Boston[train_index, ]
test_data <- Boston[-train_index, ]
```

```r
# Fit model on training data
train_model <- lm(medv ~ ., data = train_data)
predicted_values <- predict(train_model, newdata = test_data)

# Model performance
actual_values <- test_data$medv
mean((predicted_values - actual_values)^2)  # MSE
```

4.6.1 Predictor Effect Plots

A predictor effect plot graphically shows how the response variable changes with a predictor while holding others constant.

- Write out the estimated regression equation.
- Set other predictor variables to convenient values (e.g., mean for continuous predictors, reference category for categorical variables).
- Construct a line plot with the predictor on the horizontal axis and its effect on the vertical axis.
- Repeat for each quantitative predictor.
- If a predictor interacts with categorical variables, include separate lines representing each category.
- If a predictor interacts with another continuous variable, plot a series of lines for different values (e.g., quartiles).

4.6.2 Avoiding Pitfalls

- Be cautious of overfitting by keeping the model as simple as possible.
- Avoid removing variables purely based on p-values without considering context.
- Evaluate whether regression assumptions still hold in the final model.
- Ensure that predictions do not extrapolate far beyond the sample data range.
- Validate the model using new or unseen data.

Check Your Progress – 1

1. ***What is the primary objective of building a regression model?***
   A) To find the single best model that perfectly predicts the response variable
   B) To find a useful model that explains relationships and predicts outcomes
   C) To include as many predictors as possible for the highest accuracy
   D) To ensure that all predictor variables are significant

2. ***Why is it important to check for missing data before building a regression model?***

A) Missing data always leads to model overfitting
B) Missing data reduces the total usable sample size and may bias the results
C) Missing data has no effect if we have a large enough sample
D) Missing data always improves model accuracy

3. ***Which of the following is NOT a recommended approach for handling missing data?***

A) Imputing missing values with plausible estimates
B) Removing predictors with a large amount of missing values
C) Ignoring missing values and proceeding with the analysis
D) Minimizing missing data through careful data collection

4. What is the purpose of exploratory data analysis (EDA) in regression modeling?

A) To finalize the regression model before checking assumptions
B) To visualize relationships, detect patterns, and identify necessary transformations
C) To eliminate all outliers from the dataset
D) To randomly select predictor variables for the model

5. Which of the following techniques can help assess multicollinearity in a regression model?

A) Residual plots
B) Variance Inflation Factor (VIF)
C) Scatterplots
D) Adjusted R^2

6. How can interaction terms improve a regression model?

A) By allowing relationships between predictors and response variables to change based on other predictors
B) By increasing the number of predictors in the model without affecting interpretation
C) By ensuring all predictor variables remain significant
D) By reducing the need for transformations

7. What is an appropriate way to validate a regression model?

A) Using only the training dataset for evaluation
B) Checking model performance on a separate validation dataset
C) Increasing the number of predictors until model performance improves
D) Removing all insignificant predictors without checking model assumptions

8. In a predictor effect plot, what does holding all other predictors constant allow us to observe?

A) The impact of one predictor variable on the response variable
B) The effect of all predictor variables simultaneously
C) The residual variance of the model
D) The significance of each coefficient

9. Which of the following is a sign of overfitting in a regression model?

A) High performance on training data but poor performance on validation data
B) Low adjusted R^2 value for the model
C) A model with very few predictor variables
D) A model that includes only statistically significant variables

10. What is the main reason for using transformations in regression models?

A) To increase the number of predictor variables
B) To meet regression assumptions such as normality and linearity
C) To artificially increase the R^2 value
D) To remove all categorical variables from the model

4.7 LET US SUM UP

By carefully following these steps, analysts can create regression models that are not only interpretable and robust but also capable of delivering valuable insights. A crucial point to remember is that there is no single "best" model; instead, several good models can often offer similar interpretations and predictions. The iterative nature of model refinement helps ensure that the final model is both statistically rigorous and practically effective.

In this unit, we demonstrated how to implement the process of building a regression model using the Boston dataset in R. The coding steps covered data preparation, model fitting, assumption checking, and refinement. However, we have not yet completed the entire process, and interpretation of the results was not part of this discussion. We also introduced model validation techniques to ensure the reliability of predictions. For a more detailed analysis of these techniques, you can refer to the earlier block, where each step was thoroughly examined. Moving forward, we will focus on interpreting the model results to derive meaningful insights and guide decision-making.

By adhering to these guidelines and best practices, you can develop reliable and interpretable regression models that provide valuable insights for decision-making and prediction.

4.8 Check Your Progress: Possible Answers

Check Your Progress – 1

Answer 1: B) To find a useful model that explains relationships and predicts outcomes

Answer 2: B) Missing data reduces the total usable sample size and may bias the results

Answer 3: C) Ignoring missing values and proceeding with the analysis

Answer 4: B) To visualize relationships, detect patterns, and identify necessary transformations

Answer 5: B) Variance Inflation Factor (VIF)

Answer 6: A) By allowing relationships between predictors and response variables to change based on other predictors

Answer 7: B) Checking model performance on a separate validation dataset

Answer 8: A) The impact of one predictor variable on the response variable

Answer 9: A) High performance on training data but poor performance on validation data

Answer 10: B) To meet regression assumptions such as normality and linearity

4.9 Further Reading

1. Applied Regression Modeling 3rd Edition, IAIN PARDOE**,** John Wiley & Sons, Inc, December 2020
2. Regression Analysis by Example Using R 6th Edition, Ali S. Hadi and Samprit Chatterjee, Wiley Publication, October 2023

4.10 Assignment

1. What are the key steps in building a regression model?
2. Why is exploratory data analysis (EDA) important before fitting a model?
3. How can missing data impact the regression model, and what are possible solutions?
4. What are the key differences between training and validation datasets?
5. How can predictor effect plots help in model interpretation?

BAOU
Education for All

યુનિવર્સિટી ગીત

સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ
સ્વાધ્યાયઃ પરમં તપઃ

શિક્ષણ, સંસ્કૃતિ, સદ્‌ભાવ, દિવ્યબોધનું ધામ
ડૉ. બાબાસાહેબ આંબેડકર ઓપન યુનિવર્સિટી નામ;
સૌને સૌની પાંખ મળે, ને સૌને સૌનું આભ,
દશે દિશામાં સ્મિત વહે હો દશે દિશે શુભ-લાભ.

અભણ રહી અજ્ઞાનના શાને, અંધકારને પીવો ?
કહે બુદ્ધ આંબેડકર કહે, તું થા તારો દીવો;
શારદીય અજવાળા પહોંચ્યાં ગુર્જર ગામે ગામ
ધ્રુવ તારકની જેમ ઝળહળે એકલવ્યની શાન.

સરસ્વતીના મયૂર તમારે ફળિયે આવી ગહેકે
અંધકારને હડસેલીને ઉજાસના ફૂલ મહેંકે;
બંધન નહીં કો સ્થાન સમયના જવું ન ઘરથી દૂર
ઘર આવી મા હરે શારદા દૈન્ય તિમિરના પૂર.

સંસ્કારોની સુગંધ મહેંકે, મન મંદિરને ધામે
સુખની ટપાલ પહોંચે સૌને પોતાને સરનામે;
સમાજ કેરે દરિયે હાંકી શિક્ષણ કેરું વહાણ,
આવો કરીયે આપણ સૌ
ભવ્ય રાષ્ટ્ર નિર્માણ...
દિવ્ય રાષ્ટ્ર નિર્માણ...
ભવ્ય રાષ્ટ્ર નિર્માણ

DR. BABASAHEB AMBEDKAR OPEN UNIVERSITY
(Established by Government of Gujarat)
'Jyotirmay' Parisar,
Sarkhej-Gandhinagar Highway, Chharodi, Ahmedabad-382 481
Website : www.baou.edu.in