People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Constantine 1 University-Frères Mentouri
Faculty of Exact Sciences
Department of Mathematics

Linear Model: Application to Regression and ANOVA

Course Notes

Master 1 – Applied Statistics

Prepared by:
Dalel ZERDAZI

Academic Year: 2025–2026

Contents

Introduction

The linear model is a fundamental framework in applied statistics, providing a unified approach to modeling and analyzing both quantitative and qualitative data. This course focuses on the application of linear models to regression analysis and Analysis of Variance (ANOVA), allowing students to develop the skills required to study and interpret statistical relationships in carefully designed studies.

Course Overview. The content of this course is structured to cover the following major themes: the fundamentals of the linear model, the development of simple and multiple regression techniques, and their connections with classical ANOVA frameworks. It progressively introduces both the theoretical and computational aspects of model estimation using Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE), while emphasizing inference, diagnostics, and model validation. Advanced topics such as multicollinearity, model selection, and the treatment of singular cases are also discussed. Finally, the course extends to hierarchical and mixed factor models, bridging classical regression concepts with modern applications and practical implementation in R.

Learning Objectives

By completing this course, students will be able to:

- Master the theoretical foundations of linear modeling, including model assumptions and properties.
- Apply OLS and MLE methods for parameter estimation and understand their statistical characteristics.
- Detect and remedy problems such as multicollinearity, heteroscedasticity, and non-normality of errors.
- Perform hypothesis testing using linear models, including Fisher's F-test and linear contrasts.
- Analyze experimental designs with one or multiple factors, including hierarchical and mixed models.
- Extend the linear modeling framework to more general situations, preparing for advanced applications in research and professional practice.

Course Importance

The knowledge acquired in this course is essential for any student aiming to conduct rigorous statistical analyses. It provides a solid foundation for understanding complex data structures, improving model interpretation, and ensuring reliable inference. Moreover, the methods and concepts taught here are directly applicable to a wide range of fields, including economics, social sciences, biology, engineering, and many others.

Expected Outcomes

Upon successful completion of the course, students will be able to:

- Formulate and estimate linear models for real-world data.
- Select appropriate estimation and testing methods depending on the design and data characteristics.
- Diagnose and address model violations using both theoretical and practical approaches.
- Critically interpret statistical results and communicate findings effectively in academic and applied contexts.

Chapter 1

Fundamentals of the Linear Model

This chapter introduces the linear model, presenting its historical context, basic definitions, and geometric interpretation. It highlights the assumptions underlying linear regression, discusses key properties and limitations, provides a foundation for understanding extensions in more complex modeling scenarios.

1.1 Historical Background and Motivation

The linear model is one of the most fundamental and widely used tools in statistics and data analysis. Its origins date back to the early 19th century, when **Legendre** (1805) and **Gauss** (1809) introduced the method of *least squares* to solve astronomical problems. Later, statisticians such as **Karl Pearson** and **R.A. Fisher** extended the theory, making it a cornerstone for experimental design, econometrics, biostatistics, and the social sciences.

The importance of the linear model lies in its simplicity and interpretability: it describes how a response variable Y changes as a function of one or more explanatory variables X. Despite its simplicity, the linear model provides the foundation for more advanced techniques such as *generalized linear models*, *mixed models*, and modern machine learning algorithms.

1.2 Introductory Examples

Before diving into the formal definitions, let us consider a few motivating examples:

- **Economics:** Modeling the relationship between individual income (Y) and years of education (X).
- **Medicine:** Predicting a patient's blood pressure (Y) based on age and body mass index (BMI) (X_1, X_2).
- **Engineering:** Estimating the lifetime of a machine (Y) from the temperature and pressure at which it operates (X_1, X_2).
- **Social sciences:** Studying the effect of study hours (X) on student performance (Y).

These simple cases illustrate how the linear model provides a first and intuitive approach to quantify relationships between variables.

1.3 Definitions and General Form

The general form of the linear regression model can be written as:

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where:

- Y is the $n \times 1$ vector of observed responses (dependent variable);
- X is the $n \times p$ design matrix of explanatory variables (predictors);
- β is the $p \times 1$ vector of unknown coefficients (parameters);
- ε is the $n \times 1$ vector of random errors satisfying

$$\mathbb{E}[\varepsilon] = 0, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Definition 1.1. *The response variable, denoted by Y, is the variable we aim to explain or predict from the explanatory variables.*

Definition 1.2. *The explanatory variables, also called predictors or independent variables, are represented by the columns of the matrix X. They provide the information used to explain the variations in Y.*

Definition 1.3. *The coefficients β are the parameters of the linear model. Each element of β represents the partial effect of the corresponding explanatory variable on the expected value of Y.*

Definition 1.4. *The random error term ε represents the unobserved factors affecting the response variable and accounts for the discrepancy between the observed values and their conditional expectation. For exact finite-sample inference, it is commonly assumed that*

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

1.4 Geometric Intuition

The linear model can also be understood in terms of geometry:

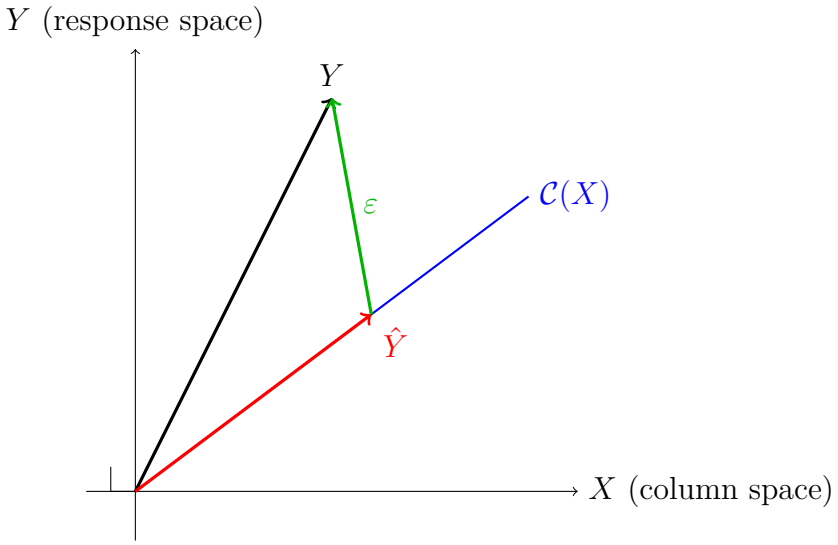

Figure 1.1: Geometric interpretation of the linear model.

Geometric interpretation.

In the linear model, the response vector Y can be decomposed into two components: the fitted value $\hat{Y} = X\hat{\beta}$, which lies in the column space of the design matrix $\mathcal{C}(X)$, and the residual vector ε, which is orthogonal to that space.

This illustrates how the model expresses Y as a combination of explanatory variables and an error term.

This interpretation provides a powerful way to visualize regression as fitting a line (or hyperplane) in high-dimensional space.

1.5 Assumptions

For valid estimation and inference, the classical linear model relies on a set of standard assumptions. These assumptions concern the structure of the model, the properties of the error term, and the design matrix.

1. **Linearity in Parameters**

   The model is linear in the unknown parameters:

   $$Y = X\beta + \varepsilon.$$

   This means that the coefficients β enter the model linearly, even if the explanatory variables themselves involve nonlinear transformations (e.g., X^2, $\log(X)$).

2. **Zero Mean of the Errors**

   The error term has zero mean:

   $$\mathbb{E}[\varepsilon] = 0.$$

   This assumption ensures that the model is correctly centered and that the expected value of the response variable is given by

   $$\mathbb{E}[Y \mid X] = X\beta.$$

3. **Homoscedasticity and No Autocorrelation**

   The variance–covariance matrix of the errors is

   $$\mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

   This implies that:

   - All errors have the same variance σ^2 (homoscedasticity),
   - The errors are uncorrelated with each other.

4. **Full Rank of the Design Matrix**

   The design matrix X has full column rank:

   $$\mathrm{rank}(X) = p.$$

   This condition guarantees that the matrix $X^T X$ is invertible, which ensures the existence and uniqueness of the Ordinary Least Squares estimator.

Additional Assumption for Statistical Inference. For hypothesis testing and the construction of confidence intervals, an additional assumption is often introduced:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Under this normality assumption, the estimators and test statistics follow exact probability distributions, allowing the use of t-tests and F-tests.

1.6 Properties

The linear model enjoys several important properties:

- The **Ordinary Least Squares (OLS)** estimator is the Best Linear Unbiased Estimator (BLUE) according to the Gauss–Markov theorem.
- The coefficients β_j have a clear interpretation: the expected change in Y when X_j increases by one unit, holding other variables constant.
- The model is **additive**: the total effect of predictors on Y is the sum of their individual effects.
- The model is **linear in parameters**: even if X contains nonlinear transformations (e.g., X^2), the coefficients still enter linearly.
- The model can be easily extended to include interaction terms and polynomial effects.
- In time series contexts, additional assumptions such as stationarity may be required.

1.7 Limitations

Despite its usefulness, the linear model also presents some limitations:

- It may not capture nonlinear relationships adequately.
- It is sensitive to outliers, which can distort estimates.
- The validity of inference depends on strong assumptions (independence, homoscedasticity, normality).
- Multicollinearity among predictors can make estimation unstable.

These limitations motivate the development of more flexible models and robust estimation techniques.

1.8 Extensions and Outlook

While the preceding discussion illustrated the linear model framework with one or more predictors, the linear model framework can be extended to:

- Multiple predictors (Multiple Linear Regression),
- Nonlinear transformations (Polynomial Regression),
- Categorical predictors (ANOVA models),
- Advanced estimation methods (Maximum Likelihood Estimation).

Thus, the linear model serves as a gateway to more sophisticated statistical and machine learning approaches.

In the following chapter, dedicated to the Simple Linear Regression Model and One-Way ANOVA, we will focus on estimating the coefficients of the model and evaluating their statistical properties.

Chapter 2

Simple Linear Regression Model and One-Way ANOVA

This chapter builds on the theoretical foundations of the linear model presented in Chapter 1 to develop estimation techniques, study their properties, and apply inference procedures.

2.1 Simple Linear Regression Model

The simplest case is when there is only one explanatory variable X. The model becomes:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{2.1}$$

where:

- β_0 is the intercept, the expected value of Y when $X = 0$;
- β_1 is the slope, measuring the expected change in Y for a one-unit increase in X;
- ε_i are the independent **random errors**, where $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Normality may be assumed for inferential purposes.

Motivating Example

Consider the relationship between daily temperature (X, in °C) and ice cream sales (Y, in units sold). The data for five summer days are given below:

Temperature (°C)	Ice Cream Sales (units)
20	120
25	150
30	200
35	250
40	300

Table 2.1: Daily temperature and ice cream sales.

The model can be expressed as in equation (2.1), where β_0 represents the baseline sales when the temperature is zero, and β_1 quantifies the increase in sales per degree Celsius.

Although simplified, this example illustrates the rationale for linear modeling: the slope quantifies the effect of temperature on ice cream demand, while the intercept represents the theoretical baseline sales when temperature is zero.

2.2 Ordinary Least Squares Method

Ordinary Least Squares (OLS) technique is a fundamental approach in simple linear regression used to determine the best-fitting line through the data.

2.2.1 Estimation of the Regression Coefficients β_0 and β_1

In this step, the regression coefficients are calculated by selecting the parameter values that minimize the total squared difference between the observed data and the model's predictions.
The OLS method seeks estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following function:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2. \tag{2.2}$$

This criterion corresponds to finding the line that best fits the data in the least squares sense, by minimizing the sum of squared vertical deviations between the observed values and the fitted regression line.
The minimization problem consists in finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $S(\beta_0, \beta_1)$. The following subsection develops this analytical approach and leads to the closed-form expressions of the OLS estimators.

Analytical Approach

The OLS estimates, obtained by solving the minimization problem in (2.2), are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{2.3}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{2.4}$$

where $\bar{X}$ and $\bar{Y}$ denote the sample means of X and Y, respectively.

Derivation of the Estimators

To minimize $S(\beta_0, \beta_1)$, we take partial derivatives with respect to β_0 and β_1, leading to the following conditions, called *normal equations*:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0,$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0.$$

Setting the derivatives equal to zero yields the first-order conditions, known as the normal equations:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i, \tag{2.5}$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2. \tag{2.6}$$

Dividing equation (2.5) by n gives:

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} \quad \Rightarrow \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}. \tag{2.7}$$

Substituting this expression into equation (2.6), we obtain:

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Thus, the slope estimator is given by equation (2.3), and the intercept is obtained from equation (2.4).

Interpretation

- $\hat{\beta}_1$ measures the estimated change in Y associated with a one-unit increase in X. Its sign indicates the direction of the relationship (positive or negative).
- $\hat{\beta}_0$ is the estimated value of Y when $X = 0$. In practice, this intercept may sometimes have limited interpretability if $X = 0$ is outside the observed data range. In such cases, the intercept primarily serves to position the regression line rather than provide substantive interpretation.

Matrix Approach

The estimation of the parameter vector β using Ordinary Least Squares (OLS) method can be compactly expressed in matrix notation. The principle remains the same: minimize the sum of squared residuals (RSS):

$$RSS(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T (Y - X\beta), \tag{2.8}$$

where:

- Y is the $n \times 1$ vector of responses,
- X is the $n \times p$ design matrix containing the predictors (including a column of ones for the intercept),
- β is the $p \times 1$ vector of coefficients.

Matrix dimensions for simple linear regression:

$$X \in \mathbb{R}^{n \times 2}, \quad Y \in \mathbb{R}^{n \times 1}, \quad \beta \in \mathbb{R}^{2 \times 1}.$$

Here, n is the number of observations, the first column of X is a column of ones for the intercept, and the second column contains the observed values of the explanatory variable X_i.

Expanding the quadratic form in (2.8) gives:

$$RSS(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta. \tag{2.9}$$

Solution of the Normal Equations

To find the minimizer, we compute the gradient of (2.8) with respect to β and set it equal to zero:

$$\frac{\partial}{\partial \beta} RSS(\beta) = -2X^T(Y - X\beta) = 0. \tag{2.10}$$

This leads to the *normal equations*:

$$X^T Y = X^T X \hat{\beta}_{OLS}. \tag{2.11}$$

Provided that $X^T X$ is invertible (i.e., the columns of X are linearly independent), the solution is uniquely given by:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y. \tag{2.12}$$

Interpretation

- The matrix form unifies the estimation procedure for both simple and multiple regression.
- The condition that $X^T X$ be invertible ensures identifiability of the coefficients. In cases of multicollinearity (when predictors are linearly dependent), this matrix is singular, and the OLS estimator cannot be computed directly.
- This compact representation highlights that OLS estimation corresponds to projecting the response vector Y onto the column space of X.

Remark 2.2.1. *While the analytical approach is essential for understanding simple linear regression, the matrix approach generalizes seamlessly to multiple regression models and forms the backbone of modern statistical inference.*

2.2.2 Properties of the OLS Estimators

Ordinary least squares (OLS) estimator plays a central role in linear regression. Its two key properties : *unbiasedness* and *efficiency*, guarantee that the estimator is both accurate (on average) and statistically optimal within a broad class of estimators.

Unbiasedness

OLS estimator is given by equation (2.12):

$$\hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T Y,$$

under the linear model

$$Y = X\beta + \varepsilon, \tag{2.13}$$

with assumptions

$$\mathbb{E}(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I_n.$$

Here, for simple linear regression, the matrix dimensions are:

$$X \in \mathbb{R}^{n \times p}, \quad Y \in \mathbb{R}^{n \times 1}, \quad \beta \in \mathbb{R}^{p \times 1}, \quad p = 2.$$

Since X is non-random and $\mathbb{E}(\varepsilon) = 0$, it follows that

$$\mathbb{E}(Y) = X\beta.$$

Substituting into (2.12) gives

$$\mathbb{E}(\hat{\beta}_{\text{OLS}}) = (X^T X)^{-1} X^T \mathbb{E}(Y) = (X^T X)^{-1} X^T X\beta = \beta.$$

Therefore, $\hat{\beta}_{\text{OLS}}$ is an **unbiased estimator** of β.

Efficiency (BLUE)

An estimator is *efficient* if it is linear, unbiased, and has the minimum variance among all linear unbiased estimators. Under the Gauss–Markov assumptions:

- Correct linear model specification,
- Zero-mean errors: $\mathbb{E}(\varepsilon) = 0$,
- Homoscedasticity: $\text{Var}(\varepsilon) = \sigma^2 I_n$,
- Uncorrelated errors: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.

The Gauss–Markov theorem asserts that $\hat{\beta}_{\text{OLS}}$ is the **B**est **L**inear **U**nbiased **E**stimator (BLUE) of β. That is, it achieves the smallest variance among all linear unbiased estimators.

Remark 2.2.2. *A detailed treatment of the Gauss–Markov theorem, including its proof, matrix formulation, and implications for multiple regression, is provided in Chapter 3.*

2.2.3 Variance of the OLS Estimator and Its Estimation

The variance of $\hat{\beta}_{\text{OLS}}$ quantifies the dispersion of the estimator around its expected value, it is defined as:

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = \mathbb{E}\Big[(\hat{\beta}_{\text{OLS}} - \mathbb{E}[\hat{\beta}_{\text{OLS}}])(\hat{\beta}_{\text{OLS}} - \mathbb{E}[\hat{\beta}_{\text{OLS}}])^T\Big] \tag{2.14}$$

By the unbiasedness of OLS : $\mathbb{E}[\hat{\beta}_{\text{OLS}}] = \beta$, expression (2.14) simplifies to:

$$\text{Var}(\hat{\beta}_{OLS}) = \mathbb{E}\Big[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)^T\Big]. \tag{2.15}$$

Substituting the OLS estimator from equation (2.12),

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y,$$

into the linear model given in equation (2.13),

$$Y = X\beta + \varepsilon,$$

we obtain :

$$\text{Var}(\hat{\beta}_{OLS}) = \mathbb{E}\Big[(X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}\Big].$$

Under the classical linear regression assumption $\mathbb{E}(\varepsilon\varepsilon^T) = \sigma^2 I_n$, this reduces to:

$$\text{Var}(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1}. \tag{2.16}$$

Estimated variance from residuals.

Since σ^2 is unknown, it is replaced by

$$s^2 = \frac{1}{n-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta}), \tag{2.17}$$

yielding the estimated variance:

$$\widehat{\mathrm{Var}}(\hat{\beta}_{OLS}) = s^2(X^TX)^{-1}. \tag{2.18}$$

Conclusion

- σ^2 is the true but unknown error variance in the population.
- s^2 is an unbiased estimator of σ^2 computed from the residuals.
- The estimated variance of $\hat{\beta}_{OLS}$ in (2.18) is the cornerstone of statistical inference.
- This result directly supports t-tests, F-tests, and confidence interval construction in regression analysis.
- It also emphasizes the importance of experimental design: more variability in X leads to more precise coefficient estimates.

Example: Study Hours and Exam Scores

Note: This example is presented in full detail to demonstrate, step by step, the complete procedure for estimating the coefficients of a simple linear regression model.

Problem Statement

A small study examines the relationship between students' study time and their exam scores. For each of five students, the number of study hours X and the corresponding exam score Y are recorded. The observed data are given in Table 2.2.

Study Hours (X)	Exam Score (Y)
1	52
2	55
3	58
4	60
5	62

Table 2.2: Study hours and exam scores of five students.

We aim to fit the simple linear regression model

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

where X is the number of study hours, Y is the exam score, β_0 the intercept, and β_1 the slope.

Questions

1. Fit the model using Ordinary Least Squares (OLS) method:
   - Compute the sample means $\bar{X}$ and $\bar{Y}$.
   - Derive the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_0$.
   - Write the fitted regression equation $\hat{Y}$.
2. Interpret the estimated slope and intercept in the context of the study.
3. Verify the results using the matrix formula $\hat{\boldsymbol{\beta}}_{OLS} = (X^\top X)^{-1} X^\top Y$.

Solution

1. Analytical estimation of the parameters
The OLS estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Step 1: Means :

$$\bar{X} = 3, \qquad \bar{Y} = 57.4.$$

Step 2: Deviations : The deviations from the means and their products are summarized in Table 2.3.

X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	52	-2	-5.4	10.8
2	55	-1	-2.4	2.4
3	58	0	0.6	0.0
4	60	1	2.6	2.6
5	62	2	4.6	9.2

Table 2.3: Deviations from means and cross-products.

Step 3: Slope :

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 25, \qquad \sum_{i=1}^{n}(X_i - \bar{X})^2 = 10,$$

$$\hat{\beta}_1 = \frac{25}{10} = 2.5.$$

Step 4: Intercept :

$$\hat{\beta}_0 = 57.4 - (2.5 \times 3) = 49.9.$$

Hence, the fitted regression line is:

$$\hat{Y} = 49.9 + 2.5X.$$

2. Interpretation of the coefficients

- The intercept $\hat{\beta}_0 = 49.9$ represents the predicted exam score for a student who studies zero hours.
- The slope $\hat{\beta}_1 = 2.5$ indicates that each additional hour of study increases the expected exam score by about 2.5 points.

3. Verification using the matrix formula
We compute

$$\hat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T Y,$$

with

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \quad Y = \begin{bmatrix} 52 \\ 55 \\ 58 \\ 60 \\ 62 \end{bmatrix}.$$

$$X^T X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 287 \\ 886 \end{bmatrix}, \quad (X^T X)^{-1} = \tfrac{1}{50} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix}.$$

Thus,

$$\hat{\boldsymbol{\beta}}_{OLS} = \begin{bmatrix} 49.9 \\ 2.5 \end{bmatrix},$$

which exactly confirms the analytical calculation.

Practical Work 1: Study Hours and Exam Scores in R

Listing 2.1: Simple linear regression of exam score on study hours in R, including regression line and legend.

```
1
2  # Create the data
3  study_hours <- c(1, 2, 3, 4, 5)
4  exam_score <- c(52, 55, 58, 60, 62)
5
6  # Create a data frame
7  data <- data.frame(study_hours, exam_score)
8
9  # Fit the linear regression model
10 model <- lm(exam_score ~ study_hours, data = data)
11
12 # Model summary
13 summary(model)
14
15 # Plot with regression line
16 plot(study_hours, exam_score,
17      main = "Study Hours vs Exam Score",
18      xlab = "Study hours (X)",
19      ylab = "Exam score (Y)",
```

```r
     pch = 19, col = "black")
abline(model, col = "blue", lwd = 2)

# Add legend
legend("topleft",
       legend = c("Data points", "Regression line"),
       col = c("black", "blue"),
       pch = c(19, NA),
       lty = c(NA, 1),
       lwd = c(NA, 2),
       bty = "n")  # removes the box around legend
```

R Output for the OLS Model

Listing 2.2: R console output of the fitted OLS model

```
Call:
lm(formula = exam_score ~ study_hours, data = data)

Residuals:
    1     2     3     4     5
 -0.4   0.1   0.6   0.1  -0.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.9000     0.5066   98.50  2.31e-06 ***
study_hours    2.5000     0.1528   16.37  0.000496 ***
---
Signif. codes: 0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

Residual standard error: 0.483 on 3 degrees of freedom
Multiple R-squared: 0.9889,    Adjusted R-squared: 0.9852
F-statistic: 267.9 on 1 and 3 DF,  p-value: 0.0004964
```

Explanation. The R console output above summarizes the fitted Ordinary Least Squares (OLS) model:

- **Residuals:** The differences between observed and predicted exam scores are all very small (close to zero), which indicates an excellent fit.
- **Coefficients:**
  - Intercept: $\hat{\beta}_0 = 49.9$ with a very small p-value (2.31×10^{-6}), confirming its strong statistical significance.
  - Slope: $\hat{\beta}_1 = 2.5$ with a p-value of 0.000496, meaning each additional study hour increases the predicted exam score by approximately 2.5 points.

- **Model Fit:** The residual standard error is 0.483, and $R^2 = 0.9889$ (adjusted $R^2 = 0.9852$), which shows that about 99% of the variance in exam scores is explained by study hours.
- **Overall Significance:** The F-statistic ($F = 267.9$, $p = 0.0004964$) demonstrates that the regression model as a whole is highly significant.

The graphical representation of the fitted model is shown in Figure 2.1. It illustrates the observed data points together with the estimated regression line.

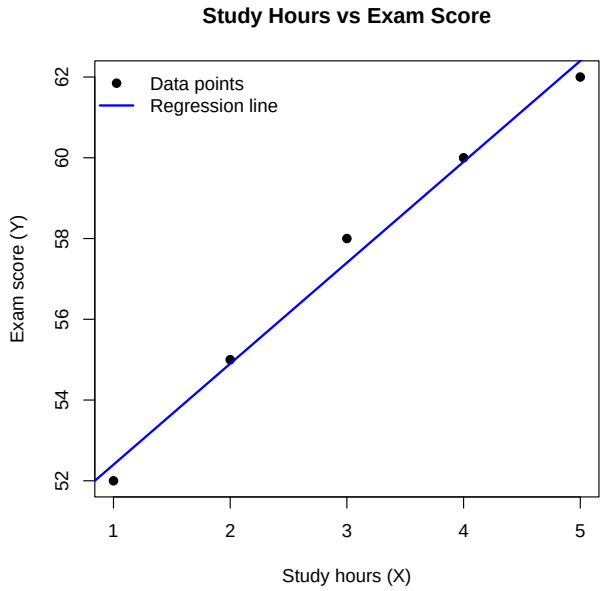

Figure 2.1: Scatter plot of study hours (X) versus exam scores (Y) with fitted regression line.

Results and Discussion

The R implementation confirms the theoretical results obtained both analytically and via the matrix formulation, showing the consistency between theory and practice. The fitted regression model is given by:

$$\hat{Y} = 49.9 + 2.5X.$$

Both coefficients are statistically significant, and the model exhibits excellent explanatory power ($R^2 = 98.89\%$).

This example highlights the agreement between analytical, matrix-based, and computational approaches to simple linear regression, and provides a clear interpretation of how study time influences exam performance.

2.3 Maximum Likelihood Estimation

2.3.1 Estimation of the Regression Coefficients β_0 and β_1

As introduced in equation (2.1), we consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus, conditionally on X_i, the response variable is normally distributed:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2).$$

The probability density function of an observation Y_i is given by:

$$f(Y_i \mid X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right). \tag{2.19}$$

The likelihood function for all n observations is therefore:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i \mid X_i).$$

Substituting (2.19), we obtain:

$$L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right).$$

Taking the logarithm, the log-likelihood becomes:

$$\log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

Maximizing the likelihood with respect to β_0 and β_1 is equivalent to minimizing the sum of squared errors:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2.$$

Matrix formulation:

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \in \mathbb{R}^{n\times 2}, \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^{n\times 1}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \in \mathbb{R}^{2\times 1}.$$

Derivative with respect to β_0

Differentiating with respect to β_0 gives:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0.$$

Rearranging yields:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}. \tag{2.20}$$

Derivative with respect to β_1

Differentiating with respect to β_1 gives:

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0.$$

Expanding and substituting the expression for β_0, we obtain the well-known formula for the slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{2.21}$$

Interpretation of coefficients:

- $\hat{\beta}_1$ indicates the expected change in Y for a one-unit increase in X.
- $\hat{\beta}_0$ represents the expected value of Y when $X = 0$ (may be outside observed range).

2.3.2 Estimation of the Variance

After estimating the regression coefficients β_0 and β_1, we now turn to the estimation of the variance σ^2 of the error term ε_i in (2.1):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Maximizing the Log-Likelihood with Respect to σ^2

Differentiating the log-likelihood with respect to σ^2 gives:

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Setting this derivative equal to zero and solving yields the MLE of the variance:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \tag{2.22}$$

Remark 2.3.1. *The Maximum Likelihood Estimator (MLE) of σ^2 divides by n, while the unbiased OLS-based estimator divides by $(n - p)$ (here $p = 2$ for simple regression), reflecting the loss of degrees of freedom due to estimating the regression coefficients. Both estimators converge as n increases.*

Example: Height–Weight Regression

Problem Statement

We investigate the relationship between height (in cm) and weight (in kg) in a sample of seven individuals. The observed data are summarized in Table 2.4.

Height (X, cm)	150	155	160	165	170	175	180
Weight (Y, kg)	50	52	55	58	60	65	68

Table 2.4: Observed data

We assume a linear regression model and aim to estimate β_0, β_1, and σ^2 using MLE.

Solution

1. Estimation of β_0 and β_1

$$\bar{X} = 165, \qquad \bar{Y} = 58.29,$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 425, \qquad \sum_{i=1}^{n} (X_i - \bar{X})^2 = 700,$$

$$\hat{\beta}_1 = \frac{425}{700} = 0.607, \quad \hat{\beta}_0 = 58.29 - (0.607 \times 165) = -41.87.$$

2. Estimation of σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{n} = \frac{3.42}{7} = 0.489.$$

The estimated regression equation is:

$$\hat{Y} = -41.87 + 0.607X.$$

Practical Work 2: Height–Weight Regression in R

Listing 2.3: Maximum Likelihood Estimation (MLE) in R with regression line and legend.

```
1
2  # Define the data
3  X <- c(150, 155, 160, 165, 170, 175, 180)
4  Y <- c(50, 52, 55, 58, 60, 65, 68)
5  n <- length(Y)  # Number of observations
6
7  # Log-likelihood function
8  log_likelihood <- function(params) {
9    beta0 <- params[1]
10   beta1 <- params[2]
11   sigma2 <- params[3]
12
13   if (sigma2 <= 0) return(Inf)  # Avoid negative variance
14
15   # Compute residuals
16   residuals <- Y - (beta0 + beta1 * X)
17
18   # Log-likelihood under normality assumption
19   logL <- - (n / 2) * log(2 * pi * sigma2) - (1 / (2 * sigma2)) *
         sum(residuals^2)
20
21   return(-logL)  # Minimize negative log-likelihood
22 }
23
24 # Initial parameter values
25 init_params <- c(beta0 = 0, beta1 = 0.1, sigma2 = 1)
26
27 # Maximize log-likelihood with constraints (sigma^2 > 0)
28 result <- optim(par = init_params,
29                 fn = log_likelihood,
30                 method = "L-BFGS-B",
31                 lower = c(-Inf, -Inf, 1e-6))
32
33 # Check convergence and extract results
34 if (result$convergence == 0) {
35   beta0_hat <- result$par[1]
36   beta1_hat <- result$par[2]
37   sigma2_hat <- result$par[3]
38
```

```r
  # Display results
  cat("Optimization successful!\n")
  cat("MLE of beta0:", beta0_hat, "\n")
  cat("MLE of beta1:", beta1_hat, "\n")
  cat("MLE of sigma^2:", sigma2_hat, "\n")

  # Plot regression line
  plot(X, Y,
       xlab = "Height (cm)",
       ylab = "Weight (kg)",
       pch = 16, col = "blue")
  abline(a = beta0_hat, b = beta1_hat, col = "red", lwd = 2)

  # Add a legend
  legend("topleft",
         legend = c("Data points", "MLE regression line"),
         col = c("blue", "red"),
         pch = c(16, NA),
         lty = c(NA, 1),
         lwd = c(NA, 2),
         bty = "n")  # no box around legend
} else {
  cat("Optimization did not converge. Check the data and initial
     values.\n")
}
```

The R code above estimates the regression parameters using Maximum Likelihood Estimation (MLE) by minimizing the negative log-likelihood. The fitted regression line is visualized below:

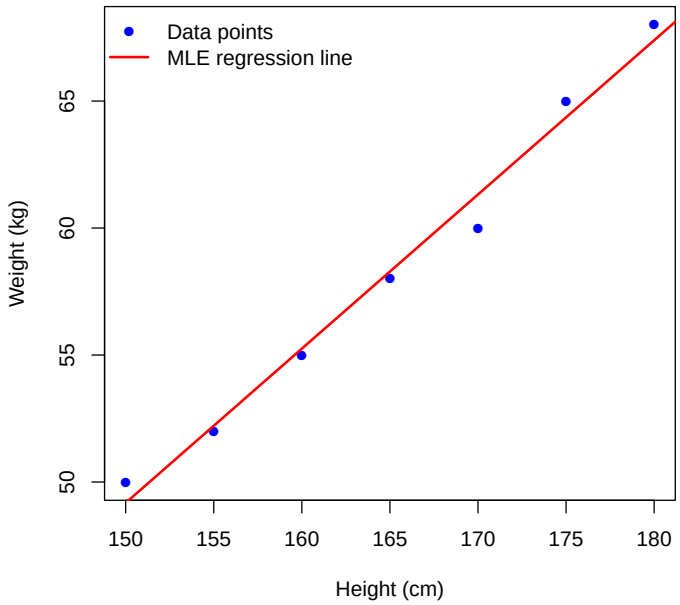

Figure 2.2: Regression line based on MLE fitted to Height–Weight data

Results and Discussion

MLE provides the optimal coefficients. The fitted regression line captures the increasing linear relationship between height (X) and weight (Y):

$$\hat{\beta}_0 = -41.87, \quad \hat{\beta}_1 = 0.607.$$

- The slope indicates that each additional cm of height corresponds to 0.61 kg increase in weight.
- The intercept is negative and has no direct physical meaning; it positions the regression line mathematically outside the observed range.

2.4 Comparison between OLS and MLE

The following table summarizes the main similarities and differences between Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) in linear regression.

Criterion	**Ordinary Least Squares (OLS)**	**Maximum Likelihood Estimation (MLE)**
Approach	Minimizes the sum of squared errors	Maximizes the likelihood function
Assumptions on ε_i	Zero mean, constant variance, independence	Normality, independence, constant variance
Estimation of σ^2	Obtained indirectly from residuals	Explicitly estimated within the model
Robustness	Does not require normality, more robust under misspecification	More precise if distributional assumptions hold
Small sample performance	Less efficient	More efficient under normality
Computational complexity	Simple and fast	More computationally demanding

Table 2.5: Comparison between OLS and MLE

Conclusion: When to choose which method?

- If the error distribution is unknown and robustness is desired, use **OLS**.
- If the errors are normally distributed, **MLE yields the same estimates as OLS** for β_0 and β_1, but also allows exact likelihood-based inference.
- To jointly estimate regression coefficients and error variance σ^2, use **MLE**.
- For small samples under normality, **MLE is more efficient**.
- For large-scale problems, **OLS is computationally lighter**.

In practice, **OLS is the standard choice** because of its simplicity and reliable performance under mild assumptions. However, **MLE is preferred** in probabilistic modeling, econometrics, and machine learning when full distributional information about the errors is essential.

Example: Modeling Annual Salary with Work Experience

Problem Statement

A company wants to study the relationship between employees' annual salary (Y, in thousands of euros) and their years of work experience (X). The company collected data from 5 employees:

X_i (Years of Experience)	Y_i (Salary in k€)
1	35
2	37
3	40
4	46
5	50

We model the relationship using the simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Questions

1. Estimate the regression parameters β_0 and β_1 using:
   - Ordinary Least Squares (OLS)
   - Maximum Likelihood Estimation (MLE)
2. Compare OLS and MLE estimates, highlighting similarities and differences in slope, intercept, and error variance.

Solution

1. Parameter Estimation

Step 1: Compute Means :

$$\bar{X} = 3, \quad \bar{Y} = 41.6$$

Step 2: Compute Slope and Intercept (OLS) :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{5}(X_i - \bar{X})^2} = \frac{39}{10} = 3.9$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 29.9$$

Step 3: Maximum Likelihood Estimation (MLE) :
For normally distributed errors, the MLE for slope and intercept is identical to OLS:

$$\hat{\beta}_0^{MLE} = 29.9, \quad \hat{\beta}_1^{MLE} = 3.9$$

Step 4: Estimate Error Variance :
Residuals: $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$:

$$\hat{Y}_i = 29.9 + 3.9X_i = [33.8, 37.7, 41.6, 45.5, 49.4]$$

$$e_i = Y_i - \hat{Y}_i = [1.2, -0.7, -1.6, 0.5, 0.6]$$

$$\sum_{i=1}^{5} e_i^2 = 5.1$$

- OLS estimate of error variance:

$$\hat{\sigma}^2_{OLS} = \frac{\sum_{i=1}^{5} e_i^2}{n-2} = \frac{5.1}{3} \approx 1.7$$

- MLE estimate of error variance:

$$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{5} e_i^2}{n} = \frac{5.1}{5} = 1.02$$

2. Comparison of OLS and MLE

- **Slope and Intercept:** Identical for OLS and MLE:

$$\hat{\beta}_0 = 29.9, \quad \hat{\beta}_1 = 3.9$$

- **Error Variance:** OLS uses $(n-2)$ in the denominator (unbiased), MLE uses n. Therefore:

$$\hat{\sigma}^2_{OLS} = 1.7 > \hat{\sigma}^2_{MLE} = 1.02$$

- Both methods yield the same fitted regression line:

$$\hat{Y} = 29.9 + 3.9X$$

Conclusion: Each additional year of experience increases the annual salary by 3.9k€. OLS and MLE give the same slope and intercept, but differ in the error variance estimate.

Practical Work 3: Modeling Annual Salary with Work Experience in R

Listing 2.4: Comparison of OLS and MLE for linear regression: estimating salary based on experience.

```

# Create the data (Example: Salary vs Experience)
experience <- c(1, 2, 3, 4, 5)
salary <- c(35, 37, 40, 46, 50)

# --- OLS regression ---
model_ols <- lm(Salary ~ Experience, data = data.frame(Experience
   =experience, Salary=salary))
summary(model_ols)

# Extract OLS coefficients
beta0_ols <- coef(model_ols)[1]
beta1_ols <- coef(model_ols)[2]
cat("OLS Estimates:\nBeta0 =", beta0_ols, "\nBeta1 =", beta1_ols,
    "\n")

# --- Maximum Likelihood Estimation (MLE) ---
log_likelihood <- function(params) {
  beta0 <- params[1]
  beta1 <- params[2]
  sigma2 <- params[3]

  if (sigma2 <= 0) return(Inf)

  residuals <- salary - (beta0 + beta1 * experience)
  logL <- - (length(salary)/2)*log(2*pi*sigma2) - (1/(2*sigma2))*
     sum(residuals^2)
  return(-logL)  # negative log-likelihood for minimization
}

init_params <- c(beta0=0, beta1=0.1, sigma2=1)
result <- optim(par=init_params, fn=log_likelihood, method="L-
   BFGS-B",
                lower=c(-Inf, -Inf, 1e-6))

if (result$convergence == 0) {
  cat("MLE Estimates:\nBeta0 =", result$par[1], "\nBeta1 =",
     result$par[2], "\nSigma2 =", result$par[3], "\n")
} else {
  cat("Optimization did not converge.\n")
}

# --- Plot OLS and MLE regression lines ---
plot(experience, salary, pch=16, col="blue", main="OLS vs MLE:
   Salary vs Experience",
     xlab="Experience (years)", ylab="Salary ( k  )", ylim=c
        (30,55))
abline(a=beta0_ols, b=beta1_ols, col="red", lwd=2, lty=1)
```

```
abline(a=result$par[1], b=result$par[2], col="green", lwd=2, lty
   =2)
legend("topleft", legend=c("OLS","MLE"), col=c("red","green"),
   lty=c(1,2), lwd=2)
```

The R implementation produces the regression lines in Figure 2.3.

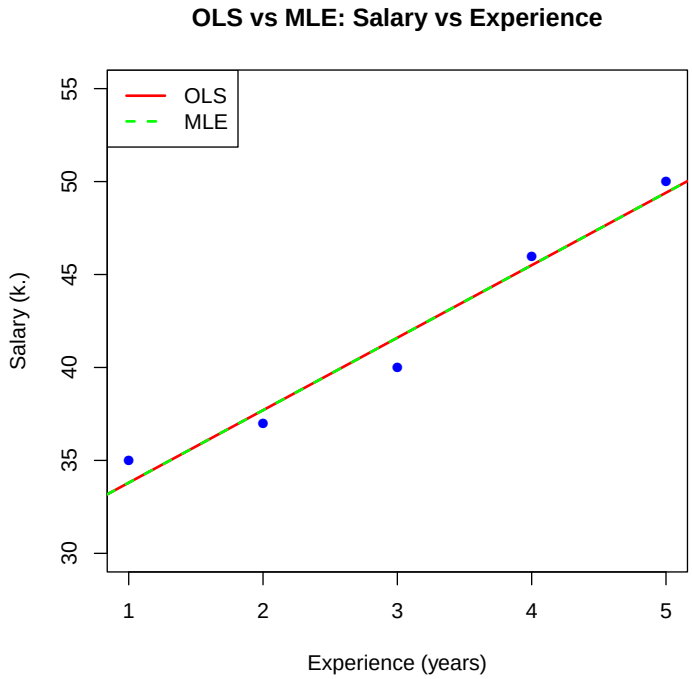

Figure 2.3: Regression lines obtained by OLS (red) and MLE (green dashed) for the salary vs experience example.

Results and Discussion

The regression of *Salary* on *Experience* was performed using both **OLS** and **MLE**. The results from R are:

- OLS: $\hat{\beta}_0 = 29.9$, $\hat{\beta}_1 = 3.9$, residual SE ≈ 1.304.
- MLE: $\hat{\beta}_0 = 29.9$, $\hat{\beta}_1 = 3.9$, $\hat{\sigma}^2 = 1.020$.

Interpretation: Each additional year of experience increases salary by 3.9 k€. The intercept represents the predicted salary for zero experience (29.9 k€).

Comparison: OLS and MLE give identical slope and intercept estimates, while variance differs due to denominator choice ($(n-2)$ vs n).

Conclusion: There is a strong positive relationship between experience and salary, and the linear model fits the data well. Both estimation methods confirm this result.

2.5 Inference and Extensions

The validity of inference procedures strongly depends on the assumptions introduced in Chapter 1, such as error normality, homoscedasticity, and independence of observations.

2.5.1 Linear Hypothesis Testing and Fisher's F-Test

In linear regression, we often want to test whether a certain combination of regression coefficients equals zero or another specified value. This is called a **linear hypothesis test**.

Formulation of a Linear Hypothesis

Consider the general linear model:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I),$$

where Y is the $n \times 1$ response vector, X is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression parameters, and $\boldsymbol{\varepsilon}$ is the vector of errors.

A linear hypothesis can be written as:

$$H_0 : C\boldsymbol{\beta} = d, \qquad H_1 : C\boldsymbol{\beta} \neq d,$$

where:

- C is a known $q \times p$ matrix specifying the linear combination of coefficients to test,
- d is a $q \times 1$ vector of constants (often zero),
- H_0 is the null hypothesis and H_1 the alternative.

Fisher's F-Test

The classical test statistic for a linear hypothesis is:

$$F = \frac{(C\hat{\boldsymbol{\beta}} - d)^\top \left[C(X^\top X)^{-1} C^\top\right]^{-1} (C\hat{\boldsymbol{\beta}} - d)/q}{\hat{\sigma}^2},$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$, $\hat{\sigma}^2$ is the unbiased estimate of the error variance, and q is the number of linear restrictions. Under H_0, this statistic follows an F-distribution:

$$F \sim \mathcal{F}(q, n - p),$$

where n is the sample size and p is the number of regression parameters.

Interpretation:

- A **large** F value indicates evidence against H_0.
- A **small** F value indicates insufficient evidence to reject H_0.
- For $q = 1$, the F-statistic is equivalent to the square of the t-statistic: $F = t^2$.

Key Quantities in Regression

Before performing inference, we define the key quantities that describe the variability of the response variable and how much is explained by the regression model:

- **Total sum of squares** (variation of Y around its mean):

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- **Regression sum of squares** (variation explained by the model):

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- **Error sum of squares** (unexplained variation):

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- **Coefficient of determination**:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

- **Mean squares**:

$$MSR = \frac{SSR}{p-1}, \quad MSE = \frac{SSE}{n-p}$$

- **Regression F-statistic**:

$$F = \frac{MSR}{MSE} \sim \mathcal{F}_{p-1,n-p}$$

Remark 2.5.1. *In simple linear regression ($p = 2$), $MSR = SSR/1$ and $MSE = SSE/(n-2)$. The regression F-statistic reduces to $F = t^2$.*

Illustrative Example: Simple Linear Regression

Consider a simple linear regression with 5 observations:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, 5.$$

Data:

X	Y
1	2
2	3
3	5
4	4
5	6

We test:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Step 1: Estimate the coefficients:

$$\bar{X} = 3, \quad \bar{Y} = 4, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{5}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{5}(X_i - \bar{X})^2} = 0.9, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 1.3$$

Step 2: Compute predicted values and sums of squares:

$$\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_5)^\top = (2.2, 3.1, 4.0, 4.9, 5.8)^\top$$

$$SSR = \sum_{i=1}^{5}(\hat{Y}_i - \bar{Y})^2 = 0.99, \quad SSE = \sum_{i=1}^{5}(Y_i - \hat{Y}_i)^2 = 1.9$$

$$MSR = SSR/1 = 0.99, \quad MSE = SSE/3 \approx 0.633$$

Step 3: Compute the F-statistic:

$$F = \frac{MSR}{MSE} = \frac{0.99}{0.633} \approx 12.79$$

Step 4: Decision rule:

- Degrees of freedom: $(1, 3)$
- Critical value at $\alpha = 0.05$: $F_{0.05,1,3} \approx 10.13$
- Since $F = 12.79 > 10.13$, we **reject** H_0.

Note: Simple linear regression with one predictor can be interpreted as a special case of one-way ANOVA. The regression F-test for $H_0 : \beta_1 = 0$ is algebraically equivalent to the ANOVA F-test: the variation explained by the predictor corresponds to between-group variability, and the residual variation corresponds to within-group variability.

2.5.2 One-Way Fixed Factor ANOVA Model

Consider a one-way ANOVA with k groups and n_i observations in group i:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

where μ is the overall mean and τ_i represents the fixed effect of group i.

The corresponding F-test is:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0 \quad \text{vs} \quad H_1 : \text{at least one } \tau_i \neq 0,$$

with the statistic:

$$F = \frac{MS_B}{MS_W} = \frac{SSB/(k-1)}{SSW/(N-k)} \sim \mathcal{F}_{k-1,\,N-k},$$

where

$$SSB = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2, \quad SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2, \quad N = \sum_{i=1}^{k} n_i.$$

Remark 2.5.2. *The one-way ANOVA is a special case of the general linear model where the explanatory variable is categorical rather than quantitative.*

Example: Weight Loss by Diet

In this example, we examine whether different diet types affect weight loss using a one-way ANOVA.

Problem Statement

We are interested in testing whether diet type has a significant effect on weight loss. The observed data (weight loss in kg) for three diet types are summarized in the following table:

Diet	Weight Loss (kg)
A	1.2, 1.5, 1.3, 1.4, 1.6
B	2.3, 2.5, 2.7, 2.6, 2.8
C	0.9, 1.1, 1.0, 1.2, 0.8

Questions

1. What are the null and alternative hypotheses?
2. What is the overall mean of all observations?
3. What are the group means and sums of squares (SSB, SSW, SST)?
4. What are the mean squares MS_B and MS_W?
5. What is the F-statistic?
6. What does the ANOVA table look like?
7. What is the conclusion at $\alpha = 0.05$ and its interpretation?

Solution

1. Hypotheses:

$H_0 : \mu_A = \mu_B = \mu_C$ (no effect of diet), H_1 : At least one group mean differs.

2. Overall Mean:

$$\bar{Y} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}}{N} = \frac{24.9}{15} = 1.66$$

3. Group Means and Sums of Squares:

$$\bar{Y}_A = 1.40, \quad \bar{Y}_B = 2.58, \quad \bar{Y}_C = 1.00$$

$$SSB = \sum_{i=1}^{k} n_i \, (\bar{Y}_i - \bar{Y})^2 = 6.748,$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = 0.348,$$

$$SST = SSB + SSW = 7.096$$

4. Mean Squares:

$$MS_B = \frac{SSB}{k-1} = 3.374, \qquad MS_W = \frac{SSW}{N-k} = 0.029$$

5. F-Test:

$$F = \frac{MS_B}{MS_W} = 116.34$$

6. ANOVA Table:

Source	SS	df	MS	F
Between groups	6.748	2	3.374	116.34
Within groups	0.348	12	0.029	
Total	**7.096**	**14**		

7. Conclusion: At a significance level of $\alpha = 0.05$, the critical value is $F_{0.05,2,12} = 3.89$. Since $116.34 > 3.89$, we **reject** H_0.
Interpretation: Diet type has a highly significant effect on weight loss.

Practical Work 4: Weight Loss by Diet in R

Listing 2.5: One-Way ANOVA – Weight Loss by Diet.

```
# ================================
# Example: One-Way ANOVA      Weight Loss by Diet
# ================================

# 1. Create the data
diet <- factor(c(rep('A',5), rep('B',5), rep('C',5)))
weight_loss <- c(1.2,1.5,1.3,1.4,1.6,
                 2.3,2.5,2.7,2.6,2.8,
                 0.9,1.1,1.0,1.2,0.8)

data <- data.frame(diet, weight_loss)

# 2. Descriptive statistics
tapply(weight_loss, diet, mean)
tapply(weight_loss, diet, sd)

# 3. Perform one-way ANOVA
anova_result <- aov(weight_loss ~ diet, data = data)
summary(anova_result)

# 4. Visualize the data
# ------------------------------
# Boxplot with mean points and colors

boxplot(weight_loss ~ diet,
        data = data,
        col = c("#99CCFF", "#FFCC99", "#CCFFCC"),
```

```
        border = "gray40",
        xlab = "Diet Type",
        ylab = "Weight Loss (kg)",
        cex.main = 1.2, cex.lab = 1)

# Add group means as red points
group_means <- tapply(weight_loss, diet, mean)
points(1:3, group_means, col = "red", pch = 19, cex = 1.3)
lines(1:3, group_means, col = "red", lty = 2)

# 5. Add F statistic and p-value on the plot
anova_summary <- summary(anova_result)[[1]]
F_value <- round(anova_summary["F value"][1], 2)
p_value <- signif(anova_summary["Pr(>F)"][1], 3)

legend("topleft",
       legend = paste("F =", F_value, ", p =", p_value),
       bty = "n", text.col = "black", cex = 0.9)
```

The R code above (Listing 2.5) generates the boxplot shown in Figure 2.4, which visually compares the weight loss distributions across the three diet types and highlights the group means in red.

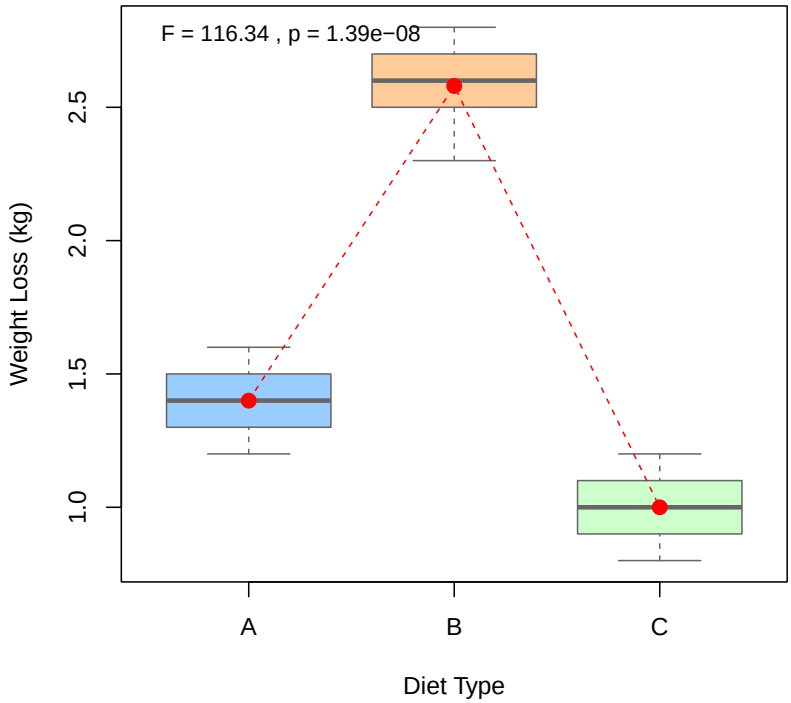

Figure 2.4: Comparison of Weight Loss Across Diet Types

Results and Discussion

The R output confirms the calculations. The ANOVA table shows that the diet factor is highly significant ($F(2, 12) \approx 116.3$, $p < 0.001$), providing very strong evidence that the mean weight loss differs across the three diets.
The boxplot illustrates the distribution of weight loss for each diet group (A, B, and C). Diet B exhibits the highest median weight loss (around 2.6 kg) and the smallest variability, suggesting it is the most effective among the three. Diet A shows moderate weight loss (median around 1.4 kg), while Diet C results in the lowest weight loss (median around 1.0 kg). These visual results are consistent with the ANOVA test, which indicates a statistically significant difference between the diets.

Having covered simple linear regression and one-way ANOVA, we now extend these concepts to multiple linear regression and two-way ANOVA, allowing the analysis of several predictors and the interaction of two factors.

Chapter 3

Multiple Linear Regression Model and Two-Factor ANOVA

This chapter introduces multiple linear regression and two-factor ANOVA, extending simple linear regression and one-way ANOVA to situations involving several explanatory variables and two categorical factors.

3.1 Multiple Linear Regression Model

The fundamental principles of multiple linear regression extend those introduced in simple linear regression (SLR). Multiple linear regression is one of the most widely used statistical tools for analyzing multidimensional data. As a special case of the general linear model, it represents the natural extension of simple regression.

Multiple linear regression models a dependent variable Y as a linear function of several explanatory variables $X_1, X_2, \ldots, X_p$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{3.1}$$

In matrix form:

$$Y = X\beta + \varepsilon. \tag{3.2}$$

Where:

- Y: $(n \times 1)$ vector of observations of the dependent variable,
- X: $(n \times (p+1))$ matrix of explanatory variables whose first column consists of 1s (intercept),
- β: $(p+1) \times 1$ vector of unknown parameters,
- ε: $(n \times 1)$ vector of random errors.

where n denotes the sample size and p the number of explanatory variables.
We assume that the error vector satisfies the following conditions:

$$E(\varepsilon) = 0, \qquad Var(\varepsilon) = \sigma^2 I_n.$$

These assumptions imply that the error terms have mean zero, constant variance σ^2, and are mutually uncorrelated.

If the errors are additionally assumed to be normally distributed, then the model can be written as

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{3.3}$$

Illustrative Example

Suppose we want to predict the academic performance Y of a student using two explanatory variables:

- X_1: number of study hours per week,
- X_2: number of class absences.

The multiple linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

where:

- Y_i: academic performance of student i,
- X_{i1}: study hours per week of student i,
- X_{i2}: number of absences of student i,
- β_0: expected academic performance when $X_1 = 0$ and $X_2 = 0$ (intercept),
- β_1: marginal effect of study hours on performance. For example, if $\beta_1 = 2$, an additional study hour increases expected performance by 2 points, holding absences constant,
- β_2: effect of absences. For instance, if $\beta_2 = -1$, each additional absence decreases expected performance by 1 point, holding study hours constant,
- ε_i: random error representing the difference between observed and predicted performance.

This example illustrates how multiple regression allows us to evaluate the partial effect of each explanatory variable on the response variable.

3.2 Ordinary Least Squares Method

3.2.1 Estimation of the Regression Coefficient β

We use the following matrix formulation (3.2):

$$Y = X\beta + \varepsilon.$$

The OLS estimator of the parameter vector β is obtained by minimizing the sum of squared errors:

$$\hat{\beta}_{OLS} = (X^\intercal X)^{-1} X^\intercal Y.$$

Although the model now includes multiple explanatory variables, the principle remains the same: minimizing the sum of squared errors provides a natural way to estimate the parameters of the multiple linear regression model, just as in the simple linear regression case.

This estimator has an important optimality property, which is formalized in the ***Gauss–Markov theorem****.*

3.2.2 Gauss–Markov Theorem

The theorem states that each component $\hat{\beta}_j$ of $\hat{\beta}$ for $j = 0, 1, \ldots, p$ has minimum variance among all linear unbiased estimators:

$$\mathrm{Var}(\tilde{\beta}_j) \geq \mathrm{Var}(\hat{\beta}_j).$$

Here, $\tilde{\beta}_j$ denotes any other unbiased estimator of β_j.

Proof of the Gauss–Markov Theorem

Let the OLS estimator be

$$\hat{\beta} = AY, \quad \text{where } A = (X^\mathsf{T}X)^{-1}X^\mathsf{T}.$$

Consider another estimator $\tilde{\beta}$ that is linear in Y:

$$\tilde{\beta} = (A + C)Y = \Big[(X^\mathsf{T}X)^{-1}X^\mathsf{T} + C\Big]Y.$$

Substituting $Y = X\beta + \varepsilon$ gives

$$\tilde{\beta} = \Big[(X^\mathsf{T}X)^{-1}X^\mathsf{T} + C\Big](X\beta + \varepsilon)$$

$$= (X^\mathsf{T}X)^{-1}X^\mathsf{T}X\beta + CX\beta + (X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon + C\varepsilon.$$

Since $(X^\mathsf{T}X)^{-1}X^\mathsf{T}X = I_{p+1}$, we obtain

$$\tilde{\beta} = \beta + CX\beta + (X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon + C\varepsilon.$$

The estimator $\tilde{\beta}$ is unbiased if

$$\mathbb{E}[\tilde{\beta}] = \beta,$$

which holds if and only if

$$CX = 0.$$

Under this condition,

$$\tilde{\beta} = \beta + \Big[(X^\mathsf{T}X)^{-1}X^\mathsf{T} + C\Big]\varepsilon.$$

$$\tilde{\beta} - \beta = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon + C\varepsilon,$$

Variance of $\tilde{\beta}$

$$\mathrm{Var}(\tilde{\beta}) = \mathbb{E}\Big[(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\mathsf{T}}\Big]$$

$$= \mathbb{E}\Big[\Big((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\varepsilon + C\varepsilon\Big)\Big((X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\varepsilon + C\varepsilon\Big)^{\mathsf{T}}\Big].$$

Because

$$\mathbb{E}[\varepsilon\varepsilon^{\mathsf{T}}] = \sigma^2 I_n,$$

we obtain

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2\Big[(X^{\mathsf{T}}X)^{-1} + (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}C^{\mathsf{T}} + CX(X^{\mathsf{T}}X)^{-1} + CC^{\mathsf{T}}\Big].$$

Since $CX = 0$, the cross terms vanish and we obtain

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2\Big[(X^{\mathsf{T}}X)^{-1} + CC^{\mathsf{T}}\Big].$$

Thus each component satisfies

$$\mathrm{Var}(\tilde{\beta}_j) \geq \mathrm{Var}(\hat{\beta}_j), \qquad j = 0, 1, \ldots, p.$$

Therefore $\hat{\beta}$ is the linear unbiased estimator with minimum variance, that is, the BLUE (Best Linear Unbiased Estimator).

Remark 3.2.1. *For the Gauss–Markov theorem to hold, it is required that the errors have zero mean,* $\mathbb{E}[\varepsilon] = 0$*, that* $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$*, and that the design matrix* X *has full column rank.*

Remark 3.2.2. *The properties of the OLS estimator in multiple linear regression mirror those in simple regression. They rely on key assumptions such as zero-mean errors, homoscedasticity, and full column rank of the design matrix. Under these conditions, the OLS estimator is unbiased, consistent, and efficient among linear unbiased estimators.*

3.2.3 Properties of the OLS Estimator

Unbiasedness

The OLS estimator is *unbiased* under the assumption that the error term has zero mean:

$$\mathbb{E}[\varepsilon] = 0.$$

Under this condition,

$$\mathbb{E}[\hat{\beta}] = \beta.$$

This means that, over repeated samples, the mean of $\hat{\beta}$ equals the true parameter value.

Efficiency

The OLS estimator is *efficient* among all linear unbiased estimators. According to the Gauss–Markov theorem, if the following assumptions hold:

- The errors have zero mean: $\mathbb{E}[\varepsilon] = 0$,
- The errors have constant variance (homoscedasticity),

$$\mathrm{Var}(\varepsilon) = \sigma^2 I,$$

- The design matrix X has full column rank,

then the OLS estimator is the Best Linear Unbiased Estimator (BLUE), meaning that it has the smallest variance among all linear unbiased estimators.

3.2.4 Estimation of the Variance of $\hat{\beta}_{\mathrm{OLS}}$

The variance–covariance matrix of the OLS estimator is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^\mathsf{T} X)^{-1}.$$

This result holds under the classical assumption that the error vector satisfies

$$\mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Severe multicollinearity among explanatory variables makes the matrix $X^\mathsf{T} X$ nearly singular, which increases the variance of the OLS estimates and reduces the precision of the estimation.

Estimation of the Error Variance s^2

Since the true error variance σ^2 is unknown, it is estimated by the residual variance:

$$s^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Here n denotes the number of observations and p the number of explanatory variables (excluding the intercept). The denominator $n-p-1$ corresponds to the degrees of freedom remaining after estimating the $p+1$ regression parameters.
This adjustment ensures that s^2 is an unbiased estimator of the error variance σ^2.

Example: Student Performance Regression

Problem Statement

We want to model the academic performance Y of students as a function of two explanatory variables:

- X_1: number of study hours per week,
- X_2: number of tutoring sessions attended.

The observed data for five students are:

Study Hours (per week)	Tutoring Sessions	Score (out of 100)
5	0	65
10	1	75
8	2	78
12	1	85
15	3	95

The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \qquad i = 1, \ldots, 5.$$

The goal is to estimate $\beta_0, \beta_1, \beta_2$ using Ordinary Least Squares (OLS) and compute the coefficient of determination R^2.

Questions

1. Write the regression model in matrix form.
2. Compute $X^\mathsf{T}X$ and $X^\mathsf{T}Y$.
3. Find the OLS estimator $\hat{\beta}_{OLS} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$.
4. Provide the estimated regression equation.
5. Interpret the coefficients.
6. Compute R^2 and interpret the result.

Solution

1. Matrix Form

$$Y = X\beta + \varepsilon, \quad X = \begin{bmatrix} 1 & 5 & 0 \\ 1 & 10 & 1 \\ 1 & 8 & 2 \\ 1 & 12 & 1 \\ 1 & 15 & 3 \end{bmatrix}, \quad Y = \begin{bmatrix} 65 \\ 75 \\ 78 \\ 85 \\ 95 \end{bmatrix}.$$

2. Intermediate Calculations The required sums (for $i = 1, \ldots, 5$) are

$$\sum_{i=1}^{5} X_{i1} = 50, \qquad \sum_{i=1}^{5} X_{i2} = 7,$$

$$\sum_{i=1}^{5} X_{i1}^2 = 558, \qquad \sum_{i=1}^{5} X_{i2}^2 = 15, \qquad \sum_{i=1}^{5} X_{i1}X_{i2} = 83.$$

Thus,

$$X^\mathsf{T}X = \begin{bmatrix} 5 & 50 & 7 \\ 50 & 558 & 83 \\ 7 & 83 & 15 \end{bmatrix}.$$

Next,

$$\sum_{i=1}^{5} Y_i = 398,$$

$$\sum_{i=1}^{5} X_{i1}Y_i = 4144, \qquad \sum_{i=1}^{5} X_{i2}Y_i = 601.$$

Hence

$$X^{\mathsf{T}}Y = \begin{bmatrix} 398 \\ 4144 \\ 601 \end{bmatrix}.$$

3. OLS Estimator

$$\hat{\beta}_{OLS} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$$

which gives approximately

$$\hat{\beta}_{OLS} \approx \begin{bmatrix} 53.92 \\ 2.14 \\ 3.08 \end{bmatrix}.$$

4. Estimated Regression Equation

$$\hat{Y}_i \approx 53.92 + 2.14\, X_{i1} + 3.08\, X_{i2}.$$

5. Interpretation

- $\hat{\beta}_0 = 53.92$: The baseline expected score when a student has 0 study hours and 0 tutoring sessions.
- $\hat{\beta}_1 = 2.14$: Each additional study hour per week increases the expected score by about 2.14 points, holding tutoring sessions constant.
- $\hat{\beta}_2 = 3.08$: Each additional tutoring session increases the expected score by about 3.08 points, holding study hours constant.

6. Goodness of Fit

The mean score is

$$\bar{Y} = \frac{65 + 75 + 78 + 85 + 95}{5} = 79.6.$$

The total sum of squares is

$$SST = \sum_{i=1}^{5} (Y_i - \bar{Y})^2 = 503.2.$$

Using the predicted values obtained from the regression model, the residual sum of squares is

$$SSR \approx 17.8.$$

Therefore,

$$R^2 = 1 - \frac{SSR}{SST}$$

$$R^2 = 1 - \frac{17.8}{503.2} \approx 0.9646.$$

Conclusion. The model explains about 96.5% of the variability in student scores. Both study hours and tutoring sessions have a positive effect on performance, and the estimated coefficients are consistent with the results obtained using the `lm()` function in `R`.

Practical Work 5: Student Performance Regression in R

Listing 3.1: Multiple Linear Regression in R: Student Performance Example.

```r
# 1. Creating the Data
study_hours <- c(5,10,8,12,15)
tutoring    <- c(0,1,2,1,3)
score       <- c(65,75,78,85,95)

data <- data.frame(study_hours, tutoring, score)

# 2. Fitting the Multiple Linear Regression Model
model <- lm(score ~ study_hours + tutoring, data = data)

# 3. Model Summary and Coefficients
summary(model)
coef(model)

# 4. Predicted Values
predicted_values <- predict(model)

# 5. Observed vs. Predicted with Residuals
results <- data.frame(
  Observed  = score,
  Predicted = predicted_values,
  Residual  = score - predicted_values
)
print(results)

# 6. Manual Computation of R^2
Y_mean <- mean(score)
SS_tot <- sum((score - Y_mean)^2)
SS_res <- sum((score - predicted_values)^2)
R2 <- 1 - SS_res / SS_tot
cat("Manually Computed R^2 =", R2, "\n")

# 7. Scatterplot of Predicted vs Observed Scores
plot(predicted_values, score,
```

```
      xlab = "Predicted Scores",
      ylab = "Observed Scores",
      pch = 19, col = "blue")
abline(a = 0, b = 1, col = "red", lwd = 2, lty = 2)
```

The R code produces a scatterplot comparing the predicted scores with the observed scores obtained from the regression model. The resulting graphical representation is displayed in Figure 3.1.

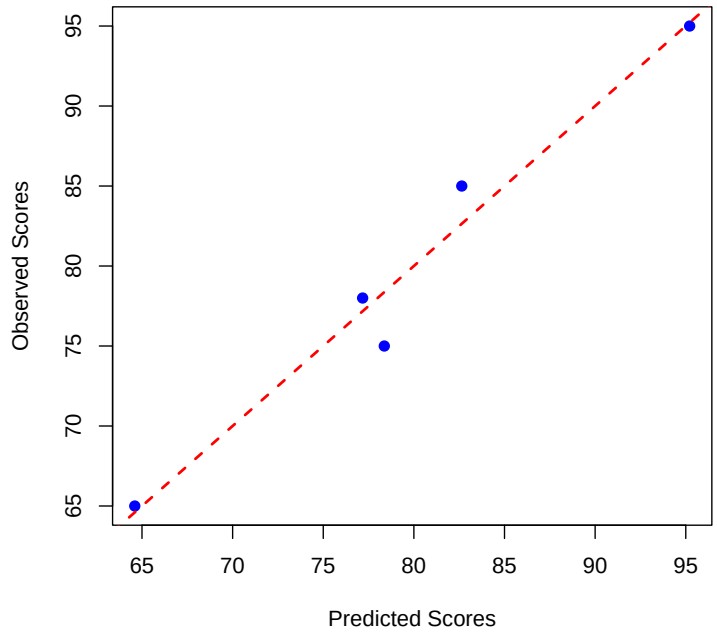

Figure 3.1: Observed vs. Predicted Student Scores from the Multiple Linear Regression Model

Results and Discussion

The figure above compares the observed student scores with the scores predicted by the multiple linear regression model. Each point represents one observation.
The dashed line corresponds to the reference line $y = x$, which represents perfect predictions. Points lying close to this line indicate that the predicted values are very close to the observed scores.
In this example, most observations lie near the diagonal line, indicating that the regression model provides a good approximation of the actual student scores. This visual assessment is consistent with the high coefficient of determination $R^2 = 0.9646$, which shows that approximately 96.5% of the variability in student performance is explained by the explanatory variables (study hours and tutoring sessions).
The small deviations from the reference line correspond to the residuals of the model, representing the difference between the observed and predicted values.

3.3 Maximum Likelihood Estimation

When the errors are assumed to be Gaussian, the Maximum Likelihood estimators of the regression coefficients coincide with the Ordinary Least Squares (OLS) estimators.

Let $Y_i \in \mathbb{R}$ be the response variable and $X_i \in \mathbb{R}^p$ the vector of explanatory variables (including the intercept) for observation i, $i = 1, \ldots, n$.
The model is assumed to satisfy:

$$Y_i = X_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \text{i.i.d.}$$

In matrix form:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathcal{N}_n(0, \sigma^2 I_n).$$

Where:

- $Y \in \mathbb{R}^n$: vector of observations,
- $X \in \mathbb{R}^{n \times p}$: design matrix (full rank p),
- $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$: vector of errors.

Under the Normality Assumption, $Y \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n)$. Its density is:

$$f(Y; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta})\right)$$

The Likelihood Function is:

$$L(\boldsymbol{\beta}, \sigma^2) = f(Y; \boldsymbol{\beta}, \sigma^2)$$

And the Log-Likelihood:

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta})$$

Remark 3.3.1. *To Obtain the Maximum Likelihood estimators, we maximize* $\log L(\boldsymbol{\beta}, \sigma^2)$ *with respect to* $\boldsymbol{\beta}$ *and* σ^2.

3.3.1 Estimation of the Regression Coefficient β

Maximizing the Log-Likelihood is equivalent to minimizing:

$$(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta})$$

which yields:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$$

This is exactly the OLS estimator.
Note: This requires that the regression matrix X *has full column rank so that* $(X^\top X)^{-1}$ *exists.*

3.3.2 Estimation of the Variance

By substituting $\hat{\boldsymbol{\beta}}$ into the Log-Likelihood and maximizing with respect to σ^2, we obtain:

$$\hat{\sigma}^2 = \frac{1}{n}(Y - X\hat{\boldsymbol{\beta}})^T(Y - X\hat{\boldsymbol{\beta}})$$

Note: This estimator is biased. The unbiased estimator divides by $(n-k)$ instead of n, where $k = p+1$ is the number of estimated regression parameters including the intercept. This estimator is consistent.

Intuitive Interpretation

Maximum Likelihood Estimation is based on a simple idea: *Choose the parameters that make the observed data the most probable.* In our case:

- The Normality assumption of the errors provides an explicit form of the probability density of Y.
- Maximizing this density amounts to minimizing the sum of squared residuals.

Thus, MLE and OLS coincide for $\boldsymbol{\beta}$, and $\hat{\sigma}^2$ measures the dispersion of the residuals around the fitted regression line.

Example: Two-Predictor Regression

Problem statement

We illustrate the Maximum Likelihood Estimation (MLE) in the context of a *multiple linear regression* with two predictors. The response variable Y depends on

$$X_1 : \text{Predictor 1}, \qquad X_2 : \text{Predictor 2}.$$

The observed data are

Observation	X_1	X_2	Y
1	1	2	6
2	2	1	7
3	3	4	15
4	4	3	16
5	5	5	22

The model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

with independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The goal is to obtain the Maximum Likelihood estimators of $\beta_0, \beta_1, \beta_2$ and σ^2.

Questions

1. Write the regression model in matrix form.
2. Derive the log-likelihood of the model.

3. Show that maximizing the log-likelihood w.r.t. $\boldsymbol{\beta}$ is equivalent to minimizing the sum of squared residuals.
4. Compute the estimator $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$.
5. Compute the Residuals and R^2.
6. Find the estimator of σ^2.

Solution

1. Matrix Form :

$$Y = X\boldsymbol{\beta} + \varepsilon, \quad X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 3 & 4 \\ 1 & 4 & 3 \\ 1 & 5 & 5 \end{bmatrix}, \quad Y = \begin{bmatrix} 6 \\ 7 \\ 15 \\ 16 \\ 22 \end{bmatrix}.$$

2. Log-Likelihood : Under the Gaussian assumption:

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta}).$$

3. Maximization w.r.t. $\boldsymbol{\beta}$: Maximizing $\log L$ with respect to $\boldsymbol{\beta}$ is equivalent to minimizing $(Y - X\boldsymbol{\beta})^T(Y - X\boldsymbol{\beta})$. Thus:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$$

4. Estimator $\hat{\boldsymbol{\beta}}$:

$$X^T X = \begin{bmatrix} 5 & 15 & 15 \\ 15 & 55 & 53 \\ 15 & 53 & 55 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 66 \\ 239 \\ 237 \end{bmatrix}.$$

Hence:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.1333 \\ 2.7222 \\ 1.7222 \end{bmatrix}.$$

5. Residuals and R^2 : The fitted values are:

$$\hat{Y} = X\hat{\boldsymbol{\beta}} = \begin{bmatrix} 6.0333 \\ 7.0333 \\ 14.9222 \\ 15.9222 \\ 22.0889 \end{bmatrix}.$$

The residuals are:

$$\hat{\varepsilon} = Y - \hat{Y} = \begin{bmatrix} -0.0333 \\ -0.0333 \\ 0.0778 \\ 0.0778 \\ -0.0889 \end{bmatrix}.$$

The coefficient of determination is:

$$R^2 = 1 - \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (Y_i - \bar{Y})^2} = 0.9998757.$$

6. Estimator of Variance :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0.00495.$$

(Unbiased version: $\tilde{\sigma}^2 = \frac{1}{n-3}\sum \hat{\varepsilon}_i^2 = 0.01237$.)

Conclusion : The Maximum Likelihood estimators are:

$$\boxed{\hat{\beta}_0 = -0.1333}, \quad \boxed{\hat{\beta}_1 = 2.7222}, \quad \boxed{\hat{\beta}_2 = 1.7222}.$$

Thus, the fitted regression model is:

$$\hat{Y} = -0.1333 + 2.7222X_1 + 1.7222X_2,$$

with an excellent fit ($R^2 = 0.9999$).

Practical Work 6: Two-Predictor Regression in R

Listing 3.2: R code: Maximum Likelihood / OLS estimation.

```r
# --- Data
   ------------------------------------------------------------
X1 <- c(1, 2, 3, 4, 5)
X2 <- c(2, 1, 4, 3, 5)
Y  <- c(6, 7, 15, 16, 22)

data <- data.frame(Y, X1, X2)

# Fit the multiple linear regression
model <- lm(Y ~ X1 + X2, data = data)
summary(model)

# Matrix calculations for manual verification
X <- cbind(1, X1, X2)
XtX <- t(X) %*% X
XtY <- t(X) %*% Y
beta_hat <- solve(XtX) %*% XtY
beta_hat

# Compute residuals and R^2 manually
residuals <- Y - X %*% beta_hat
R2 <- 1 - sum(residuals^2)/sum((Y - mean(Y))^2)
residuals
R2
```

Results and Discussion

The model provides an **excellent fit** to the data, with estimated coefficients:

$$\hat{\beta}_0 = -0.1333, \quad \hat{\beta}_1 = 2.7222, \quad \hat{\beta}_2 = 1.7222.$$

An increase of one unit in X_1 raises Y by about 2.72 units, and one unit in X_2 increases Y by approximately 1.72 units. The intercept being close to zero indicates that the regression plane nearly passes through the origin.

$$\hat{Y} = \begin{bmatrix} 6.0333 \\ 7.0333 \\ 14.9222 \\ 15.9222 \\ 22.0889 \end{bmatrix}, \quad \hat{\varepsilon} = \begin{bmatrix} -0.0333 \\ -0.0333 \\ 0.0778 \\ 0.0778 \\ -0.0889 \end{bmatrix}.$$

The model achieves $R^2 = 0.9999$ and an estimated variance $\hat{\sigma}^2 = 0.00495$, confirming that nearly all the variation in Y is explained by the predictors and that residual errors are minimal.

Obs	$\mathbf{Y_i}$	$\hat{\mathbf{Y}}_\mathbf{i}$	$\hat{\varepsilon}_\mathbf{i}$
1	6	6.0333	-0.0333
2	7	7.0333	-0.0333
3	15	14.9222	0.0778
4	16	15.9222	0.0778
5	22	22.0889	-0.0889

Table 3.1: Observed vs. Predicted Values and Residuals

The near-zero residuals demonstrate the precision of the Maximum Likelihood estimation under the Gaussian error assumption.

3.4 Comparison Between OLS and MLE

We observe that the estimators of $\boldsymbol{\beta}$ obtained by Maximum Likelihood Estimation (MLE) and by Ordinary Least Squares (OLS) are identical. However, the estimators of the variance σ^2 differ: the OLS estimator is **unbiased**, while the MLE estimator is **biased** (though it is consistent as $n \to \infty$).

Method	**Estimator of β**	**Estimator of σ^2**
OLS	$(X^TX)^{-1}X^TY$	$\hat{\sigma}^2_{\text{OLS}} = \dfrac{\text{SSR}}{n-k}$
MLE	$(X^TX)^{-1}X^TY$	$\hat{\sigma}^2_{\text{MLE}} = \dfrac{\text{SSR}}{n}$

Remark 3.4.1. *Bias of the Variance Estimator in MLE.*
The Maximum Likelihood estimator of the variance is:

$$\hat{\sigma}^2_{MLE} = \frac{SSR}{n},$$

where $SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ *is the sum of squared residuals.*
This estimator is ***biased*** *because it does not account for the degrees of freedom lost in estimating the* k *parameters of* $\boldsymbol{\beta}$*. To obtain an* ***unbiased*** *estimator, we use the Ordinary Least Squares estimator:*

$$\hat{\sigma}^2_{OLS} = \frac{SSR}{n-k},$$

which corrects for this loss by dividing by $(n-k)$*, the effective number of degrees of freedom.*

Intuitive interpretation: *The MLE estimator treats* $\boldsymbol{\beta}$ *as if it were known (fixed), which underestimates the true variance. The OLS estimator adjusts for the fact that* $\boldsymbol{\beta}$ *is estimated from the data.*

Example: Modeling Student Exam Scores

Problem statement

We model the final score Y (out of 100) as a function of three predictors:

- X_1: hours of study per week,
- X_2: number of practice exercises completed,
- X_3: hours of sleep per night.

Observed data (five students):

Student	X_1 (Study)	X_2 (Exercises)	X_3 (Sleep)	Y (Score)
1	2	5	7	50
2	4	4	8	65
3	6	6	6	80
4	8	5	7	90
5	10	7	8	100

We assume the multiple linear regression model (3.3):

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n).$$

Questions

1. Write the regression model in matrix form.
2. Compute the OLS estimator $\hat{\boldsymbol{\beta}}$.
3. Estimate the variance σ^2 using both OLS and MLE methods.
4. Compare the predicted values with the observed scores.

Solution

1. Matrix form Let $n = 5$. Define

$$X = \begin{pmatrix} 1 & 2 & 5 & 7 \\ 1 & 4 & 4 & 8 \\ 1 & 6 & 6 & 6 \\ 1 & 8 & 5 & 7 \\ 1 & 10 & 7 & 8 \end{pmatrix}, \qquad Y = \begin{pmatrix} 50 \\ 65 \\ 80 \\ 90 \\ 100 \end{pmatrix}.$$

2. OLS estimator

$$\hat{\boldsymbol{\beta}}_{OLS} = (X^\top X)^{-1} X^\top Y$$

Compute $X^\top X$ and $X^\top Y$

$$X^\top X = \begin{pmatrix} 5 & 30 & 27 & 36 \\ 30 & 220 & 172 & 218 \\ 27 & 172 & 151 & 194 \\ 36 & 218 & 194 & 262 \end{pmatrix}, \qquad X^\top Y = \begin{pmatrix} 385 \\ 2560 \\ 2140 \\ 2780 \end{pmatrix}.$$

Numerical approximation:

$$\hat{\boldsymbol{\beta}}_{OLS} \approx \begin{pmatrix} 58.75 \\ 6.6667 \\ -1.25 \\ -2.0833 \end{pmatrix}.$$

3. Variance estimates

Residuals and predicted values:

$$\hat{Y} = X\hat{\boldsymbol{\beta}} \approx \begin{pmatrix} 51.25 \\ 63.75 \\ 78.75 \\ 91.25 \\ 100.00 \end{pmatrix}, \qquad r = Y - \hat{Y} \approx \begin{pmatrix} -1.25 \\ 1.25 \\ 1.25 \\ -1.25 \\ 0.00 \end{pmatrix}.$$

$$\mathrm{RSS} = \sum_{i=1}^{n} r_i^2 = 6.25.$$

Since $n = 5$ and $k = 4$ (including the intercept), we have $n - k = 1$. Thus,

$$\hat{\sigma}^2_{\mathrm{OLS}} = \frac{\mathrm{RSS}}{n-k} = \frac{6.25}{1} = 6.25, \qquad \hat{\sigma}^2_{\mathrm{ML}} = \frac{\mathrm{RSS}}{n} = \frac{6.25}{5} = 1.25.$$

The residual standard error is:

$$\widehat{\mathrm{RSE}} = \sqrt{\hat{\sigma}^2_{\mathrm{OLS}}} = \sqrt{6.25} = 2.5.$$

4. Comparison of predicted vs observed scores

The predicted values $\hat{Y}$ are very close to the observed scores Y. The residuals are small, and the model explains nearly all variability in Y ($R^2 \approx 0.996$).

Conclusion. The estimated regression equation is:

$$\hat{Y} = 58.75 + 6.67X_1 - 1.25X_2 - 2.08X_3.$$

The OLS and MLE for $\hat{\boldsymbol{\beta}}$ coincide, while the variance estimates differ only due to the degrees-of-freedom adjustment.

Remark: In small samples, when the number of observations n is close to the number of parameters k, the OLS variance estimator $\hat{\sigma}^2_{\text{OLS}} = \frac{\text{RSS}}{n-k}$ can become very unstable. It is important not to interpret the difference between OLS and MLE as a methodological anomaly in such cases.

Practical Work 7: Modeling Student Exam Scores in R

Listing 3.3: R code for computing OLS and MLE for the multiple regression example.

```r
# Data
study <- c(2,4,6,8,10)
exercises <- c(5,4,6,5,7)
sleep <- c(7,8,6,7,8)
score <- c(50,65,80,90,100)
data <- data.frame(study, exercises, sleep, score)

# Matrices
X <- cbind(1, study, exercises, sleep)
Y <- matrix(score, ncol=1)

# OLS via matrices
B_OLS <- solve(t(X) %*% X) %*% t(X) %*% Y

# Predicted values and residuals
Y_hat <- X %*% B_OLS
residuals <- Y - Y_hat

# RSS and variance estimates
n <- nrow(X)
k <- ncol(X)
RSS <- sum(residuals^2)
sigma2_OLS <- RSS/(n-k)    # unbiased OLS
sigma2_ML <- RSS/n         # ML

# Print results
print("Coefficients (B_OLS):")
print(round(B_OLS,6))

print("Predicted values (Y_hat):")
print(round(as.vector(Y_hat),6))

```

```r
print("Residuals:")
print(round(as.vector(residuals),6))

cat("RSS =", RSS, "\n")
cat("sigma2_OLS =", sigma2_OLS, "\n")
cat("sigma2_ML =", sigma2_ML, "\n")

# Check with lm()
model <- lm(score ~ study + exercises + sleep, data=data)
print(summary(model))
```

Results and Discussion

- **Estimated coefficients ($\hat{\beta}_{\mathbf{OLS}}$)**:

$$\hat{\beta}_0 = 58.75, \quad \hat{\beta}_1 = 6.67, \quad \hat{\beta}_2 = -1.25, \quad \hat{\beta}_3 = -2.08$$

  Interpretation: each additional hour of study increases the predicted score by about 6.67 points, while each additional exercise slightly decreases it by 1.25 points, and each additional hour of sleep decreases it by about 2.08 points, holding other factors constant.

- **Predicted values**:

$$\hat{Y} = (51.25,\ 63.75,\ 78.75,\ 91.25,\ 100.00)$$

- **Residuals**:

$$r = (-1.25,\ 1.25,\ 1.25,\ -1.25,\ 0)$$

- **Residual Sum of Squares (RSS)** $= 6.25, \quad \hat{\sigma}^2_{\text{OLS}} = 6.25, \quad \hat{\sigma}^2_{\text{MLE}} = 1.25$

The model provides an excellent fit, with a residual standard error of 2.5 and a coefficient of determination $R^2 \approx 0.996$. This example illustrates that while both OLS and MLE produce the same coefficient estimates, their variance estimators differ unless the fit is perfect.

3.5 Inference and Extensions

3.5.1 ANOVA with Quantitative and Qualitative Variables

Objective

In multiple linear regression, Analysis of Variance (ANOVA) provides a framework to decompose the total variability of the response variable Y and to assess the overall significance of the model. This decomposition applies whether the regressors are:

- **Quantitative variables** (e.g., income, study hours),
- **Qualitative variables** encoded as *dummy variables* (e.g., gender, region).

Model and Variance Decomposition

Consider the general multiple regression model (3.3):

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

where Y is the $n \times 1$ response vector, X the $n \times p$ design matrix (including the intercept), and $\boldsymbol{\beta}$ the $p \times 1$ vector of regression coefficients.
The total variability of Y can be decomposed as:

$$SST = SSR + SSE,$$

where:

- **Total Sum of Squares:** $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- **Regression Sum of Squares:** $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- **Residual Sum of Squares:** $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

In analogy with one-way ANOVA:

$SSR \equiv SSB$ (between-groups variability), $SSE \equiv SSW$ (within-groups variability).

ANOVA Table for Multiple Regression

Source	Sum of Squares	df	Mean Square	F
Regression	SSR	$p-1$	$MSR = \dfrac{SSR}{p-1}$	$F = \dfrac{MSR}{MSE}$
Residual	SSE	$n-p$	$MSE = \dfrac{SSE}{n-p}$	
Total	SST	$n-1$		

Global F-Test

The overall significance of the regression model is tested by:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \quad \text{vs.} \quad H_1 : \exists j \text{ such that } \beta_j \neq 0,$$

where β_0 is the intercept.
The test statistic is

$$F = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)} \sim \mathcal{F}_{p-1,\,n-p} \quad \text{under } H_0.$$

Remark 3.5.1.

- *This formulation is fully consistent with the F-test in simple regression ($p = 2$), where $F = t^2$ for a single predictor.*
- *The global F-test in multiple regression is algebraically equivalent to the ANOVA F-test: $SSR = SSB$ (between-groups) and $SSE = SSW$ (within-groups).*
- *The decomposition of variability allows us to quantify the explanatory power of the model using $R^2 = SSR/SST = 1 - SSE/SST$.*

Connection to Simple Regression and One-Way ANOVA

- In simple regression ($p = 2$), the regression F-test for $H_0 : \beta_1 = 0$ is equivalent to the one-way ANOVA F-test.
- In multiple regression with qualitative variables encoded as dummies, ANOVA provides a clear way to assess the combined effect of categorical factors on the response.

Example: Food Expenditure with Mixed Variables

Problem statement

We want to understand how **Food Expenditure** Y depends on:

- X_1: Income (quantitative),
- X_2: Region (qualitative with 3 categories: North, Center, South).

Objective: Use regression with dummy variables to model Y, decompose variance

$$SST = SSR + SSE,$$

and further split SSR into **between-groups variability** (SSB) due to region and **additional explained variability** (SSW) due to income.
We encode the qualitative variable **Region** using two dummy variables:

$$D_1 = \begin{cases} 1 & \text{if Center} \\ 0 & \text{otherwise} \end{cases}, \quad D_2 = \begin{cases} 1 & \text{if South} \\ 0 & \text{otherwise} \end{cases}$$

(North is the reference category).
The regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$$

Observed Data:

Person	X_1 (Income)	D_1	D_2	Y (Expenditure)
1	20	0	0	15
2	25	0	0	18
3	22	1	0	16
4	20	1	0	19
5	25	0	1	22
6	22	0	1	21

Questions

1. Estimate the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$.
2. Compute fitted values $\hat{Y}_i$.
3. Compute total, regression, and residual sums of squares (SST, SSR, SSE).

4. Decompose SSR into SSB (between groups, region) and SSW (additional explained by income).
5. Construct the ANOVA table.
6. Interpret the results: which factors significantly affect food expenditure?

Solution

1. Estimated Coefficients: Using OLS:

$$\hat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta}_0 \approx 9.39, \quad \hat{\beta}_1 \approx 0.32, \quad \hat{\beta}_2 \approx 1.47, \quad \hat{\beta}_3 \approx 4.68$$

2. Fitted Values:

$$\hat{Y} \approx (15.71,\ 17.29,\ 17.82,\ 17.18,\ 21.97,\ 21.03)$$

3. Variance Decomposition:

$$\bar{Y} = \frac{1}{6}\sum_{i=1}^{6} Y_i = 18.5, \quad SST = \sum_{i=1}^{6}(Y_i - \bar{Y})^2 = 37.5$$

$$SSR = \sum_{i=1}^{6}(\hat{Y}_i - \bar{Y})^2 \approx 29.89, \quad SSE = \sum_{i=1}^{6}(Y_i - \hat{Y}_i)^2 \approx 7.61$$

Check :

$$SST = SSR + SSE \quad (37.5 \approx 29.89 + 7.61)$$

4. Decomposition of SSR:

$$SSR = SSB + SSW$$

SSB ≈ 28.0 (between-groups variation due to Region),
$SSW \approx 1.89$ (additional explained by Income).

5. ANOVA Table:

Source	**SS**	**df**	**MS**	**F**
$Regression(SSR)$	29.89	3	9.96	2.62
$Residuals(SSE)$	7.61	2	3.805	
$Total(SST)$	37.50	5		

Notes: MSR $= SSR/(p-1) = 29.89/3 \approx 9.96$, $MSE = SSE/(n-p) = 7.61/2 \approx 3.805$.

$$F = \frac{MSR}{MSE} \approx 2.62 \sim \mathcal{F}_{3,2}$$

6. Interpretation:

- The model explains about 80% of the variability ($R^2 \approx 0.80$).
- With only 6 observations, the F-test ($F \approx 2.62, p \approx 0.29$) is not statistically significant.
- Coefficients suggest:
  - Each additional unit of income increases expenditure by about 0.32 units ($\hat{\beta}_1$).
  - Center households spend on average 1.47 units more than North households (at equal income).
  - South households spend on average 4.68 units more than North households (at equal income).
- None of these effects are significant at the 5% level due to the very small sample size ($df_{res} = 2$).
- Variance decomposition confirms the bounded sums: $SSB \approx 28.0$ (region effect), $SSW \approx 1.89$ (income effect), and $SSE \approx 7.61$ (residual).

Practical Work 8: Food Expenditure with Mixed Variables in R

Listing 3.4: Food Expenditure with Mixed Variables in R.

```r
# ================================
# Example: Food Expenditure with Mixed Variables
# ================================

# 1. Data preparation
income <- c(20, 25, 22, 20, 25, 22)
D1 <- c(0, 0, 1, 1, 0, 0)  # Center
D2 <- c(0, 0, 0, 0, 1, 1)  # South
expenditure <- c(15, 18, 16, 19, 22, 21)

data <- data.frame(Person = 1:6, income, D1, D2, expenditure)
print(data)

# 2. Fit the full regression model
mod <- lm(expenditure ~ income + D1 + D2, data = data)
summary(mod)

# 3. Predicted (fitted) values and residuals
data$fitted <- fitted(mod)
data$residuals <- resid(mod)
print(data)

# 4. Global variance decomposition (ANOVA)
anova(mod)
```

```r

# 5. Compute SST, SSR, and SSE manually
Ybar <- mean(data$expenditure)
SST <- sum((data$expenditure - Ybar)^2)
SSR <- sum((data$fitted - Ybar)^2)
SSE <- sum((data$expenditure - data$fitted)^2)
R2 <- SSR / SST

cat("\n--- Global Variance Decomposition ---\n")
cat("SST =", SST, " SSR =", SSR, " SSE =", SSE, "\n")
cat("Check: SST = SSR + SSE =", SSR + SSE, "\n")
cat("R-squared =", R2, "\n")

# 6. Decomposition of SSR into SSB (Region) and SSW (Income)
# Model 1: Region only (between-groups variation)
mod_region <- lm(expenditure ~ D1 + D2, data = data)

# Model 2: Region + Income (full model)
mod_full <- mod

# Compute SSR for each model
SSR_region <- sum((fitted(mod_region) - Ybar)^2)
SSR_full <- sum((fitted(mod_full) - Ybar)^2)

# Between-groups and within-groups components
SSB <- SSR_region
SSW <- SSR_full - SSR_region

cat("\n--- Decomposition of SSR ---\n")
cat("SSB (Between-Groups, Region):", SSB, "\n")
cat("SSW (Within-Groups, Income):", SSW, "\n")
cat("Check: SSB + SSW =", SSB + SSW, "    SSR =", SSR, "\n")

# Compare the two models with ANOVA
anova(mod_region, mod_full)

# 7. Plot observed vs fitted values with residuals (colors
   improved)
plot(data$Person, data$expenditure, pch = 19, col = "blue",
     ylim = c(14, 23), xlab = "Observation", ylab = "Expenditure"
        ,
     main = "Food Expenditure: Observed vs Fitted Values")
points(data$Person, data$fitted, pch = 19, col = "red")
abline(h = Ybar, col = "lightgreen", lty = 2, lwd = 2)  # mean
   line lighter green

# Draw residual lines in lighter gray
segments(data$Person, data$expenditure, data$Person, data$fitted,
         col = "lightgray", lwd = 1.5)

# Legend
```

```r
legend("topleft",
       legend = c("Observed Y", "Fitted Y", "Mean Y", "Residuals"
          ),
       col = c("blue", "red", "lightgreen", "lightgray"),
       pch = c(19, 19, NA, NA),
       lty = c(NA, NA, 2, 1),
       lwd = c(NA, NA, 2, 1),
       bty = "n")
```

The following plot illustrates the variance decomposition from the R code.

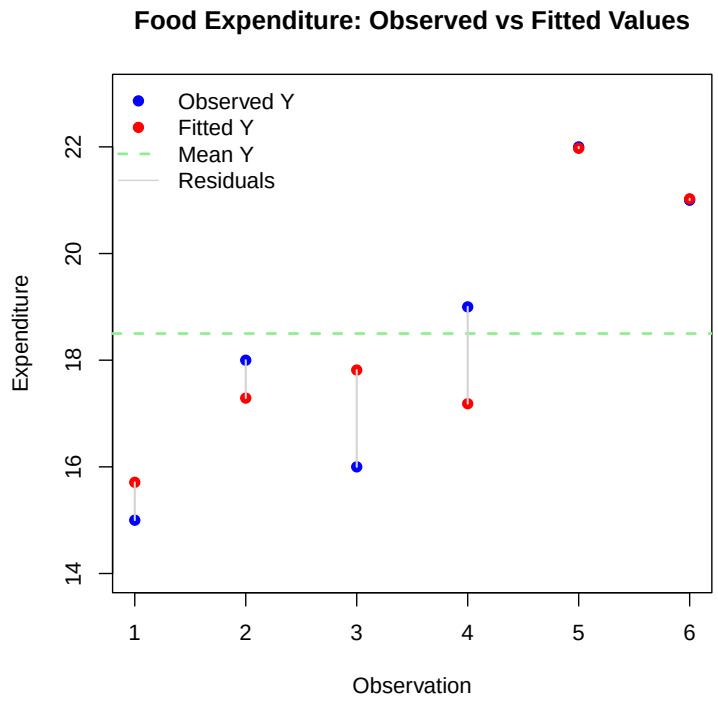

Figure 3.2: Variance decomposition: Food expenditure ($SST = SSR + SSE$)

Results and Discussion

The figure compares the observed food expenditures Y_i (blue points) with the fitted values $\hat{Y}_i$ obtained from the regression model. The horizontal dashed line represents the overall mean $\bar{Y}$.
The graph illustrates the classical variance decomposition

$$SST = SSR + SSE,$$

Graphically, the distance between Y_i and $\bar{Y}$ corresponds to the total deviation, which can be decomposed into the explained deviation $(\hat{Y}_i - \bar{Y})$ and the residual error $(Y_i - \hat{Y}_i)$. The proximity between observed and fitted values indicates that the model captures a substantial part of the variation in food expenditure.

3.5.2 Two-Factor ANOVA and Hierarchical ANOVA Models

Two-factor ANOVA extends the classical one-way ANOVA by studying the effect of two categorical factors on a quantitative response variable. It allows us to:

- Assess the main effect of each factor,
- Investigate whether there is an interaction effect between factors,
- Decompose the total variability into components attributable to each factor, their interaction, and residual error.

Hierarchical ANOVA models, also known as nested designs, are used when the levels of one factor are nested within the levels of another factor, such as students nested within schools, or plots nested within blocks.

Model Formulation: Two-Factor ANOVA

Consider two factors: A with a levels and B with b levels. Denote the observations as y_{ijk}, where $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n_{ij}$ is the replication index. The two-factor ANOVA model with interaction is written as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where:

- μ is the overall mean,
- α_i is the effect of level i of factor A,
- β_j is the effect of level j of factor B,
- $(\alpha\beta)_{ij}$ is the interaction effect between level i of A and level j of B,
- ε_{ijk} are independent random errors, assumed $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Assumptions

1. The errors are normally distributed with mean 0 and variance σ^2,
2. The observations are independent,
3. The effects are additive, unless interactions are explicitly included.

Hierarchical (Nested) ANOVA Model

When factor B is nested within factor A, the model takes the form:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},$$

where $\beta_{j(i)}$ represents the effect of level j of factor B nested within level i of factor A.

Variance Decomposition

For a two-factor ANOVA with interaction, the total sum of squares (SST) can be decomposed as:

$$SST = SSR_A + SSR_B + SSR_{AB} + SSE,$$

where:

- $SST = \sum_{i,j,k}(y_{ijk} - \bar{y}_{...})^2$ is the total variability,
- $SSR_A = nb\sum_i(\bar{y}_{i..} - \bar{y}_{...})^2$ is the variability explained by Factor A,
- $SSR_B = na\sum_j(\bar{y}_{.j.} - \bar{y}_{...})^2$ is the variability explained by Factor B,
- $SSR_{AB} = n\sum_i\sum_j(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$ is the variability explained by the interaction,
- $SSE = \sum_i\sum_j\sum_k(y_{ijk} - \bar{y}_{ij.})^2$ is the residual variability.

Degrees of Freedom and Test Statistics

The corresponding degrees of freedom are:

$$df_A = a - 1, \quad df_B = b - 1, \quad df_{AB} = (a-1)(b-1), \quad df_E = ab(n-1).$$

Mean squares are defined as:

$$MS_A = \frac{SSR_A}{df_A}, \quad MS_B = \frac{SSR_B}{df_B}, \quad MS_{AB} = \frac{SSR_{AB}}{df_{AB}}, \quad MSE = \frac{SSE}{df_E}.$$

The F-tests for significance of the effects are:

$$F_A = \frac{MS_A}{MSE}, \quad F_B = \frac{MS_B}{MSE}, \quad F_{AB} = \frac{MS_{AB}}{MSE}.$$

Example: Two-Factor ANOVA on Exam Scores

Problem Statement

We study the effect of **teaching method** (Factor A) and **study environment** (Factor B) on students' exam scores. Specifically, we aim to determine:

- Whether the teaching method significantly affects students' scores,
- Whether the study environment significantly affects scores,
- Whether there is a significant interaction between teaching method and study environment.

Data: Two teaching methods (A_1, A_2), two environments (B_1, B_2), and three students per cell.

	B_1	B_2
A_1	$50, 52, 50$	$60, 61, 59$
A_2	$55, 54, 56$	$65, 64, 66$

Solution

Step 1: Descriptive Statistics

For a two-factor design ($a = 2$, $b = 2$, $n = 3$), the cell means are:

	B_1	B_2
A_1	$\bar{y}_{11.} = 50.67$	$\bar{y}_{12.} = 60.00$
A_2	$\bar{y}_{21.} = 55.00$	$\bar{y}_{22.} = 65.00$

The marginal and overall means are:

$$\begin{aligned} \bar{y}_{1..} &= 55.33, \quad \bar{y}_{2..} = 60.00, \\ \bar{y}_{.1.} &= 52.83, \quad \bar{y}_{.2.} = 62.50, \\ \bar{y}_{...} &= 57.67. \end{aligned}$$

Step 2: Computation of Sums of Squares

The total variability is decomposed as:

$$SST = SSR_A + SSR_B + SSR_{AB} + SSE.$$

Using the classical formulas:

$$\begin{aligned} SSR_A &= nb \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2, \\ SSR_B &= na \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2, \\ SSR_{AB} &= n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2, \\ SSE &= \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2. \end{aligned}$$

Numerical results:

$$\begin{aligned} SSR_A &= 65.33, \\ SSR_B &= 280.33, \\ SSR_{AB} &= 0.33, \\ SSE &= 8.67, \\ SST &= 354.67. \end{aligned}$$

Step 3: Degrees of Freedom and Mean Squares

$$\begin{aligned} df_A &= a - 1 = 1, & MS_A &= SSR_A/df_A = 65.33, \\ df_B &= b - 1 = 1, & MS_B &= SSR_B/df_B = 280.33, \\ df_{AB} &= (a-1)(b-1) = 1, & MS_{AB} &= SSR_{AB}/df_{AB} = 0.33, \\ df_E &= ab(n-1) = 8, & MSE &= SSE/df_E = 1.083. \end{aligned}$$

Step 4: F-Tests for Significance

$$F_A = \frac{MS_A}{MSE} = 60.31, \quad F_B = \frac{MS_B}{MSE} = 258.77, \quad F_{AB} = \frac{MS_{AB}}{MSE} = 0.31.$$

The corresponding p-values (based on the $F(1, 8)$ distribution) are:

$$p_A = 5.4 \times 10^{-5}, \quad p_B = 2.2 \times 10^{-7}, \quad p_{AB} = 0.594.$$

Step 5: ANOVA Summary Table

Source	df	SS	MS	F	p-value
Factor A (Teaching method)	1	65.33	65.33	60.31	0.000054
Factor B (Environment)	1	280.33	280.33	258.77	0.00000022
Interaction A×B	1	0.33	0.33	0.31	0.594
Error	8	8.67	1.083		
Total	11	354.67			

Table 3.2: Two-Factor ANOVA results for exam scores.

Step 6: Interpretation

- The **teaching method** has a significant effect on exam scores ($F_A = 60.31$, $p < 0.001$).
- The **study environment** has an even stronger effect ($F_B = 258.77$, $p < 0.001$).
- The **interaction effect** between method and environment is not significant ($F_{AB} = 0.31$, $p = 0.594$).

Practical Work 9: Two-Factor ANOVA on Exam Scores in R

Listing 3.5: Two-Factor ANOVA on Exam Scores.

```r
# ================================
# Example: Two-Factor ANOVA on Exam Scores
# ================================

# Teaching method (A): 2 levels (A1, A2)
# Study environment (B): 2 levels (B1, B2)
# 3 students per cell

# --- 1. Create the dataset ---
teaching_method <- factor(rep(c("A1", "A2"), each = 6))
environment <- factor(rep(rep(c("B1", "B2"), each = 3), 2))
scores <- c(
  50, 52, 50,    # A1B1
  60, 61, 59,    # A1B2
```

```r
  55, 54, 56,   # A2B1
  65, 64, 66    # A2B2
)

data <- data.frame(teaching_method, environment, scores)
print(data)

# --- 2. Descriptive statistics ---
aggregate(scores ~ teaching_method + environment, data, mean)

# --- 3. Two-way ANOVA model ---
anova_model <- aov(scores ~ teaching_method * environment, data =
    data)
summary(anova_model)

# --- 4. Extract group means ---
library(dplyr)
means_table <- data %>%
  group_by(teaching_method, environment) %>%
  summarise(mean_score = mean(scores), .groups = 'drop')
print(means_table)

# --- 5. Verify overall means ---
grand_mean <- mean(data$scores)
grand_mean

# --- 6. Compute sum of squares manually (for verification) ---
A_means <- tapply(data$scores, data$teaching_method, mean)
B_means <- tapply(data$scores, data$environment, mean)
cell_means <- tapply(data$scores, list(data$teaching_method,
   data$environment), mean)

SS_A <- sum(3 * 2 * (A_means - grand_mean)^2)
SS_B <- sum(3 * 2 * (B_means - grand_mean)^2)
SS_AB <- 3 * sum((cell_means - outer(A_means, B_means, "+") +
   grand_mean)^2)
SS_Total <- sum((data$scores - grand_mean)^2)
SS_Error <- SS_Total - SS_A - SS_B - SS_AB

cat("\n--- Manual ANOVA Calculations ---\n")
cat("SS_A (Teaching Method):", SS_A, "\n")
cat("SS_B (Environment):", SS_B, "\n")
cat("SS_AB (Interaction):", SS_AB, "\n")
cat("SS_Error:", SS_Error, "\n")
cat("SS_Total:", SS_Total, "\n")

# --- 7. Degrees of freedom ---
df_A <- length(levels(data$teaching_method)) - 1
df_B <- length(levels(data$environment)) - 1
df_AB <- df_A * df_B
```

```r
df_E <- nrow(data) - length(levels(data$teaching_method)) *
   length(levels(data$environment))
df_Total <- nrow(data) - 1

# --- 8. Mean Squares and F values ---
MS_A <- SS_A / df_A
MS_B <- SS_B / df_B
MS_AB <- SS_AB / df_AB
MS_E <- SS_Error / df_E

F_A <- MS_A / MS_E
F_B <- MS_B / MS_E
F_AB <- MS_AB / MS_E

cat("\n--- F Values ---\n")
cat("F_A =", F_A, "\n")
cat("F_B =", F_B, "\n")
cat("F_AB =", F_AB, "\n")

# --- 9. Interaction Plot ---
interaction.plot(
  x.factor = data$environment,
  trace.factor = data$teaching_method,
  response = data$scores,
  type = "b",
  pch = 19,
  col = c("blue", "red"),
  xlab = "Study Environment",
  ylab = "Mean Exam Score",
  trace.label = "Teaching Method",
  main = "Interaction Plot: Teaching Method    Study Environment"
)
```

The R code above performs all necessary numerical computations for the two-way ANOVA, including manual verification of sums of squares and the generation of the interaction plot that visually represents the relationship between the two factors.

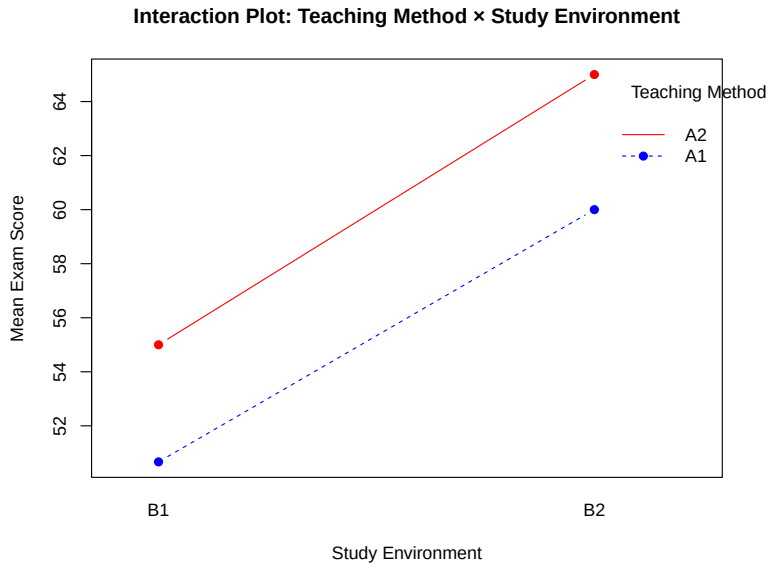

Figure 3.3: Variance decomposition of exam scores.

Results and Discussion

The interaction plot displays the mean exam scores for each combination of teaching method and study environment. For both teaching methods, the mean score increases from environment B_1 to B_2, indicating a positive effect of the study environment on students' performance.

Moreover, teaching method A_2 consistently yields higher mean scores than method A_1 in both environments, suggesting that A_2 is more effective. Since the two lines in the plot are nearly parallel, there is no strong evidence of interaction between the teaching method and the study environment.

This chapter highlighted key challenges in multiple regression, including the identification and consequences of multicollinearity and strategies for selecting suitable submodels. Understanding these issues is essential for building robust models and drawing valid conclusions from complex data.

Chapter 4

Multiple Regression Challenges: Multicollinearity and Submodel Selection

In this chapter, we extend the study of linear regression to more complex scenarios, focusing on the challenges that arise when multiple explanatory variables are included. We address issues such as multicollinearity and the selection of appropriate submodels to ensure reliable inference.

4.1 Multicollinearity

Multicollinearity arises when two or more explanatory variables are highly linearly related. In such situations, the regressors contain redundant information, making it difficult to isolate the individual effect of each variable on the dependent variable.
Multicollinearity can range from a perfect linear relationship to a near-linear dependency that distorts estimation and statistical inference.

Perfect vs. Near Multicollinearity

Perfect Multicollinearity. Perfect multicollinearity occurs when one regressor is an exact linear combination of the others:

$$X_j = \sum_{k \neq j} c_k X_k,$$

where c_k are constants.
In this situation, the matrix $X^T X$ is singular, the OLS estimator cannot be computed, and the regression parameters are not identifiable.

Example.

Suppose that one explanatory variable is a linear combination of the others:

$$X_3 = 2X_1 - 0.5X_2.$$

In this case, the coefficients of the linear combination are

$$c_1 = 2, \qquad c_2 = -0.5.$$

This means that X_3 is completely determined by X_1 and X_2, which leads to perfect multicollinearity.

Near Multicollinearity. Near multicollinearity occurs when regressors are highly, but not perfectly, linearly related:

$$X_j \approx \sum_{k \neq j} c_k X_k.$$

In this case, the OLS estimates exist but become unstable because their variances are inflated.

Example.

$\text{Corr}(\text{Income}, \text{Expenditure}) \approx 0.99$. This suggests the presence of near multicollinearity.

4.1.1 Consequences for Estimation and Inference

Estimation. The OLS estimator is given by

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X^T X)^{-1} X^T Y, \qquad \text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2 (X^T X)^{-1}.$$

Near multicollinearity implies that $X^T X$ is nearly singular. As a result:

- Small changes in the data may produce large changes in the estimated coefficients.
- The variances $\text{Var}(\hat{\beta}_j)$ become large, leading to imprecise estimates.

Inference.

- Large standard errors lead to wide confidence intervals.
- Individual t-tests may appear insignificant even when the overall F-test is significant.
- Estimated coefficients may have unexpected signs or magnitudes.

Example.

If $\text{Corr}(\text{HoursStudied}, \text{PracticeTests}) = 0.98$, both variables may jointly influence performance, but neither appears individually significant due to inflated variances.

4.1.2 Detection Tools

Correlation Matrix. High pairwise correlations (e.g. $|\rho| > 0.9$) may indicate multicollinearity, although low pairwise correlations do not necessarily rule it out.

Variance Inflation Factor (VIF).

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}.$$

Here R_j^2 is the coefficient of determination obtained by regressing X_j on the remaining regressors. A large R_j^2 indicates that X_j is well explained by other regressors and therefore has a large VIF.
Rules of thumb:

$$\mathrm{VIF} \leq 5 \text{ (acceptable)}, \qquad \mathrm{VIF} > 10 \text{ (problematic)}.$$

Condition Number. Another diagnostic measure for detecting multicollinearity is the **condition number** of the design matrix. It is based on the eigenvalues of $X^T X$:

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $X^T X$.
Interpretation:

$\kappa < 10$ (weak collinearity), $\kappa \geq 30$ (serious collinearity), $\kappa > 100$ (severe instability).

Example: Detecting Multicollinearity in House Price Data

Problem statement

We study the relationship between the house price Y and two predictors: `Size` (m^2) and `Rooms`. Use the following dataset:

Observation	Size (m^2)	Rooms
1	120	5
2	150	6
3	160	6
4	200	7
5	220	8
6	250	9

Questions

1. Compute the sample means of `Size` and `Rooms`.
2. For each observation, calculate the deviations from the means: $(X_i - \bar{X}_{\mathrm{Size}})$ and $(R_i - \bar{X}_{\mathrm{Rooms}})$.
3. Compute:
   (a) the variance of `Size`,
   (b) the variance of `Rooms`,
   (c) and the covariance between `Size` and `Rooms`.

4. Deduce the correlation coefficient between `Size` and `Rooms`.

5. Interpret the correlation value. What does it tell you about the relationship between the two predictors?

Solution

1. Compute Means.

$$\bar{X}_{\text{Size}} = 183.33, \quad \bar{X}_{\text{Rooms}} = 6.83.$$

2. Compute Deviations.

$$(X_i - \bar{X}_{\text{Size}}), \quad (R_i - \bar{X}_{\text{Rooms}}).$$

Observation	$(X_i - \bar{X}_{\text{Size}})$	$(R_i - \bar{X}_{\text{Rooms}})$
1	−63.33	−1.83
2	−33.33	−0.83
3	−23.33	−0.83
4	16.67	0.17
5	36.67	1.17
6	66.67	2.17

3. Compute Variances and Covariance.

$$\text{Var(Size)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_{\text{Size}})^2}{n-1} = \frac{11733.33}{5} = 2346.67,$$

$$\text{Var(Rooms)} = \frac{\sum_{i=1}^{n}(R_i - \bar{X}_{\text{Rooms}})^2}{n-1} = \frac{10.83}{5} = 2.17,$$

$$\text{Cov(Size, Rooms)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_{\text{Size}})(R_i - \bar{X}_{\text{Rooms}})}{n-1} = \frac{353.33}{5} = 70.67.$$

4. Compute Correlation.

$$\rho_{\text{Size, Rooms}} = \frac{\text{Cov(Size, Rooms)}}{\sqrt{\text{Var(Size)}\,\text{Var(Rooms)}}} = \frac{70.67}{\sqrt{2346.67 \times 2.17}} = 0.991.$$

5. Interpretation. The correlation between `Size` and `Rooms` is extremely high (0.991), showing strong linear dependence. This indicates near multicollinearity, which can make $(X^T X)$ nearly singular and inflate the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$.

Practical Work 10: Detecting Multicollinearity in House Price Data in R

Listing 4.1: Detecting Multicollinearity: House Price Example

```
# ============================================
# Example: Detecting Multicollinearity
# ============================================

# 1. Data
size  <- c(120, 150, 160, 200, 220, 250)
rooms <- c(5, 6, 6, 7, 8, 9)
price <- c(220, 245, 260, 310, 340, 360)  # Optional outcome

data <- data.frame(size, rooms, price)
print(data)

# 2. Compute means
mean_size  <- mean(size)
mean_rooms <- mean(rooms)
mean_size
mean_rooms

# 3. Compute deviations
dev_size  <- size  - mean_size
dev_rooms <- rooms - mean_rooms

# 4. Compute variances and covariance (sample version)
var_size  <- var(size)
var_rooms <- var(rooms)
cov_sr    <- cov(size, rooms)

var_size
var_rooms
cov_sr

# 5. Correlation between Size and Rooms
cor_sr <- cor(size, rooms)
cor_sr  # Expected     0.991

# 6. Cross-product matrix X'X
X <- cbind(1, size, rooms)
t(X) %*% X

# 7. Regression to observe instability
model_full <- lm(price ~ size + rooms, data = data)
summary(model_full)

# 8. Compare with reduced model (no Rooms)
model_reduced <- lm(price ~ size, data = data)
summary(model_reduced)

# 9. Visualize the correlation
plot(size, rooms, pch = 19, col = "steelblue",
     main = "Strong Correlation between Size and Rooms",
```

```r
     xlab = "Size (m )", ylab = "Rooms")
abline(lm(rooms ~ size), col = "red", lwd = 2)

# ============================================
# 10. Compute Variance Inflation Factor (VIF)
# ============================================
# Install car package if not already installed
# install.packages("car")
library(car)
vif(model_full)  # High VIF confirms near multicollinearity
```

The following plot confirms a very high correlation between `Size` and `Rooms`:

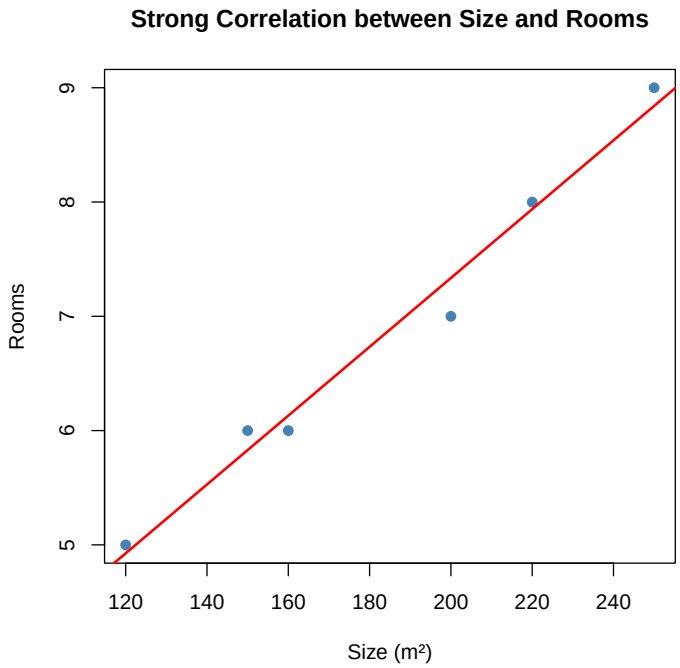

Figure 4.1: Scatter plot of `Size` versus `Rooms`.

Results and Discussion

The results reveal a very high correlation between `Size` and `Rooms`, with $r = 0.991$. This indicates a strong linear dependence between these two predictors: larger houses almost always have more rooms. Such a near-perfect relationship implies the presence of **strong multicollinearity**.

As a consequence, the matrix $X^T X$ becomes nearly singular, leading to unstable and unreliable estimates of the regression coefficients for `Size` and `Rooms`. This is also confirmed numerically by the Variance Inflation Factor (VIF), which is very high for both predictors (VIF ≈ 56), indicating extreme multicollinearity.

In practice, including both `Size` and `Rooms` in the regression model may result in large standard errors and insignificant t-tests for individual coefficients, even though the overall model explains the dependent variable well.

4.2 Remedies to Multicollinearity

4.2.1 Variable Selection, Transformation, and Variance Stabilization

One practical approach to multicollinearity is to **remove or combine redundant predictors**, thereby simplifying the model and reducing dependence among regressors. For example, highly correlated variables such as `Size` and `Rooms` can be replaced by a single composite variable (e.g., `Living_Area`) that captures the dominant information about house characteristics.
Centering and scaling predictors improve numerical conditioning and interpretability, but they do not eliminate structural multicollinearity between regressors.
In some situations, appropriate transformations of variables (logarithmic, polynomial, or interaction terms) may help reduce dependence among predictors or improve model interpretability, although such transformations may also introduce additional correlations and should therefore be used with care.

4.2.2 Principal Component Regression (PCR)

PCR uses the principal components of the predictor matrix X instead of the original correlated predictors. The first few components explain most of the variance in the predictors, while discarding components associated with near-zero eigenvalues alleviates collinearity. This approach also **stabilizes the variance** of coefficient estimates by focusing on the main directions of variability, reducing noise from redundant or highly correlated variables. PCR is particularly useful when the number of predictors is large relative to the number of observations.

4.2.3 Ridge Regression

Ridge regression **modifies the least squares criterion by adding an L_2 penalty**:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2 \right\}.$$

Where $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage.
This approach **shrinks coefficients**, reduces the impact of multicollinearity, and stabilizes the variance of the estimates when predictors are highly correlated. Ridge regression is especially effective when all predictors are potentially relevant, but high correlations make standard OLS estimates unstable. By tuning the penalty parameter λ, one can find an optimal balance between bias and variance.

4.2.4 Lasso Regression

Lasso regression introduces an L_1 **penalty**:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\},$$

leading to both **shrinkage and automatic variable selection**. By forcing some coefficients exactly to zero, Lasso selects a subset of predictors, which not only simplifies the model but also indirectly stabilizes the variance of the remaining coefficient estimates.

Lasso is particularly valuable in high-dimensional settings where the number of predictors exceeds the number of observations, or where interpretability is critical.

Example: Linear Dependence Between Regressors

Problem statement

Suppose we have a dataset with 5 observations ($n = 5$) and 3 explanatory variables: X_1 and X_2 are strongly correlated, and Y is the response variable. The data are as follows:

Observation	X_1	X_2	Y
1	1	2	3
2	2	4	5
3	3	6	7
4	4	8	9
5	5	10	11

The variable X_2 is simply double X_1, i.e. $X_2 = 2 \cdot X_1$, which introduces **perfect multicollinearity** in the data. We then set up a multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

The goal is to estimate the coefficients $\beta_0, \beta_1, \beta_2$ of the model, but because of multicollinearity, the matrix $X^T X$ will be **non-invertible**.

Solution: Non-Invertible Matrix

The design matrix X is:

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \\ 1 & 4 & 8 \\ 1 & 5 & 10 \end{pmatrix}$$

The matrix $X^T X$ is:

$$X^T X = \begin{pmatrix} 5 & 15 & 30 \\ 15 & 55 & 110 \\ 30 & 110 & 220 \end{pmatrix}$$

The determinant of $X^T X$ is:

$$\det(X^T X) = 5\,(55 \times 220 - 110 \times 110) - 15\,(15 \times 220 - 30 \times 110) + 30\,(15 \times 110 - 30 \times 55)$$

$$\det(X^T X) = 0$$

This shows that $X^T X$ is **singular** and **non-invertible**. Thus:

- Perfect multicollinearity makes the OLS estimator undefined because the matrix $X^T X$ is singular.

- Ridge and Lasso are robust alternatives.

This R code illustrates how a multicollinearity problem makes the matrix X^TX non-invertible, thus preventing the classical estimation of regression coefficients. It also shows how Ridge and Lasso regularization methods can estimate the coefficients even in this context.

Practical Work 11: Linear Dependence Between Regressors in R

Listing 4.2: R code illustrating multicollinearity and Ridge/Lasso regression.

```r
# Load necessary library
library(glmnet)     # For Ridge (L2) and Lasso (L1) regressions

# Simulate data
X1 <- c(1, 2, 3, 4, 5)    # Explanatory variable X1
X2 <- 2 * X1              # X2 is perfectly collinear with X1
Y  <- c(3, 5, 7, 9, 11)   # Response variable Y

# Create a data frame
data <- data.frame(Y, X1, X2)

# Construct the design matrix (without intercept)
X <- as.matrix(data[, c("X1", "X2")])

# Add intercept column to X
X_with_intercept <- cbind(1, X)

# Compute X'X (cross-product matrix)
XtX <- t(X_with_intercept) %*% X_with_intercept
print("Matrix X'X :")
print(XtX)

# Determinant of X'X (if zero, not invertible)
det_XtX <- det(XtX)
print(paste("Determinant of X'X:", det_XtX))

# Classical OLS estimation
lm_model <- lm(Y ~ X1 + X2, data = data)
summary(lm_model)  # May fail or produce unstable coefficients

# Ridge Regression (L2 penalty)
ridge_model <- glmnet(X, Y, alpha = 0, lambda = 1)
print("Ridge coefficients:")
print(coef(ridge_model))

# Lasso Regression (L1 penalty)
lasso_model <- glmnet(X, Y, alpha = 1, lambda = 1)
```

```
print("Lasso coefficients:")
print(coef(lasso_model))
```

- `glmnet` fits a penalized model.
- `alpha = 0` specifies an L_2 penalty → **Ridge regression**.
- `alpha = 1` specifies an L_1 penalty → **Lasso regression**.
- `lambda = 1` sets the penalty strength.
- `coef()` displays the estimated coefficients.

Results and Discussion

The results confirm that the design matrix is singular due to perfect collinearity between X_1 and X_2 ($X_2 = 2X_1$). The determinant of X^TX equals zero, implying that the matrix is non-invertible and that the ordinary least squares estimator cannot be uniquely computed because $(X^TX)^{-1}$ does not exist. Consequently, the classical `lm()` function in R detects a singularity and cannot estimate both regression coefficients simultaneously.

Ridge and Lasso regression overcome this difficulty by introducing regularization terms. Ridge regression stabilizes the estimates by shrinking the coefficients toward zero, while Lasso performs both shrinkage and automatic variable selection by forcing some coefficients to zero.

4.3 Singular Case, Identifiability Constraints, and Estimable Functions

Identifiability Constraints in Regression Models

In linear regression, parameters are not identifiable when the columns of the design matrix X are linearly dependent. A classical example is when X contains dummy variables for all levels of a categorical factor plus an intercept. For instance, if X has dummies for groups A and B along with the intercept, we have:

$$\text{Dummy}_A + \text{Dummy}_B = \mathbf{1},$$

which introduces **perfect collinearity**.

Analytical Treatment. In such cases, the matrix X^TX is singular ($\det(X^TX) = 0$) and the ordinary least squares estimator:

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$$

is **not defined**. To achieve identifiability, we must impose a constraint, for example:

$$\beta_A + \beta_B = 0 \quad \text{or} \quad \beta_B = 0.$$

Even when individual parameters are not identifiable, certain linear combinations (contrasts) remain **estimable**.

Estimable Functions and Their Interpretation

A linear combination $c^T\beta$ is **estimable** if there exists a vector a such that:

$$c^T\beta = a^T X\beta.$$

Equivalently, c must belong to the **row space of** X. In practice, contrasts such as $\beta_A - \beta_B$ are often estimable even when the full set of parameters is not.

Example: Department Salary Model

Problem Statement

Consider modeling salaries by department to understand differences among departments A, B, and C:

$$Y_i = \beta_0 + \beta_1 D_{A_i} + \beta_2 D_{B_i} + \beta_3 D_{C_i} + \varepsilon_i,$$

where $D_{A_i}, D_{B_i}, D_{C_i}$ are dummy variables for each department. Including an intercept together with all three dummies makes the design matrix **singular**, since

$$D_{A_i} + D_{B_i} + D_{C_i} = 1.$$

Questions

1. Why is the model not identifiable if all dummies and an intercept are included?
2. What is the rank of the design matrix in this case?
3. Suggest a way to make the model identifiable.
4. Which contrasts between departments remain estimable?

Solution

1. The model is not identifiable because the three dummy variables and the intercept are linearly dependent:

   $$D_A + D_B + D_C = 1.$$

   One column of the design matrix can be expressed as a linear combination of the others, making $X^T X$ singular and $(X^T X)^{-1}$ undefined.

2. The rank of the design matrix is 3 (number of dummies) + 1 (intercept) -1 (redundancy) $= 3$ independent columns.

3. To make the model identifiable, remove one dummy variable (choose a **reference category**). For example, take Department A as reference and keep D_B and D_C:

   $$Y_i = \beta_0 + \beta_1 D_{B_i} + \beta_2 D_{C_i} + \varepsilon_i.$$

   Here, β_0 represents the mean salary in Department A, while β_1 and β_2 represent deviations from that baseline.

4. The estimable contrasts are:

   $$\text{B vs A: } \beta_1, \quad \text{C vs A: } \beta_2, \quad \text{C vs B: } \beta_2 - \beta_1.$$

   Even though β_A is not explicitly included, all meaningful differences between departments can be estimated.

Practical Work 12: Department Salary Model in R

The salary values are chosen to reflect realistic differences between departments, ensuring that Department A has the lowest average salary, Department B slightly higher, and Department C the highest. This makes the contrasts between departments easy to interpret in the regression model.

Listing 4.3: Checking model identifiability and estimable contrasts in the Department Salary Model.

```
1
2  # Create a small example dataset
3  salary_data <- data.frame(
4    Salary = c(3000, 3200, 2800, 3500, 3700, 4000),
5    Dept   = factor(c("A", "A", "B", "B", "C", "C"))
6  )
7
8  # Display design matrix when including all dummies + intercept
9  X_full <- model.matrix(~ Dept, data = salary_data)
10 t(X_full) %*% X_full   # Check for singularity
11 det(t(X_full) %*% X_full)  # Determinant should be 0 (singular)
12
13 # Fit model automatically handled by R (drops one dummy)
14 lm_salary <- lm(Salary ~ Dept, data = salary_data)
15 summary(lm_salary)
16
17 # Check which contrasts (differences) are estimable
18 library(emmeans)
19 emmeans(lm_salary, pairwise ~ Dept)
```

Results and Discussion

R automatically excludes one dummy variable (Department A) to ensure model identifiability. The estimated regression equation is:

$$\widehat{Salary}_i = 3100 + 50D_{B_i} + 750D_{C_i}.$$

This implies that employees in Department B earn, on average, \$50 more than those in Department A, while Department C employees earn approximately \$750 more. The contrast between Departments C and B ($\beta_2 - \beta_1 = 700$) indicates a substantial salary gap. These results illustrate how dropping one dummy variable restores identifiability without losing interpretability, as all meaningful contrasts between departments remain estimable.

4.4 Submodel Selection in Multiple Regression

Theoretical Background for Submodel Selection

Before applying model selection techniques, it is essential to recall the main theoretical quantities used to compare submodels in multiple regression.

Model Setup

As introduced in Chapter 3, for a response variable Y and $(p-1)$ predictors collected in matrix X, the linear model (3.3) is:

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

where β contains the regression coefficients and σ^2 is the error variance.
The least squares estimator is:

$$\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T Y,$$

and the fitted values and residuals are:

$$\widehat{Y} = X\widehat{\beta}, \qquad e = Y - \widehat{Y}.$$

The Residual Sum of Squares (RSS) measures the unexplained variation:

$$\text{RSS} = e^T e = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

Goodness of Fit Measures

The Total Sum of Squares (SST) quantifies the total variability in Y:

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

and the Coefficient of Determination is

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}.$$

Because R^2 always increases when adding predictors, the **adjusted coefficient of determination** is preferred:

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p},$$

where p is the total number of estimated parameters (including the intercept).
This adjustment penalizes adding predictors that do not improve the model significantly.

Information Criteria

Information criteria balance model fit and complexity:

$$\text{AIC} = n \ln\left(\frac{\text{RSS}}{n}\right) + 2p, \qquad \text{BIC} = n \ln\left(\frac{\text{RSS}}{n}\right) + p \ln n.$$

- Smaller AIC or BIC indicates a better submodel.
- BIC penalizes the number of parameters more heavily than AIC, favoring simpler models when n is moderate.

Mallows' C_p Criterion

Mallows' C_p evaluates each submodel S relative to the full model:

$$C_p(S) = \frac{\text{RSS}_S}{\widehat{\sigma}^2_{\text{full}}} - (n - 2p_S),$$

where

$$\widehat{\sigma}^2_{\text{full}} = \frac{\text{RSS}_{\text{full}}}{n - p_{\text{full}}}$$

is the unbiased estimate of the error variance from the full model, and p_S is the number of parameters in the submodel (including the intercept).
A good submodel satisfies:

$$C_p(S) \approx p_S,$$

indicating a balance between bias (underfitting) and variance (overfitting).

Decision Principles for Submodel Selection

- Prefer the model with the largest **adjusted** R^2, as it accounts for model complexity.
- Prefer the model with the smallest **AIC or BIC**.
- Prefer submodels where $C_p \approx p_S$, showing that residual variance is consistent with the number of parameters.
- When several models perform similarly, choose the simpler (more **parsimonious**) model to enhance interpretability and reduce overfitting risk.

These theoretical principles are now applied to a concrete dataset in the following example.

Example: Evaluating Competing Regression Submodels

We observe $n = 8$ units with one response variable Y and three candidate predictors X_1, X_2, X_3. The data are:

i	Y_i	X_{1i}	X_{2i}	X_{3i}
1	5	1	2	5
2	7	2	1	3
3	9	3	4	6
4	10	4	3	4
5	12	5	5	7
6	13	6	4	5
7	15	7	6	8
8	16	8	5	6

We consider linear models with intercept:

$$Y = \beta_0 + \sum_{j \in S} \beta_j X_j + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

for different subsets $S \subseteq \{1, 2, 3\}$.

Questions

1. For each candidate submodel S (null, single predictor, pairs, full model), compute the Residual Sum of Squares

$$\mathrm{RSS}_S = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i,S})^2,$$

the adjusted coefficient of determination adj R^2, AIC, and BIC.

2. Using the full model residuals, compute Mallows' C_p for each submodel:

$$C_p(S) = \frac{\mathrm{RSS}_S}{\hat{\sigma}^2_{\text{full}}} - (n - 2p_S),$$

where p_S is the number of parameters (including intercept), and $\hat{\sigma}^2_{\text{full}} = \mathrm{RSS}_{\text{full}}/(n - p_{\text{full}})$.

3. Based on the criteria (adjusted R^2, AIC, BIC, C_p), which submodel would you select? Provide reasoning.

4. Give a short interpretation of the chosen model.

Solution

1. Compute OLS fits and RSS_S

All submodels were fitted (intercept always included). The summary statistics are:

Model S	p_S	RSS_S	R^2_{adj}	AIC	BIC
Null (no X)	1	102.8750	0.0000	22.433	22.512
$\{X_1\}$	2	0.7262	0.9918	−15.195	−15.036
$\{X_2\}$	2	25.8718	0.7066	13.390	13.549
$\{X_3\}$	2	63.8611	0.2758	20.618	20.777
$\{X_1, X_2\}$	3	0.4244	0.9942	−17.492	−17.254
$\{X_1, X_3\}$	3	0.5250	0.9929	−15.790	−15.552
$\{X_2, X_3\}$	3	9.1935	0.8749	7.112	7.351
$\{X_1, X_2, X_3\}$	4	0.3750	0.9936	−16.482	−16.164

Formulas used:

$$\mathrm{RSS}_S = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i,S})^2, \quad R^2 = 1 - \frac{\mathrm{RSS}_S}{\mathrm{SST}}, \quad R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p_S},$$

$$\mathrm{AIC} = n \ln\left(\frac{\mathrm{RSS}_S}{n}\right) + 2p_S, \quad \mathrm{BIC} = n \ln\left(\frac{\mathrm{RSS}_S}{n}\right) + p_S \ln n.$$

2. Compute $\hat{\sigma}^2_{\text{full}}$ and Mallows' C_p

From the full model $\{X_1, X_2, X_3\}$:

$$\text{RSS}_{\text{full}} = 0.3750, \quad p_{\text{full}} = 4 \quad \text{(intercept + 3 predictors)},$$

$$\hat{\sigma}^2_{\text{full}} = \frac{\text{RSS}_{\text{full}}}{n - p_{\text{full}}} = \frac{0.3750}{8 - 4} = 0.09375.$$

Mallows' C_p for each model:

$$C_p(S) = \frac{\text{RSS}_S}{0.09375} - (8 - 2p_S)$$

Model S	$C_p(S)$
Null	1091.333
$\{X_1\}$	3.746
$\{X_2\}$	271.966
$\{X_3\}$	677.185
$\{X_1, X_2\}$	2.527
$\{X_1, X_3\}$	3.600
$\{X_2, X_3\}$	96.065
$\{X_1, X_2, X_3\}$	4.000

Recall: a good model has $C_p \approx p_S$.

3. Model comparison and selection

- **Adjusted** R^2: highest for $\{X_1, X_2\}$ (0.9942), then full model (0.9936), then $\{X_1, X_3\}$ (0.9929).
- **AIC**: lowest (best) for $\{X_1, X_2\}$ (-17.492), then full model, then $\{X_1\}$.
- **BIC**: lowest (best) for $\{X_1, X_2\}$ (-17.254), then full model.
- **Mallows'** C_p: $\{X_1, X_2\}$ yields $C_p \approx 2.53$ (close to $p = 3$); full model gives $C_p = 4$ with $p = 4$. Single predictor $\{X_1\}$ has $C_p \approx 3.75$ with $p = 2$ (less close).

All criteria indicate models containing X_1 are far superior. Among these, $\{X_1, X_2\}$ is slightly favored due to lowest AIC/BIC and C_p close to p. The full model adds one parameter for marginal improvement (RSS drops $0.4244 \to 0.3750$). Since BIC penalizes complexity more strongly, $\{X_1, X_2\}$ is the best compromise.

Selected model

$$\boxed{S = \{X_1, X_2\}}$$

Rationale: lowest AIC/BIC, C_p close to p, very high adjusted R^2. Adding X_3 provides marginal gain relative to complexity penalty.

4. Interpretation of the chosen model For the model $\{X_1, X_2\}$ (intercept + X_1, X_2):

- β_0: estimated mean of Y when $X_1 = X_2 = 0$ (if meaningful).
- β_1: expected change in Y for one unit increase in X_1, holding X_2 constant.
- β_2: expected change in Y for one unit increase in X_2, holding X_1 constant.

Because X_1 explains most variability (low RSS) and X_2 adds small but meaningful contribution, the two-variable model is parsimonious and interpretable.

Conclusion. This example demonstrates how computing RSS, adjusted R^2, AIC/BIC, and Mallows' C_p for each candidate submodel allows an evidence-based selection of a parsimonious model. In practice, residual diagnostics, multicollinearity checks (e.g., VIF), and assumption verification should also be performed before finalizing the model.

Practical Work 13: Evaluating Competing Regression Submodels in R

Listing 4.4: Submodel Selection in Multiple Regression.

```r
# ============================================================
# Submodel Selection in Multiple Regression
# ============================================================
# 1. Create the exact dataset
Y  <- c(5, 7, 9, 10, 12, 13, 15, 16)
X1 <- c(1, 2, 3, 4, 5, 6, 7, 8)
X2 <- c(2, 1, 4, 3, 5, 4, 6, 5)
X3 <- c(5, 3, 6, 4, 7, 5, 8, 6)
data <- data.frame(Y, X1, X2, X3)
# 2. Fit all possible submodels (intercept always included)
null_model <- lm(Y ~ 1, data=data)
m1 <- lm(Y ~ X1, data=data)
m2 <- lm(Y ~ X2, data=data)
m3 <- lm(Y ~ X3, data=data)
m12 <- lm(Y ~ X1 + X2, data=data)
m13 <- lm(Y ~ X1 + X3, data=data)
m23 <- lm(Y ~ X2 + X3, data=data)
full_model <- lm(Y ~ X1 + X2 + X3, data=data)
# 3. Compute RSS, adjusted R  , AIC, BIC
models <- list(null_model, m1, m2, m3, m12, m13, m23, full_model)
model_names <- c("Null", "X1", "X2", "X3", "X1+X2", "X1+X3", "X2+
   X3", "Full")

compute_metrics <- function(mod) {
  RSS <- sum(residuals(mod)^2)
  adjR2 <- summary(mod)$adj.r.squared
  n <- length(residuals(mod))
  p <- length(coef(mod))
  AIC_val <- n*log(RSS/n) + 2*p
```

```r
  BIC_val <- n*log(RSS/n) + p*log(n)
  return(c(RSS=RSS, adjR2=adjR2, AIC=AIC_val, BIC=BIC_val))
}

metrics <- t(sapply(models, compute_metrics))
rownames(metrics) <- model_names
print(metrics)
# 4. Compute Mallows' Cp
RSS_full <- sum(residuals(full_model)^2)
p_full <- length(coef(full_model))
sigma2_full <- RSS_full / (nrow(data) - p_full)

Cp <- sapply(models, function(mod) {
  pS <- length(coef(mod))
  RSS_S <- sum(residuals(mod)^2)
  RSS_S / sigma2_full - (nrow(data) - 2*pS)
})
names(Cp) <- model_names
print(Cp)
# 5. Optional: visualize adjusted R
library(ggplot2)
df_plot <- data.frame(Model=model_names, Adj_R2=metrics[,"adjR2"
  ])
ggplot(df_plot, aes(x=Model, y=Adj_R2)) +
  geom_col(fill="steelblue") +
  geom_text(aes(label=round(Adj_R2,3)), vjust=-0.5) +
  labs(title="Adjusted R  Comparison of Submodels",
       x="Model", y="Adjusted R ") +
  theme_minimal()
```

The following figure summarizes the comparison of Adjusted R^2 values for the different submodels.

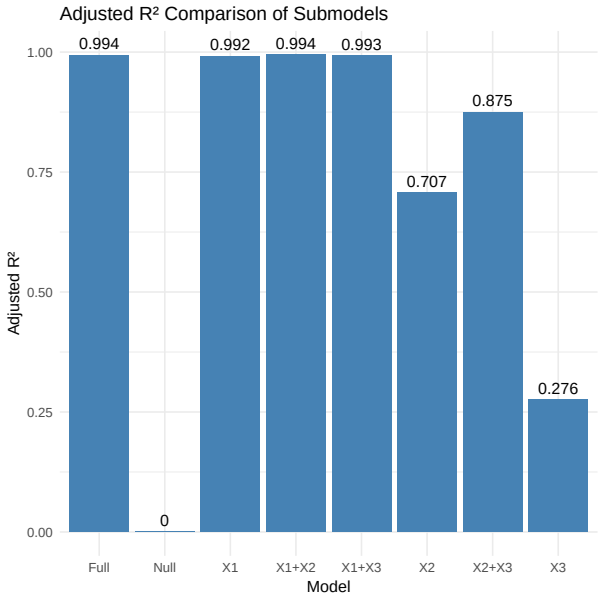

Figure 4.2: Comparison of Adjusted R^2 for different submodels of exam score regression.

Results and Discussion

Figure 4.2 presents a comparison of the adjusted R^2 values for all considered submodels in the multiple regression analysis. Adjusted R^2 accounts for the number of predictors in each model, providing a corrected measure of model fit that penalizes unnecessary complexity.
From the plot, several observations can be made:

- The submodel including both $X1$ and $X2$ achieves the highest adjusted R^2, indicating that these two predictors together explain the largest proportion of variability in Y while accounting for model complexity. This confirms the theoretical expectation that $X1$ and $X2$ are the most informative variables for predicting Y.
- Models including only $X1$ or only $X2$ have slightly lower adjusted R^2 values. While $X1$ alone explains a substantial amount of variation, adding $X2$ further improves the fit, highlighting the complementary contribution of these two predictors.
- The inclusion of $X3$, either alone or in combination with other variables, yields relatively lower adjusted R^2 values. This suggests that $X3$ contributes minimally to explaining the response variable, consistent with its lower theoretical relevance in the model.
- The null model, which includes only the intercept, unsurprisingly has an adjusted R^2 of zero, confirming that none of the variability in Y is explained without predictors.

The study of multicollinearity and submodel selection provides a foundation for assessing model adequacy. The last chapter examines regression assumptions, diagnostics, and remedies to improve model reliability.

Chapter 5

Diagnostics, Remedies, and Extensions of the Linear Model

Building on the previous discussion of multiple regression and submodel selection, this chapter focuses on assessing the validity of linear models. We will investigate how violations of classical assumptions affect inference, present diagnostic techniques for detecting such issues, and introduce methods to correct or extend the basic linear model framework.

5.1 Diagnostics of Assumptions, Remedies, and Advanced Applications

5.1.1 Consequences of Assumption Violations

In Chapter 1, we introduced the classical linear regression assumptions. When these assumptions fail, the consequences include:

- **Endogeneity (exogeneity violation):** estimators become biased and inconsistent when regressors correlate with the error term. This may result from omitted variables, measurement errors, or simultaneity.
- **Heteroscedasticity and multicollinearity:** estimators remain unbiased under OLS but are inefficient; standard errors are incorrect, which compromises tests and confidence intervals. Consequently, t- and F-tests may provide misleading results.
- **Autocorrelation and non-normality:** inference based on usual t- and F-statistics can be invalid; prediction intervals may be misleading.

5.1.2 Diagnostics of Model Assumptions

Having outlined the possible consequences of assumption violations, we now turn to the methods used to detect such problems in practice.

Residual analysis

Residuals are defined by

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \equiv e_i, \qquad i = 1, \ldots, n,$$

and are the primary tool for diagnosing violations. Useful graphical diagnostics include:

- **Residuals vs. Fitted:** detect nonlinearity and heteroscedasticity; a fan-shaped pattern suggests increasing variance with fitted values.
- **Scale–Location (spread) plot:** visualize variance changes with fitted values; an upward trend indicates heteroscedasticity.
- **Normal Q–Q plot:** assess residual normality; points should lie approximately on the diagonal line.
- **Histogram of residuals:** complements Q–Q plot for detecting skewness and kurtosis.

Influence and leverage

Large-leverage or influential observations can distort estimates:

- **Hat matrix:** $H = X(X^T X)^{-1} X^T$, leverage $h_{ii} = H_{ii}$ measures how far observation i's predictors are from the center of the data.
- **Cook's distance:**
$$D_i = \frac{e_i^2}{p\,\widehat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2},$$
quantifies the influence of observation i on fitted values. A commonly used rule-of-thumb is
$$D_i > \frac{4}{n},$$
but depending on the sample size, some authors suggest
$$D_i > 1 \quad \text{or} \quad D_i > \frac{3}{n}.$$
Values above these thresholds suggest strong influence on the estimated coefficients.
- **DFBETAS, DFFITS:** give parameter- and prediction-specific influence measures; for DFBETAS, a common guideline is $|DFBETA_i| > 2/\sqrt{n}$.

Formal tests

When a graphical diagnostic suggests a violation, formal tests help:

- **Breusch–Pagan test:** regress e_i^2 on regressors; test statistic nR^2_{aux}. Assumes residuals are approximately normal.
- **White's test:** more general test allowing nonlinear and interaction terms of regressors.
- **Durbin–Watson / Ljung–Box:** test autocorrelation in residuals. Durbin–Watson is mainly for first-order autocorrelation; Ljung–Box is preferable for longer series.
- **Variance Inflation Factor (VIF):** detect multicollinearity; VIF values above 10 often indicate severe multicollinearity.

5.1.3 Model Remedies, Transformations, and Variance Stabilization

Motivation

When diagnostic analyses indicate violations of the classical regression assumption such as heteroscedasticity, nonlinearity, or the presence of influential points: model adequacy can often be restored through suitable remedies or transformations. The goal is to stabilize variance, improve linearity, and reduce the influence of outliers while maintaining interpretability.

Transformations of Variables

Transforming the dependent or explanatory variables is often an effective way to achieve homoscedasticity and approximate normality of residuals. Common transformations include:

- **Logarithmic transformation:** useful when variability increases proportionally with the mean or when relationships are multiplicative.
- **Square-root transformation:** appropriate for count-type data where variance is roughly proportional to the mean.
- **Reciprocal transformation:** helps linearize inverse relationships.
- **Box–Cox family:** provides a continuum of power transformations,

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log(Y), & \lambda = 0, \end{cases}$$

  where λ is estimated by maximizing the model's log-likelihood. Note that Y must be strictly positive.

The interpretation of regression coefficients changes with the transformation. For example, in a log–log model, coefficients represent elasticities; in a log–linear model, they correspond to semi-elasticities. Thus, all inferences must be expressed on the transformed scale or appropriately back-transformed for reporting.

Weighted and Robust Approaches

If transformation alone does not adequately address heteroscedasticity or outliers, several alternative estimation strategies can be used:

- **Weighted Least Squares (WLS):** assigns weights inversely proportional to the estimated error variance.
- **Robust regression (M-estimators):** downweights extreme residuals to limit outlier influence.
- **Heteroscedasticity-consistent standard errors (HC3):** adjust inference while retaining OLS point estimates.

Generalized and Mixed Model Extensions

When assumption violations cannot be adequately corrected, more general frameworks extend the linear model's flexibility:

- **Generalized Least Squares (GLS):** explicitly models correlated or non-constant variance structures.
- **Generalized Linear Models (GLMs):** handle non-normal response distributions through appropriate link functions.
- **Mixed and hierarchical models:** incorporate random effects to account for grouped or nested data.

5.1.4 Advanced ANCOVA Applications

Model and assumptions

ANCOVA combines categorical factors and continuous covariates to compare adjusted group means. A basic ANCOVA model is

$$Y_i = \beta_0 + \beta_1 X_i + \gamma Z_i + \varepsilon_i,$$

where Z_i encodes a categorical factor (e.g. by dummies), X_i is a quantitative covariate, and $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. A key assumption is *homogeneity of regression slopes*: the effect of X on Y is the same across factor levels.

Interaction (test of homogeneity of slopes)

Test homogeneity by adding an interaction:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma Z_i + \delta(X_i \times Z_i) + \varepsilon_i.$$

Testing $H_0 : \delta = 0$ is equivalent to testing parallelism of group regression lines. If H_0 is not rejected, report adjusted group means from the reduced model; otherwise present group-specific slopes.

Example: Parallel Regression Lines Model with a Two-Level Factor

Problem Statement

We consider a dataset of $n = 10$ observations containing a continuous response Y, a quantitative covariate X, and a two-level factor $Z \in \{A, B\}$:

i	X_i	Z_i	Y_i
1	1	A	3.7
2	2	A	4.6
3	3	A	6.1
4	4	A	6.9
5	5	A	8.2
6	1	B	4.1
7	2	B	6.0
8	3	B	6.8
9	4	B	8.0
10	5	B	14.2 (possible outlier)

We adopt the ANCOVA model:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma D_i + \varepsilon_i, \qquad i = 1, \ldots, 10,$$

where $D_i = 1(Z_i = B)$ is the dummy variable for group B.

Questions

1. Fit the OLS model and estimate the parameters $(\beta_0, \beta_1, \gamma)$.
2. Evaluate the model fit using residuals, RSS, and estimate the error variance.
3. Test for heteroscedasticity using the Breusch–Pagan test.
4. Identify influential observations using Cook's distance and leverage.
5. Apply Weighted Least Squares (WLS) to correct for groupwise heteroscedasticity, if necessary.
6. Test the homogeneity of slopes by adding the interaction term $X \times Z$.

Solution

1. OLS estimation

The normal equations give:

$$\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T Y = \begin{pmatrix} 0.8750 \\ 1.6750 \\ 1.9200 \end{pmatrix}.$$

Hence, the fitted model is:

$$\widehat{Y}_i = 0.8750 + 1.6750 X_i + 1.9200 D_i.$$

2. Residuals and error variance

The residual sum of squares (RSS) and the estimated variance are:

$$\text{RSS} = 15.6755, \qquad \widehat{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{15.6755}{7} \approx 2.239.$$

Residuals are mostly small except for the last observation, suggesting one potential outlier.

3. Breusch–Pagan test

Auxiliary regression of squared residuals gives:

$$R^2_{\text{aux}} = 0.4272, \qquad \text{BP} = nR^2_{\text{aux}} = 4.27, \quad p \approx 0.118.$$

Conclusion: no strong evidence against homoscedasticity at the 5% level, but possible mild heteroscedasticity.

4. Influence diagnostics

The leverages and Cook's distances are:

$$h = (0.40, 0.25, 0.20, 0.25, 0.40, 0.40, 0.25, 0.20, 0.25, 0.40),$$

$$D = (0.219, 0.009, 0.002, 0.030, 0.182, 0.023, 0.001, 0.048, 0.148, 1.518).$$

Since $4/n = 0.4$, observation 10 ($D_{10} = 1.52$) is clearly influential and should be examined carefully.

5. Weighted Least Squares (WLS)

Although the Breusch–Pagan test is not significant at the 5% level, WLS is illustrated in order to show how groupwise heteroscedasticity could be handled.
Estimated groupwise residual variances:

$$\widehat{v}_A = 0.6123, \quad \widehat{v}_B = 2.5229.$$

Using inverse-variance weights yields:

$$\widehat{\beta}_{\text{WLS}} = \begin{pmatrix} 1.8714 \\ 1.3429 \\ 1.9200 \end{pmatrix}.$$

Interpretation: group B (higher variance) receives less weight, slightly changing intercept and slope.

6. ANCOVA: test for interaction

Fit $Y = \beta_0 + \beta_1 X + \gamma D + \delta(X \times D) + \varepsilon$.
Comparing reduced (no interaction) and full models:

$$F = \frac{(15.6755 - 9.7350)/1}{9.7350/6} = 3.6613, \qquad p = 0.1042.$$

Conclusion: homogeneity of slopes not rejected at 5% (marginal at 10%).

Practical Work 14: Parallel Regression Lines Model with a Two-Level Factor in R

Listing 5.1: Diagnostics and ANCOVA example.

```r
# Data
X <- c(1,2,3,4,5,1,2,3,4,5)
Z <- factor(c(rep("A",5), rep("B",5)))
Y <- c(3.7,4.6,6.1,6.9,8.2,4.1,6.0,6.8,8.0,14.2)
data <- data.frame(Y, X, Z)
# 1. Baseline OLS
model_ols <- lm(Y ~ X + Z, data = data)
summary(model_ols)
# 2. Residual plots
plot(model_ols$fitted.values, resid(model_ols),
     xlab="Fitted values", ylab="Residuals",
     main="Residuals vs Fitted")
abline(h=0, lty=2)
qqnorm(resid(model_ols)); qqline(resid(model_ols))
# 3. Breusch-Pagan test
library(lmtest)
bptest(model_ols)
# 4. Cook's distance and leverage
data.frame(obs=1:10, Y, X, Z,
           resid=resid(model_ols),
           hat=round(hatvalues(model_ols),3),
           cooks=round(cooks.distance(model_ols),3))
# 5. Weighted Least Squares (if needed)
by_resid_var <- tapply(resid(model_ols)^2, Z, mean)
weights <- ifelse(Z=="A", 1/by_resid_var["A"], 1/by_resid_var["B"
   ])
model_wls <- lm(Y ~ X + Z, data = data, weights = weights)
summary(model_wls)
# 6. ANCOVA: test interaction
model_int <- lm(Y ~ X * Z, data = data)
anova(model_ols, model_int)
```

The plot generated above shows the Normal Q–Q plot of residuals from OLS regression.

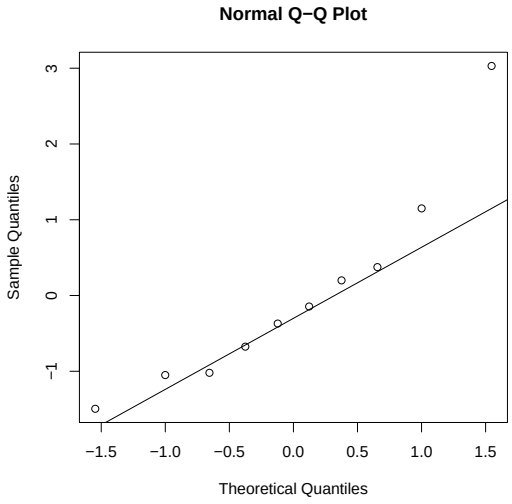

Figure 5.1: Normal Q–Q plot of residuals for the OLS regression model.

Results and Discussion

Figure 5.1 presents the Normal Q–Q plot of the residuals from the OLS regression model. Most of the residuals lie close to the theoretical straight line, indicating that the normality assumption is reasonably satisfied. However, slight deviations are observed at the upper tail, suggesting the presence of one or two influential observations with larger residuals (notably observation 10, as indicated by the Cook's distance analysis). Despite these minor departures, the overall distribution of residuals appears approximately normal, supporting the validity of the inferential results obtained from the OLS estimation.

5.2 Hierarchical and Mixed Factor Models

Mathematical Formulation

Hierarchical ANOVA is used when experimental units are **nested** within higher-level groups. For example:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2),$$

where:

- μ is the overall mean,
- α_i is the effect of factor A (the higher-level factor),
- $\beta_{j(i)}$ is the effect of factor B nested within A,
- ε_{ijk} is the random error term.

Example. Suppose we measure students' test scores in several schools. Here, schools represent the higher-level factor A, and classes within schools are the nested factor B. Then Y_{ijk} represents the score of student k in class j of school i. The nested term $\beta_{j(i)}$ captures class-specific variability, while ε_{ijk} represents individual student-level deviations.

Mixed models extend this idea by allowing some effects to be treated as random. This formulation captures both fixed effects (systematic differences among groups) and random effects (natural variability between higher-level units):

$$Y_{ij} = \mu + \alpha_i + u_j + \varepsilon_{ij}, \qquad u_j \sim \mathcal{N}(0, \sigma_u^2),$$

where u_j represents the random variation among groups (e.g., classes or individuals).

Interpretation. In this mixed model:

- α_i is a fixed effect, comparing specific levels of factor A.
- u_j is a random effect, modeling variability between sampled groups, assumed to come from a normal distribution with variance σ_u^2.
- ε_{ij} is the residual (within-group) variance.

Variance decomposition. The total variability in Y_{ij} can be conceptually decomposed into:

$$\mathrm{Var}(Y_{ij}) = \underbrace{\sigma_u^2}_{\text{between-group}} + \underbrace{\sigma^2}_{\text{within-group}},$$

highlighting how mixed models separate systematic group effects from random variability.

Example: Hierarchical Models for Classroom Data

Problem Statement

Consider a small educational dataset illustrating a **hierarchical structure**, where **students are nested within classes**. Two classes each contain three students, and their scores on a standardized test are recorded as follows:

$$\text{Class 1: } Y_{11} = 70,\ Y_{12} = 75,\ Y_{13} = 80$$

$$\text{Class 2: } Y_{21} = 65,\ Y_{22} = 68,\ Y_{23} = 72$$

We assume the following hierarchical model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, 2;\ j = 1, 2, 3,$$

where:

- μ denotes the **true overall mean**, in practice, μ is unknown and is estimated by the sample grand mean, denoted $\bar{Y}$,
- α_i is the random effect associated with class i (between-class variability),
- ε_{ij} is the residual error of student j within class i (within-class variability).

Questions

1. Compute the estimator $\bar{Y}$ of the overall mean μ (grand mean) across all students.
2. Compute the mean score for each class, $\bar{Y}_1$ and $\bar{Y}_2$.
3. Compute the Total Sum of Squares (SST) and decompose it into the Between-Class Sum of Squares (SSB) and Within-Class Sum of Squares (SSW).
4. Estimate the variance components $\sigma^2_{\text{between}}$ and σ^2_{within}.

Solution

1. Compute Overall Mean

$$\bar{Y} = 71.67$$

2. Compute Class Means

$$\bar{Y}_1 = 75, \qquad \bar{Y}_2 = 68.33$$

3. Compute Sum of Squares (SST, SSB, SSW) Total Sum of Squares (SST):

$$SST = \sum_{i,j}(Y_{ij} - \bar{Y})^2 = 141.31$$

Between-Class Sum of Squares (SSB):

$$SSB = \sum_{i} n_i(\bar{Y_i} - \bar{Y})^2 = 66.66$$

Within-Class Sum of Squares (SSW):

$$SSW = \sum_{i}\sum_{j}(Y_{ij} - \bar{Y_i})^2 = 74.65$$

$$\text{Check: } SST = SSB + SSW = 66.66 + 74.65 \approx 141.31$$

4. Variance Components Classical ANOVA:

$$\sigma^2_{\text{between (ANOVA)}} = \frac{SSB}{k-1} = \frac{66.66}{1} = 66.66, \qquad \sigma^2_{\text{within}} = \frac{SSW}{N-k} = \frac{74.65}{4} = 18.66$$

Hierarchical Mixed Model (REML estimate):

$$\sigma^2_{\text{between (REML)}} = \frac{MS_B - MS_W}{n} = \frac{66.66 - 18.66}{3} = 16, \qquad \sigma^2_{\text{within (REML)}} = MS_W = 18.66$$

Remark 5.2.1. *ANOVA between-class variance (66.66) reflects the sample-level difference of class means. Mixed model separates the true population variance between classes (16) from the sampling effect, keeping the same within-class variance (18.66) as in ANOVA.*

Practical Work 15: Hierarchical Models for Classroom Data in R

Listing 5.2: Hierarchical Mixed Model and Visualization - Stable Graph

```
# 1. Load required packages
library(lme4)
library(ggplot2)

# 2. Create dataset
data <- data.frame(
  Score = c(70, 75, 80, 65, 68, 72),
  Class = factor(rep(1:2, each = 3)),
  Student = factor(1:6)
)

# 3. Fit random intercept model using REML
mixed_model <- lmer(Score ~ 1 + (1 | Class), data = data, REML =
   TRUE)

```

```r
# 4. Classical one-way ANOVA
anova_model <- anova(lm(Score ~ Class, data = data))

# 5. Extract key components
overall_mean <- mean(data$Score)  # stable overall mean
var_between  <- as.numeric(VarCorr(mixed_model)$Class)
var_within   <- sigma(mixed_model)^2
n_per_class  <- table(data$Class)[1]
MSB_like     <- var_within + n_per_class * var_between

# 6. Create list-style summary
comparison <- list(
  "Overall Mean" = overall_mean,
  "Between-Class Variance (Mixed Model)" = var_between,
  "Residual Variance (Within-Class)" = var_within,
  "Adjusted Between-Class MSB (Mixed Model)" = MSB_like,
  "ANOVA Between-Class MSB" = anova_model$`Mean Sq`[1],
  "ANOVA Within-Class Variance" = anova_model$`Mean Sq`[2]
)

# 7. Print comparison
print(comparison)

# 8. Visualize hierarchical structure (stable)
class_means <- tapply(data$Score, data$Class, mean)  # class means

ggplot(data, aes(x = Class, y = Score)) +
  geom_jitter(width = 0.1, height = 0, color = "blue", size = 3) +  # individual points
  geom_point(aes(y = class_means[Class]), shape = 15, size = 5, color = "red") +  # class means
  geom_hline(yintercept = overall_mean, linetype = "dashed", color = "green", linewidth = 1) +  # overall mean
  labs(title = "Hierarchical Structure: Students Nested in Classes",
       x = "Class", y = "Score") +
  theme_minimal()
```

The plot generated from the R code above (Figure 5.2) intuitively illustrates the hierarchical structure:

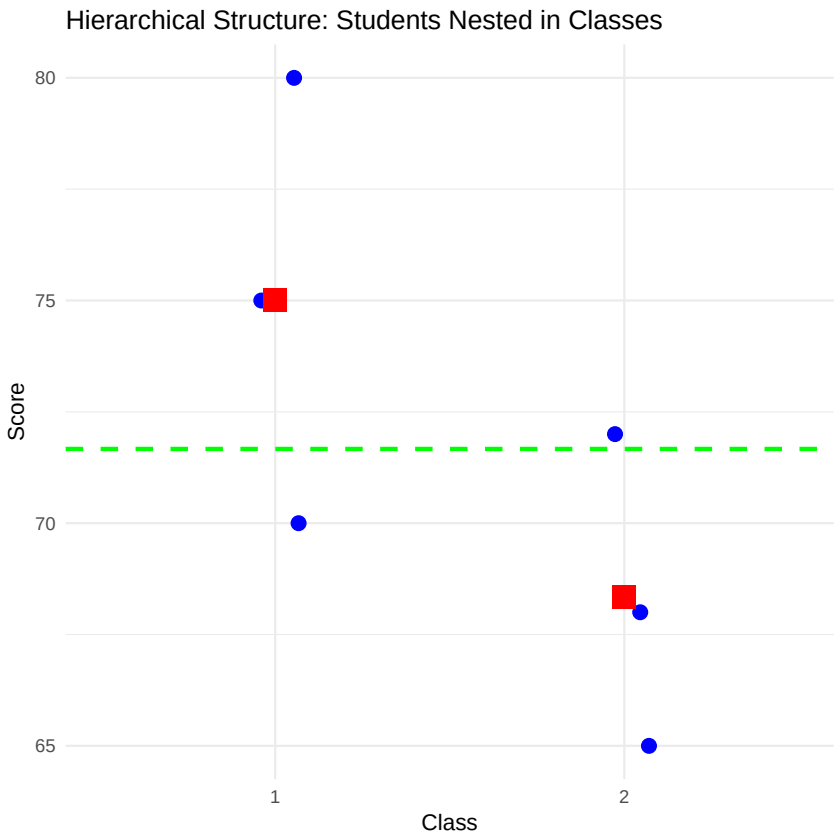

Figure 5.2: Graphical illustration of the hierarchical model

Results and Discussion

The figure illustrates a hierarchical structure where students (level 1) are nested within classes (level 2). The horizontal axis represents the classes, while the vertical axis shows the students' test scores. The blue points correspond to the individual student scores in each class, showing the within-class variability. The red squares represent the mean score of each class (75 for class 1 and 68.33 for class 2). The green dashed line indicates the overall mean (grand mean) calculated across all students. The difference between the class means and the grand mean reflects the between-class variability, while the dispersion of the individual scores around their class mean represents the within-class variability. Thus, the graph visually illustrates the decomposition of the total variability into between-class and within-class components, which is the basis of ANOVA and hierarchical (mixed) models.

This example demonstrates that a hierarchical (mixed) model separates the variability due to differences between classes from the variability among students within each class, quantifying both as variance components.

Bibliography

[1] Belsley, D. A., Kuh, E., & Welsch, R. E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, 1980.

[2] Chatterjee, S., & Hadi, A. S., *Regression Analysis by Example*, 4th Edition, Wiley, 2006.

[3] Chatterjee, S., & Simonoff, J. S., *Handbook of Regression Analysis with Applications in R*, Wiley, 2019.

[4] Faraway, J. J., *Linear Models with R*, 3rd Edition, CRC Press, 2025.

[5] Fox, J., *Applied Regression Analysis and Generalized Linear Models*, 3rd Edition, Sage, 2015.

[6] Hoaglin, D. C., et al., *How to Interpret the ANOVA*, Sage, 2000.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R., *An Introduction to Statistical Learning*, Springer, 2013.

[8] Kutner, M. H., Nachtsheim, C. J., & Neter, J., *Applied Linear Statistical Models*, 5th Edition, McGraw-Hill, 2005.

[9] Matzner-Løber, É., *Régression : Théorie et applications*, Springer Science & Business Media, 2007.

[10] Montgomery, D. C., *Design and Analysis of Experiments*, 8th Edition, Wiley, 2012.

[11] Rao, C. R., & Toutenburg, H., *Linear Models*, Springer New York, 1995, pp. 3-18.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025. https://www.R-project.org/

[13] Tibshirani, R., *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society: Series B, 58(1), 267–288, 1996.

[14] Tokpavi Sessi, *Le modèle de régression linéaire multiple et la méthode des moindres carrés ordinaires*, EconomiX-CNRS, Université Paris Ouest.

[15] Wooldridge, J. M., *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2010.

[16] Zerdazi, D., *L'approche neuronale de l'inférence statistique*, Thèse de doctorat, Université Frères Mentouri - Constantine 1, 2017.