STAT8561 - Linear Statistical Analysis I

Chi-Kuang Yeh

2026-03-28

Table of contents

Preface

The topic of this course includes statistical inference, Multivariate normal distribution, distribution of quadratic forms, linear models, regression models and experimental design models.

Prerequisites

Math 4751/6751 Mathematical Statistics I and Math 4752/6752 Mathematics and Statistics II.

Instructor

Chi-Kuang Yeh, Assistant Professor in the Department of Mathematics and Statistics, Georgia State University.

- Office: Suite 1407, 25 Park Place.
- Email: cyeh@gsu.edu.

Office Hour

TBA

Grade Distribution

- TBA

Assignment

- ☐ A1, TBA

Exam

- TBA

Topics and Corresponding Lectures

Those chapters are based on the lecture notes. This part will be updated frequently.

Status	Chapter	Topic	Lecture
	Ch. 1	Welcome and Overview	1

Recommended Textbooks

- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2021. (Montgomery et al. 2021)
- Seber, George AF, and Alan J. Lee. *Linear Regression Analysis*. John Wiley & Sons, 2003. (Seber and Lee 2003)

Side Readings

- Montgomery, Douglas C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons. (Montgomery 2017)

1 Introduction

This is an introduction to the course for linear statistical analysis. It will give you an overview of what we will be covering in the course and how to get the most out of it. We will be covering a wide range of topics in linear statistical analysis, including linear regression, generalized linear models, and mixed effects models. We will also be discussing the assumptions underlying these models and how to check them.

The course will be structured around lectures, homework assignments, and a final project. The lectures will cover the theoretical aspects of the material, while the homework assignments will give you the opportunity to apply what you have learned to real data sets. The final project will allow you to explore a topic of your choice in more depth and present your findings to the class.

To get the most out of this course, it is important to attend all lectures and complete all homework assignments. It is also important to ask questions and participate in class discussions. The more you engage with the material, the more you will learn.

I am looking forward to a great semester and I hope you are too!

To illustrate the concepts we will be covering in this course, let's consider a simple example. Suppose we have a data set of heights and weights of individuals. We want to understand the relationship between height and weight, and we can use linear regression to model this relationship.

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

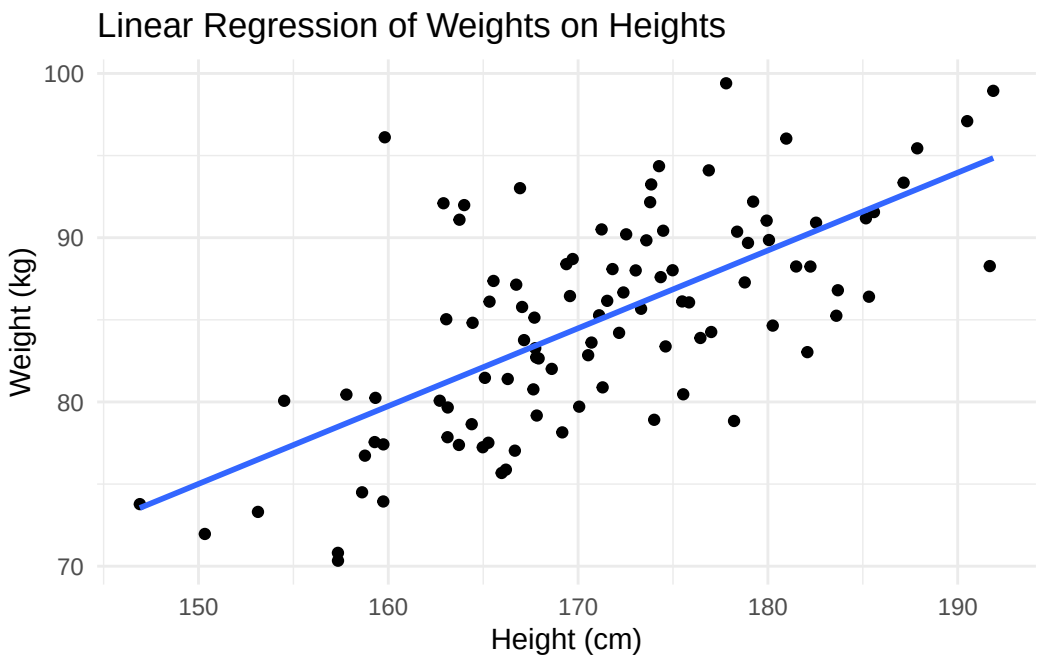
```r
# Create a sample data set
set.seed(123)

theme_set(theme_minimal())

n <- 100
heights <- rnorm(n, mean = 170, sd = 10)
weights <- 0.5 * heights + rnorm(n, mean = 0, sd
 = 5)
data <- data.frame(heights, weights)
# Fit a linear regression model
model <- lm(weights ~ heights, data = data)
# Summarize the model
summary(model)
```

```
Call:
lm(formula = weights ~ heights, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5367 -3.4175 -0.4375  2.9032 16.4520

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.94607    9.14588   0.431    0.667
heights      0.47376    0.05344   8.865  3.5e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.854 on 98 degrees of freedom
Multiple R-squared:  0.4451,    Adjusted R-squared:  0.4394
F-statistic:  78.6 on 1 and 98 DF,  p-value: 3.497e-14
```

```r
# Plot the data and the fitted line
ggplot(data, aes(x = heights, y = weights)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  labs(title = "Linear Regression of Weights on Heights",
       x = "Height (cm)",
       y = "Weight (kg)")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



In this example, we generated a data set of heights and weights, fitted a linear regression model to the data, and visualized the relationship between height and weight. This is just a simple example, but it illustrates the types of analyses we will be doing in this course. We will be covering much more complex models and data sets as we progress through the semester.

2 Week 1 Overview

In this first week, we introduce the mathematical language and statistical framework that will be used throughout the course. Our focus is on the notation of vectors and matrices, vectors of random variables, expectation and covariance operators, and the basic form of the linear regression model.

2.1 Learning Objectives

By the end of this week, students should be able to:

- use standard matrix notation for linear statistical models;
- distinguish between scalars, vectors, matrices, random variables, and random vectors;
- compute expectations and covariance matrices for random vectors;
- interpret the linear regression model in matrix form;
- understand why projection and least squares will play a central role in this course.

2.2 Reading

Recommended reading for this week:

- Seber and Lee, Chapter 1:
  - 1.1 Notation
  - 1.2 Statistical Models
  - 1.3 Linear Regression Models
  - 1.4 Expectation and Covariance Operators
- Optional preview:
  - Chapter 2: Multivariate Normal Distribution

2.3 Why Linear Statistical Analysis?

Linear statistical analysis is one of the central foundations of graduate statistics. Many methods that at first look different are built on the same underlying structure:

- regression,
- analysis of variance,
- analysis of covariance,
- prediction,
- model comparison,
- and parts of generalized linear modelling.

A major goal of this course is to see these topics under a unified framework.

2.4 A unifying point of view

A large part of the course can be summarized by the model

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

where

- Y is a response vector,
- X is a design matrix,
- β is an unknown parameter vector,
- ε is a random error vector.

This compact expression contains a great deal of statistical structure. Over the semester, we will study how to estimate β, quantify uncertainty, test hypotheses, diagnose model failures, and make predictions.

2.5 Basic Notation

2.6 Scalars, vectors, and matrices

We use the following conventions throughout the course:

- scalars are written in lowercase italic letters, such as a, b, n;
- vectors are written in bold lowercase letters, such as $\mathbf{x}$, $\mathbf{y}$;
- matrices are written in bold uppercase letters, such as $\mathbf{X}$, $\mathbf{A}$;

- random variables are often written in uppercase letters, such as Y;
- realizations of random variables are written in lowercase letters, such as y.

A column vector in $\mathbb{R}^n$ is written as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

An $n \times p$ matrix is written as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

2.7 Transpose, inverse, and rank

If $\mathbf{A}$ is a matrix, then:

- $\mathbf{A}^\top$ denotes its transpose;
- $\mathbf{A}^{-1}$ denotes its inverse, when it exists;
- $\text{rank}(\mathbf{A})$ denotes its rank;
- $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.

2.8 Inner product and norm

For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the inner product is

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

The Euclidean norm is

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}.$$

These ideas are fundamental because least squares estimation is based on minimizing squared Euclidean distance.

2.9 Random Vectors

2.9.1 Definition

A random vector is a vector whose entries are random variables. For example,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

is an n-dimensional random vector.

In linear models, the response is naturally treated as a random vector.

2.10 Mean vector

The mean vector of $\mathbf{Y}$ is

$$\mathbb{E}[\mathbf{Y}] = \begin{bmatrix} \mathbb{E}[Y_1] \\ \mathbb{E}[Y_2] \\ \vdots \\ \mathbb{E}[Y_n] \end{bmatrix}.$$

We often write

$$\mathbb{E}[\mathbf{Y}] = \mu.$$

2.11 Covariance matrix

The covariance matrix of $\mathbf{Y}$ is

$$\mathrm{Var}(\mathbf{Y}) = \mathbb{E}\left[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^\top\right].$$

Equivalently,

$$\mathrm{Var}(\mathbf{Y}) = \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] - \mu\mu^\top.$$

If

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix},$$

then

$$\mathrm{Var}(\mathbf{Y}) = \begin{bmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, Y_2) & \mathrm{Cov}(Y_1, Y_3) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Var}(Y_2) & \mathrm{Cov}(Y_2, Y_3) \\ \mathrm{Cov}(Y_3, Y_1) & \mathrm{Cov}(Y_3, Y_2) & \mathrm{Var}(Y_3) \end{bmatrix}.$$

##Expectation and Covariance Operators

2.11.1 Linearity of expectation

If $\mathbf{A}$ is a constant matrix and $\mathbf{b}$ is a constant vector, then

$$\mathbb{E}[\mathbf{A}\mathbf{Y} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{Y}] + \mathbf{b}.$$

This is one of the most useful identities in the course.

2.11.2 Covariance of linear transformations

If $\mathbf{A}$ is a constant matrix, then

$$\mathrm{Var}(\mathbf{A}\mathbf{Y}) = \mathbf{A}\,\mathrm{Var}(\mathbf{Y})\,\mathbf{A}^\top.$$

More generally, if $\mathbf{Y}$ and $\mathbf{Z}$ are random vectors, then

$$\mathrm{Cov}(\mathbf{A}\mathbf{Y}, \mathbf{B}\mathbf{Z}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{Y}, \mathbf{Z})\,\mathbf{B}^\top.$$

These formulas are essential for deriving the variance of estimators later.

2.11.3 Special case: independent components

If $Y_1, \ldots, Y_n$ are independent and each has variance σ^2, then

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

This is the most common starting assumption in classical linear regression.

2.12 Statistical Models

A statistical model is a set of probability distributions that may plausibly describe the data-generating mechanism.

2.12.1 General idea

Suppose we observe data y from a random quantity Y. A model introduces assumptions about the distribution of Y, often indexed by an unknown parameter θ.

For example:

$$Y \sim N(\mu, \sigma^2)$$

with unknown parameters μ and σ^2.

In regression, the model is not only about the marginal distribution of the response but also about how the mean changes with explanatory variables.

2.12.2 Deterministic part and random part

A useful way to think about a statistical model is:

$$\text{data} = \text{systematic part} + \text{random part}.$$

For linear regression, this becomes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Here,

- $\beta_0 + \beta_1 x_i$ is the systematic part;
- ε_i is the random part.

2.13 The Linear Regression Model

2.13.1 Simple linear regression

The simplest regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n.$$

Typical assumptions are

$$\mathbb{E}[\varepsilon_i] = 0, \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2, \qquad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j.$$

2.13.2 Matrix form

We can write the model compactly as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

This notation will allow us to treat simple regression, multiple regression, ANOVA, and ANCOVA in one common language.

2.13.3 Mean and variance under the model

If

$$\mathbb{E}[\varepsilon] = \mathbf{0} \qquad \text{and} \qquad \mathrm{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n,$$

then

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

and

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

These are immediate consequences of the expectation and covariance rules above.

2.13.4 Geometry Preview

A central idea in linear regression is projection.

The fitted values $\hat{\mathbf{Y}}$ will later be obtained by projecting $\mathbf{Y}$ onto the column space of $\mathbf{X}$.

2.13.5 Column space

The column space of $\mathbf{X}$ is

$$\mathcal{C}(\mathbf{X}) = \{\mathbf{X}\beta : \beta \in \mathbb{R}^p\}.$$

This is the set of all mean vectors that the model can represent.

2.14 Why projection matters

The least squares estimator chooses $\widehat{\beta}$ so that

$$\|\mathbf{Y} - \mathbf{X}\beta\|^2$$

is minimized.

So regression is not only algebra. It is also geometry.

3 Worked Example by Hand

Suppose we observe the following data:

x_i	0	1	2	3
y_i	1	3	3	5

Then

$$\mathbf{Y} = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 5 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

We will learn next week that the least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

For now, the main point is to understand how the model is written and how the data are represented in matrix form.

4 R Demonstration

```r
x <- c(0, 1, 2, 3)
y <- c(1, 3, 3, 5)

fit <- lm(y ~ x)
summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4
-0.2  0.6 -0.6  0.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2000     0.5292   2.268   0.1515
x             1.2000     0.2828   4.243   0.0513 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6325 on 2 degrees of freedom
Multiple R-squared:    0.9, Adjusted R-squared:   0.85
F-statistic:    18 on 1 and 2 DF,  p-value: 0.05132
```

```r
# Inspecting the design matrix
model.matrix(fit)
```

```
  (Intercept) x
1           1 0
2           1 1
3           1 2
```

```
4           1 3
attr(,"assign")
[1] 0 1
```

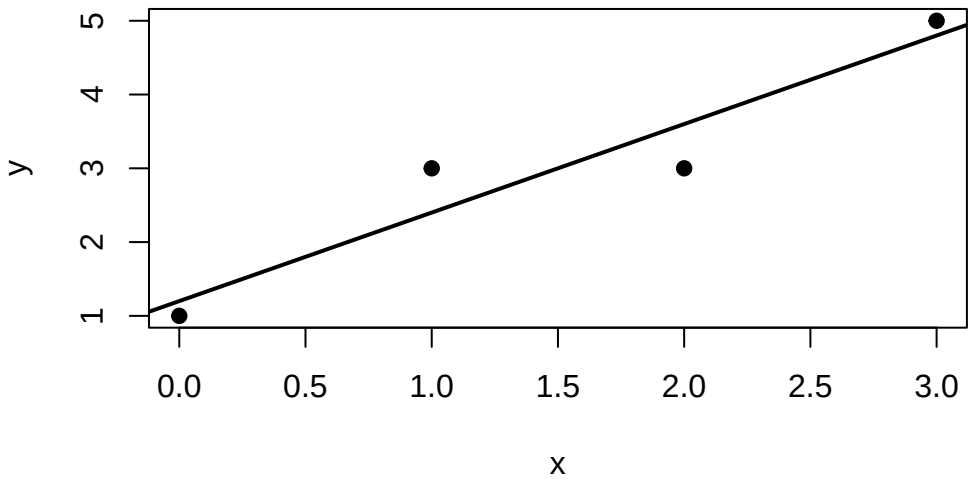
```r
# Fitted values and residuals
fitted(fit)
```

```
  1   2   3   4
1.2 2.4 3.6 4.8
```

```r
residuals(fit)
```

```
   1    2    3    4
-0.2  0.6 -0.6  0.2
```

```r
# quick plot
plot(x, y, pch = 19, xlab = "x", ylab = "y")
abline(fit, lwd = 2)
```



5 In-Class Discussion Questions

. Why is it helpful to write regression models in matrix form rather than only scalar notation? 2. What does the covariance matrix tell us that separate variances do not? 3. What is the interpretation of the column space of $\mathbf{X}$? 4. In what sense is regression a projection problem?

6 Practice Problems

Conceptual 1. Explain the difference between a random variable and a random vector. 2. Explain why $\mathrm{Var}(\mathbf{Y})$ must be a symmetric matrix. 3. Give an example of a statistical model outside regression.

Computational

Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

with

$$\mathbb{E}[\mathbf{Y}] = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \mathrm{Var}(\mathbf{Y}) = \begin{bmatrix} 4 & 1 \\ 1 & 9 \end{bmatrix}.$$

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

Compute:

1. $\mathbb{E}[\mathbf{AY} + \mathbf{b}]$
2. $\mathrm{Var}(\mathbf{AY})$

Regression setup

For the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

write out the matrices $\mathbf{Y}$, $\mathbf{X}$, β, and ε for $n = 5$ observations.

6.1 Suggested Homework

Complete the following:

- review matrix multiplication and transpose rules;
- derive $\mathbb{E}[\mathbf{AY} + \mathbf{b}]$ from first principles;
- derive $\text{Var}(\mathbf{AY})$ using the definition of covariance;
- write the simple linear regression model in matrix form for a dataset of your choice;
- fit a simple regression in R and report:
- the estimated coefficients,
- fitted values,
- residuals,
- and a scatterplot with the fitted line.

6.2 Summary

This week introduced the notation and basic probabilistic tools needed for the rest of the course. We defined random vectors, mean vectors, covariance matrices, and the matrix form of the linear regression model. These ideas will support everything that follows.

Next week, we will study least squares estimation and the geometry of projection in more detail.

For some optional review of the matrix algebra, see Chapter 41

7 Week 2 Overview

In this week, we study the core idea behind linear regression: **least squares estimation**. We derive the ordinary least squares estimator, interpret it geometrically as a projection, and introduce the fitted values, residuals, and the hat matrix.

7.1 Learning Objectives

By the end of this week, students should be able to:

- define the least squares criterion for a linear model;
- derive the normal equations;
- obtain the ordinary least squares estimator when the design matrix has full column rank;
- interpret least squares as an orthogonal projection;
- define the fitted values, residuals, and hat matrix;
- explain the orthogonality properties of residuals.

7.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - Chapter 3: sections on least squares estimation
- Montgomery, Peck, and Vining:
  - introductory sections on estimation in linear regression

8 1. Review of the Linear Model

Recall the linear model from Week 1:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

where

- $\mathbf{Y}$ is an $n \times 1$ response vector,
- $\mathbf{X}$ is an $n \times p$ design matrix,
- β is a $p \times 1$ parameter vector,
- ε is an $n \times 1$ error vector.

Under the classical setup, we often assume

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

Hence,

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta, \qquad \mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

Our goal is to estimate β from the observed data.

9 2. The Least Squares Criterion

9.1 2.1 Motivation

For any candidate value β, the model predicts

$$\mathbf{X}\beta.$$

The discrepancy between the observed response $\mathbf{Y}$ and the model mean $\mathbf{X}\beta$ is

$$\mathbf{Y} - \mathbf{X}\beta.$$

This vector is called the **residual vector** for the candidate β.

A natural idea is to choose β so that this discrepancy is as small as possible.

9.2 2.2 Residual sum of squares

The least squares criterion is

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta).$$

Equivalently,

$$S(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

We choose $\widehat{\beta}$ to minimize $S(\beta)$.

9.3 2.3 Expansion of the criterion

Expanding the quadratic form gives

$$S(\beta) = \mathbf{Y}^\top \mathbf{Y} - 2\beta^\top \mathbf{X}^\top \mathbf{Y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta.$$

This is a quadratic function of β.

10 3. Derivation of the Normal Equations

To minimize $S(\beta)$, differentiate with respect to β:

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\beta.$$

Setting this equal to zero yields the **normal equations**:

$$\mathbf{X}^\top \mathbf{X}\widehat{\beta} = \mathbf{X}^\top \mathbf{Y}.$$

These are the equations that define the ordinary least squares estimator.

10.1 3.1 Full column rank case

If $\mathbf{X}$ has full column rank p, then $\mathbf{X}^\top \mathbf{X}$ is invertible, and the unique least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}.$$

This is the ordinary least squares estimator, or OLS estimator.

10.2 3.2 When is this formula valid?

The closed-form expression above requires

$$\operatorname{rank}(\mathbf{X}) = p.$$

This means the columns of $\mathbf{X}$ are linearly independent. If they are not, then $\mathbf{X}^\top \mathbf{X}$ is singular, and special care is needed. We will discuss rank deficiency later in the course.

11 4. Geometric Interpretation

11.1 4.1 Column space of the design matrix

Recall the column space of $\mathbf{X}$:

$$\mathcal{C}(\mathbf{X}) = \{\mathbf{X}\beta : \beta \in \mathbb{R}^p\}.$$

This is the set of all vectors that can be represented by the linear model.

11.2 4.2 Least squares as projection

The fitted value vector is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

Since $\hat{\mathbf{Y}} \in \mathcal{C}(\mathbf{X})$, least squares chooses the vector in the column space of $\mathbf{X}$ that is closest to $\mathbf{Y}$ in Euclidean distance.

Thus, $\hat{\mathbf{Y}}$ is the **orthogonal projection** of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$.

11.3 4.3 Residual vector

The residual vector is

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}.$$

Geometrically, $\mathbf{e}$ is the part of $\mathbf{Y}$ orthogonal to the model space $\mathcal{C}(\mathbf{X})$.

12 5. Orthogonality Properties

A central property of least squares is that the residual vector is orthogonal to every column of $\mathbf{X}$.

12.1 5.1 Algebraic proof

Starting from the normal equations,

$$\mathbf{X}^\top \mathbf{X}\widehat{\beta} = \mathbf{X}^\top \mathbf{Y},$$

we rearrange to obtain

$$\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \mathbf{0}.$$

Since $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$, this becomes

$$\mathbf{X}^\top \mathbf{e} = \mathbf{0}.$$

Therefore, $\mathbf{e}$ is orthogonal to every column of $\mathbf{X}$.

12.2 5.2 Consequences

This implies:

- the residuals sum to zero if the model includes an intercept;
- fitted values and residuals are orthogonal;
- the least squares fit is a projection onto the model space.

If the first column of $\mathbf{X}$ is $\mathbf{1}$, then

$$\mathbf{1}^\top \mathbf{e} = 0,$$

so

$$\sum_{i=1}^{n} e_i = 0.$$

13 6. The Hat Matrix

13.1 6.1 Definition

In the full rank case,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Define the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

Then

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}.$$

The matrix $\mathbf{H}$ is called the **hat matrix** because it puts the "hat" on $\mathbf{Y}$.

13.2 6.2 Properties of the hat matrix

The hat matrix satisfies two important properties:

13.2.1 Symmetry

$$\mathbf{H}^\top = \mathbf{H}.$$

13.2.2 Idempotence

$$\mathbf{H}^2 = \mathbf{H}.$$

A matrix that is both symmetric and idempotent is the matrix of an orthogonal projection. Thus, $\mathbf{H}$ projects vectors onto $\mathcal{C}(\mathbf{X})$.

13.3 6.3 Residual maker matrix

Define

$$\mathbf{M} = \mathbf{I}_n - \mathbf{H}.$$

Then the residual vector can be written as

$$\mathbf{e} = \mathbf{M}\mathbf{Y}.$$

The matrix $\mathbf{M}$ is also symmetric and idempotent, and it projects onto the orthogonal complement of $\mathcal{C}(\mathbf{X})$.

14 7. Sum of Squares Decomposition

Because $\hat{\mathbf{Y}}$ and $\mathbf{e}$ are orthogonal, we have

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}$$

with

$$\hat{\mathbf{Y}}^\top \mathbf{e} = 0.$$

Hence,

$$\|\mathbf{Y}\|^2 = \|\hat{\mathbf{Y}}\|^2 + \|\mathbf{e}\|^2.$$

This is a Pythagorean identity.

In regression with an intercept, a more familiar decomposition is

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

Later we will call these:

- total sum of squares (SST),
- regression sum of squares (SSR),
- error sum of squares (SSE).

15 8. Statistical Properties of the OLS Estimator

Assume

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon] = \mathbf{0}, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

15.1 8.1 Expectation

Using

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

we get

$$\mathbb{E}[\hat{\beta}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{Y}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta = \beta.$$

Thus, OLS is unbiased.

15.2 8.2 Variance

Also,

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathrm{Var}(\mathbf{Y}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Since $\mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, this becomes

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

This formula will be fundamental for confidence intervals and hypothesis tests later.

16 9. Worked Example by Hand

Consider the simple regression dataset

x_i	0	1	2	3
y_i	1	3	3	5

Then

$$\mathbf{Y} = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 5 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

First compute

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}, \qquad \mathbf{X}^\top \mathbf{Y} = \begin{bmatrix} 12 \\ 24 \end{bmatrix}.$$

Hence,

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Since

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix},$$

we obtain

$$\widehat{\beta} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 12 \\ 24 \end{bmatrix} = \begin{bmatrix} 1.2 \\ 1.2 \end{bmatrix}.$$

Thus, the fitted line is

$$\widehat{Y} = 1.2 + 1.2x.$$

17 10. R Demonstration

17.1 10.1 Fit the model

```
x <- c(0, 1, 2, 3)
y <- c(1, 3, 3, 5)

fit <- lm(y ~ x)
coef(fit)
```

```
(Intercept)           x
        1.2         1.2
```

```
X <- model.matrix(fit)
X
```

```
  (Intercept) x
1           1 0
2           1 1
3           1 2
4           1 3
attr(,"assign")
[1] 0 1
```

```
Y <- matrix(y, ncol = 1)
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y
beta_hat
```

```
            [,1]
(Intercept)  1.2
x            1.2
```

```r
y_hat <- X %*% beta_hat
e <- Y - y_hat

y_hat
```

```
  [,1]
1  1.2
2  2.4
3  3.6
4  4.8
```

```r
e
```

```
  [,1]
1 -0.2
2  0.6
3 -0.6
4  0.2
```

```r
H <- X %*% solve(t(X) %*% X) %*% t(X)
round(H, 4)
```

```
     1   2   3    4
1  0.7 0.4 0.1 -0.2
2  0.4 0.3 0.2  0.1
3  0.1 0.2 0.3  0.4
4 -0.2 0.1 0.4  0.7
```

```r
round(H - t(H), 8)
```

```
  1 2 3 4
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
```

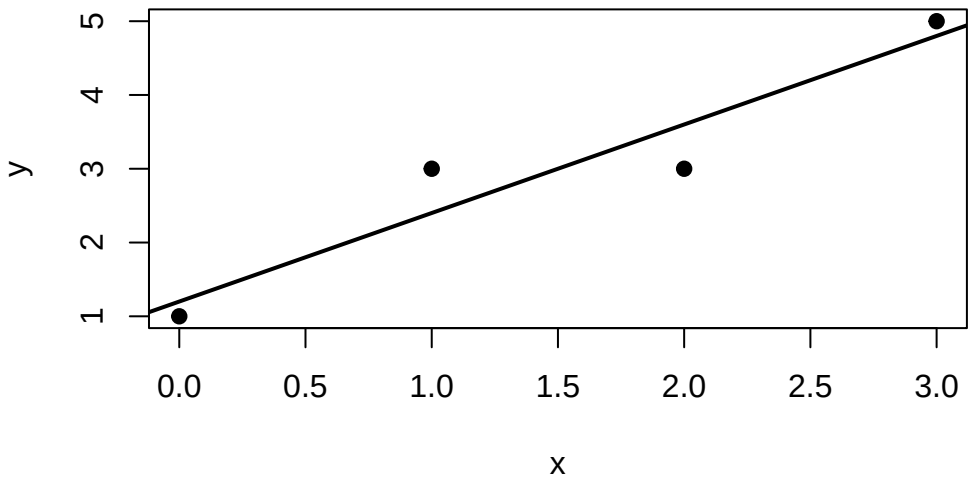
```r
round(H %*% H - H, 8)
```

```
  1 2 3 4
1 0 0 0 0
2 0 0 0 0
3 0 0 0 0
4 0 0 0 0
```

```
round(t(X) %*% e, 8)
```

```
            [,1]
(Intercept)    0
x              0
```

```
plot(x, y, pch = 19, xlab = "x", ylab = "y")
abline(fit, lwd = 2)
```



11. Interpretation of the Geometry

The vector $\mathbf{Y}$ lives in $\mathbb{R}^n$. The model space $\mathcal{C}(\mathbf{X})$ is a p-dimensional subspace of $\mathbb{R}^n$ when $\mathbf{X}$ has full rank.

Least squares finds the point in this subspace that is closest to $\mathbf{Y}$.

This is why • fitted values are projections, • residuals are orthogonal to the model space, • sum of squares decompositions are geometric identities.

12. In-Class Discussion Questions

    1. Why does minimizing $|\mathbf{Y} - \mathbf{X}\beta|^2$ lead to orthogonality?
    2. Why is the hat matrix called a projection matrix?
    3. Why does the inclusion of an intercept imply that the residuals sum to zero?
    4. What fails when $\mathbf{X}$ does not have full column rank?

13. Practice Problems

Conceptual 1. Explain why $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$ is a geometric statement. 2. Give an interpretation of $\mathcal{C}(\mathbf{X})$ in the context of regression. 3. Explain the difference between $\mathbf{H}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$.

Computational

Let

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 2 \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}.$$

1. Compute $\mathbf{X}^\top \mathbf{X}$.
2. Compute $\mathbf{X}^\top \mathbf{Y}$.
3. Find $\widehat{\beta}$.
4. Compute $\widehat{\mathbf{Y}}$ and $\mathbf{e}$.
5. Verify that $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$.

Proof-based

Show that the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

is symmetric and idempotent.

14. Suggested Homework

Complete the following tasks:

- derive the normal equations from the least squares criterion;
- prove that $\widehat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$;
- prove that $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$;
- verify that $\mathbf{H}$ is symmetric and idempotent;
- fit a simple regression model in R and compute: $+\ \widehat{\beta}$, $+\ \widehat{\mathbf{Y}}$, $+\ \mathbf{e}$, $\bullet\ \mathbf{H}$.

15. Summary

In this week, we introduced the least squares estimator and showed that it solves the normal equations

$$\mathbf{X}^\top \mathbf{X}\widehat{\beta} = \mathbf{X}^\top \mathbf{Y}.$$

When $\mathbf{X}$ has full column rank,

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}.$$

Geometrically, least squares is projection onto the column space of $\mathbf{X}$. This leads naturally to the hat matrix, residual orthogonality, and sum of squares decompositions.

Next week, we will study the distribution theory of OLS under the normal error model, including estimation of σ^2, standard errors, and inference.

Appendix: Matrix Calculus Facts Used This Week

For a vector β and constant matrix $\mathbf{A}$,

$$\frac{\partial}{\partial \beta}(\mathbf{a}^\top \beta) = \mathbf{a},$$

and if $\mathbf{A}$ is symmetric,

$$\frac{\partial}{\partial \beta}(\beta^\top \mathbf{A}\beta) = 2\mathbf{A}\beta.$$

These identities justify the derivative of the least squares criterion.

18 Week 3: Distribution Theory of OLS and Inference

In this week, we study the sampling distribution of the ordinary least squares estimator under the normal linear model. This allows us to quantify uncertainty, estimate the error variance, construct confidence intervals, perform hypothesis tests, and distinguish between inference for the mean response and prediction for a future observation.

18.1 Learning Objectives

By the end of this week, students should be able to:

- state the normal linear model;
- derive the distribution of the OLS estimator;
- obtain an unbiased estimator of σ^2;
- understand the role of chi-square, t, and F distributions in linear regression;
- construct confidence intervals for regression coefficients and mean responses;
- perform hypothesis tests for individual coefficients and general linear hypotheses;
- distinguish between confidence intervals for the mean response and prediction intervals for a new observation.

18.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on the distribution theory of least squares estimators
  - estimation of the error variance
  - inference for regression coefficients
- Montgomery, Peck, and Vining:
  - sections on confidence intervals, hypothesis testing, and prediction in linear regression

19 1. Review of the Linear Model

Recall the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

where

- $\mathbf{Y}$ is an $n \times 1$ response vector,
- $\mathbf{X}$ is an $n \times p$ design matrix,
- β is a $p \times 1$ unknown parameter vector,
- ε is an $n \times 1$ error vector.

From Week 2, when $\mathbf{X}$ has full column rank, the ordinary least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

We also know that

$$\mathbb{E}[\widehat{\beta}] = \beta, \qquad \mathrm{Var}(\widehat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1},$$

provided that

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

To obtain exact finite-sample inference, we now strengthen the model assumptions.

20 2. The Normal Linear Model

We assume

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n).$$

Equivalently,

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

This is called the **normal linear model**.

Under this assumption, exact sampling distributions can be derived for the OLS estimator, residual sum of squares, and many test statistics.

20.1 Why normality matters

Without normality, OLS is still unbiased under the standard moment assumptions, but exact t and F inference generally no longer holds in finite samples.

Normality gives us:

- exact distribution of $\widehat{\beta}$;
- exact chi-square distribution for the residual sum of squares;
- exact t and F tests.

21 3. Distribution of the OLS Estimator

Since

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

the estimator is a linear transformation of $\mathbf{Y}$.

Because a linear transformation of a multivariate normal vector is again multivariate normal, we immediately obtain:

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}) \,.$$

Thus,

- the mean of $\hat{\beta}$ is β;
- the covariance matrix of $\hat{\beta}$ is $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

21.1 Individual coefficients

For the jth coefficient,

$$\hat{\beta}_j \sim N\big(\beta_j, \sigma^2 c_{jj}\big) \,,$$

where c_{jj} is the jth diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Hence,

$$\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{c_{jj}}} \sim N(0,1).$$

This is useful, but it still depends on the unknown σ.

22 4. Residual Sum of Squares and Estimation of σ^2

22.1 Residual vector

The residual vector is

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

is the hat matrix.

22.2 Residual sum of squares

The residual sum of squares is

$$\text{SSE} = \mathbf{e}^\top \mathbf{e} = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{Y} - \mathbf{X}\widehat{\beta}).$$

Equivalently,

$$\text{SSE} = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

22.3 Distribution of SSE

Under the normal linear model,

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p}.$$

The degrees of freedom are $n - p$ because we estimate p parameters.

22.4 Unbiased estimator of σ^2

Since the mean of a chi-square random variable with k degrees of freedom is k, we have

$$\mathbb{E}[\text{SSE}] = (n-p)\sigma^2.$$

Therefore,

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p}$$

is an unbiased estimator of σ^2.

This quantity is also called the **mean squared error**:

$$\text{MSE} = \frac{\text{SSE}}{n-p}.$$

23 5. Independence Between $\widehat{\beta}$ and SSE

A crucial result under the normal linear model is that

$$\hat{\beta} \quad \text{and} \quad \text{SSE}$$

are independent.

This is special and extremely useful. It is what allows us to replace the unknown σ with $\hat{\sigma}$ and obtain exact t and F distributions.

Intuitively:

- $\hat{\beta}$ depends on the projected part of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$;
- SSE depends on the orthogonal residual part.

Because these parts are orthogonal and jointly normal, they are independent.

24 6. Inference for a Single Coefficient

24.1 6.1 t statistic

Since

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj}),$$

and since

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p},$$

with independence between $\hat{\beta}_j$ and $\hat{\sigma}^2$, it follows that

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t_{n-p}.$$

24.2 6.2 Confidence interval

A $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{1-\alpha/2,\, n-p}\, \hat{\sigma}\sqrt{c_{jj}}.$$

24.3 6.3 Hypothesis test

To test

$$H_0 : \beta_j = \beta_{j,0} \qquad \text{versus} \qquad H_1 : \beta_j \neq \beta_{j,0},$$

we use

$$T = \frac{\hat{\beta}_j - \beta_{j,0}}{\hat{\sigma}\sqrt{c_{jj}}}.$$

Under H_0,

$$T \sim t_{n-p}.$$

For a two-sided test, reject H_0 if

$$|T| > t_{1-\alpha/2,\, n-p}.$$

25 7. Inference for Linear Combinations

Often we are interested not in a single coefficient, but in a linear combination

$$a^\top \beta,$$

where a is a fixed $p \times 1$ vector.

For example:

- a single coefficient;
- the difference between two coefficients;
- the mean response at a given covariate value.

Since

$$a^\top \widehat{\beta} \sim N(a^\top \beta, \sigma^2 a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a)\,,$$

we obtain

$$\frac{a^\top \widehat{\beta} - a^\top \beta}{\widehat{\sigma}\sqrt{a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a}} \sim t_{n-p}.$$

Thus, a confidence interval for $a^\top \beta$ is

$$a^\top \widehat{\beta} \pm t_{1-\alpha/2,\, n-p} \widehat{\sigma}\sqrt{a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a}.$$

26 8. General Linear Hypotheses and F Tests

26.1 8.1 Linear hypothesis

Suppose we want to test

$$H_0 : \mathbf{C}\beta = \mathbf{d},$$

where $\mathbf{C}$ is an $r \times p$ matrix of rank r, and $\mathbf{d}$ is an $r \times 1$ vector.

This is called a **general linear hypothesis**.

Examples include:

- testing one coefficient;
- testing several coefficients simultaneously;
- testing equality of coefficients.

26.2 8.2 F statistic

Under H_0, the appropriate test statistic is

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

Under H_0,

$$F \sim F_{r,\, n-p}.$$

This gives the general framework for multi-parameter tests.

26.3 8.3 Relationship between t and F

When $r = 1$, the F test reduces to the square of the corresponding t test:

$$F = T^2.$$

27 9. Confidence Intervals for the Mean Response

Suppose we want to estimate the mean response at a covariate vector x_0.

Let

$$\mu(x_0) = x_0^\top \beta.$$

The estimator is

$$\hat{\mu}(x_0) = x_0^\top \hat{\beta}.$$

Its variance is

$$\mathrm{Var}(\hat{\mu}(x_0)) = \sigma^2 x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0.$$

Therefore,

$$\frac{x_0^\top \hat{\beta} - x_0^\top \beta}{\hat{\sigma}\sqrt{x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0}} \sim t_{n-p}.$$

A $100(1-\alpha)\%$ confidence interval for the **mean response** at x_0 is

$$x_0^\top \hat{\beta} \pm t_{1-\alpha/2,\, n-p} \hat{\sigma} \sqrt{x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0}.$$

28 10. Prediction Interval for a New Observation

Now suppose we want to predict a **new future response** at x_0:

$$Y_{\text{new}} = x_0^\top \beta + \varepsilon_{\text{new}},$$

where

$$\varepsilon_{\text{new}} \sim N(0, \sigma^2)$$

and is independent of the original data.

The prediction error is

$$Y_{\text{new}} - x_0^\top \hat{\beta}.$$

Its variance is

$$\text{Var}(Y_{\text{new}} - x_0^\top \hat{\beta}) = \sigma^2 \left(1 + x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0\right).$$

Hence, a $100(1-\alpha)\%$ prediction interval for a new observation is

$$x_0^\top \hat{\beta} \pm t_{1-\alpha/2,\, n-p} \hat{\sigma} \sqrt{1 + x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} x_0}.$$

28.1 Key difference

- Confidence interval for the mean response: uncertainty about the regression mean;
- Prediction interval: uncertainty about the regression mean **plus** random individual variation.

Therefore, prediction intervals are always wider.

29 11. Worked Example by Hand

Consider again the dataset

x_i	0	1	2	3
y_i	1	3	3	5

with design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 5 \end{bmatrix}.$$

From Week 2, we found

$$\hat{\beta} = \begin{bmatrix} 1.2 \\ 1.2 \end{bmatrix}.$$

The fitted values are

$$\hat{\mathbf{Y}} = \begin{bmatrix} 1.2 \\ 2.4 \\ 3.6 \\ 4.8 \end{bmatrix},$$

so the residuals are

$$\mathbf{e} = \begin{bmatrix} -0.2 \\ 0.6 \\ -0.6 \\ 0.2 \end{bmatrix}.$$

Thus,

$$\text{SSE} = (-0.2)^2 + 0.6^2 + (-0.6)^2 + 0.2^2 = 0.8.$$

Since $n = 4$ and $p = 2$,

$$\hat{\sigma}^2 = \frac{0.8}{4-2} = 0.4, \qquad \hat{\sigma} = \sqrt{0.4}.$$

Also,

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}.$$

Hence the estimated variance of $\hat{\beta}_1$ is

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = 0.4 \cdot \frac{4}{20} = 0.08,$$

so the standard error is

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{0.08}.$$

To test

$$H_0 : \beta_1 = 0,$$

the t statistic is

$$T = \frac{1.2}{\sqrt{0.08}}.$$

30 12. R Demonstration

30.1 12.1 Fit the model

```
x <- c(0, 1, 2, 3)
y <- c(1, 3, 3, 5)

fit <- lm(y ~ x)
summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
   1    2    3    4
-0.2  0.6 -0.6  0.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2000     0.5292   2.268   0.1515
x             1.2000     0.2828   4.243   0.0513 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6325 on 2 degrees of freedom
Multiple R-squared:    0.9, Adjusted R-squared:   0.85
F-statistic:    18 on 1 and 2 DF,  p-value: 0.05132
```

```
coef(fit)
```

```
(Intercept)           x
        1.2         1.2
```

```
vcov(fit)
```

```
            (Intercept)     x
(Intercept)        0.28 -0.12
x                 -0.12  0.08
```

```
sqrt(diag(vcov(fit)))
```

```
(Intercept)           x
  0.5291503   0.2828427
```

```
deviance(fit)
```

```
[1] 0.8
```

```
sigma(fit)^2
```

```
[1] 0.4
```

```
sigma(fit)
```

```
[1] 0.6324555
```

```
confint(fit)
```

```
                 2.5 %   97.5 %
(Intercept) -1.07674982 3.476750
x           -0.01697397 2.416974
```

```
summary(fit)$coefficients
```

```
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept)      1.2  0.5291503 2.267787 0.1514719
x                1.2  0.2828427 4.242641 0.0513167
```

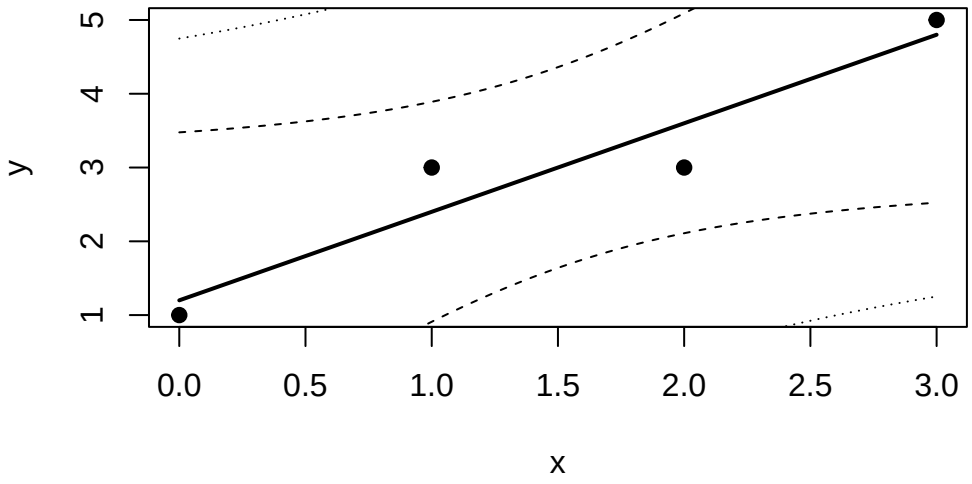
```r
newdat <- data.frame(x = 2)
predict(fit, newdata = newdat, interval = "confidence")
```

```
  fit      lwr      upr
1 3.6 2.109517 5.090483
```

```r
predict(fit, newdata = newdat, interval = "prediction")
```

```
  fit      lwr      upr
1 3.6 0.497313 6.702687
```

```r
xg <- seq(min(x), max(x), length.out = 100)
out_conf <- predict(fit, newdata = data.frame(x = xg), interval = "confidence")
out_pred <- predict(fit, newdata = data.frame(x = xg), interval = "prediction")

plot(x, y, pch = 19, xlab = "x", ylab = "y")
lines(xg, out_conf[, "fit"], lwd = 2)
lines(xg, out_conf[, "lwr"], lty = 2)
lines(xg, out_conf[, "upr"], lty = 2)
lines(xg, out_pred[, "lwr"], lty = 3)
lines(xg, out_pred[, "upr"], lty = 3)
```



31 13. Interpretation of Standard Output

In regression output from summary(lm(…)), the key columns are:

- `Estimate`: the estimated coefficient;
- `Std. Error`: the estimated standard deviation of the estimator;
- `t value`: the test statistic for testing whether the coefficient equals zero;
- `Pr(>|t|)`: the corresponding p-value.

The output also reports:

- residual standard error;
- degrees of freedom;
- R^2 and adjusted R^2;
- an overall F test.

We will discuss the overall ANOVA-style decomposition more formally soon.

31.1 14. In-Class Discussion Questions

1. Why does normality lead to exact finite-sample inference?
2. Why do we divide SSE by $n-p$ rather than n?
3. Why are prediction intervals wider than confidence intervals for the mean response?
4. Why is independence between $\widehat{\beta}$ and SSE so important?

31.2 15. Practice Problems

Conceptual 1. Explain the difference between the sampling distribution of $\widehat{\beta}_j$ and the distribution of Y_i. 2. Explain why $\hat{\sigma}^2 = \text{SSE}/(n-p)$ is unbiased. 3. Explain why the t distribution appears instead of the normal distribution.

Computational

Suppose

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} 0.5 & 0.1 & 0.1 & 0.2 \end{bmatrix},$$

$\hat{\beta} = (2, -1)^\top$, and $\hat{\sigma}^2 = 4$. 1. Find the standard error of $\hat{\beta}_2$. 2. Construct a confidence interval for β_2 using a generic critical value $t^\star$. 3. For $x_0 = (1, 3)^\top$, compute the estimated variance of the fitted mean. 4. Write down the form of the prediction interval at x_0.

Proof-based

Show that if

$$\hat{\beta} \sim N_p!\left(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right),$$

then for any fixed vector a,

$$a^\top\hat{\beta} \sim N!\left(a^\top\beta, \sigma^2 a^\top(\mathbf{X}^\top\mathbf{X})^{-1}a\right).$$

16. Suggested Homework

Complete the following tasks: • derive the distribution of $\hat{\beta}$ under the normal linear model; • prove that $\hat{\sigma}^2 = \mathrm{SSE}/(n-p)$ is unbiased; • derive the t statistic for one coefficient; • construct a confidence interval for the mean response at a chosen covariate value; • construct a prediction interval for a future observation at the same covariate value; • fit a regression model in R and interpret all coefficient-level inferential output.

17. Summary

In this week, we moved from estimation to inference under the normal linear model.

We showed that

$$\hat{\beta} \sim N_p!\left(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right),$$

and that

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-p}, \qquad \hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-p}.$$

These results, together with independence between $\hat{\beta}$ and SSE, lead to exact t and F inference.

Next week, we will develop the ANOVA decomposition in regression and study the overall significance test, nested models, and extra sum of squares.

Appendix: Useful Distribution Facts

If

$$Z \sim N(0,1), \qquad U \sim \chi^2_\nu,$$

and Z and U are independent, then

$$\frac{Z}{\sqrt{U/\nu}} \sim t_\nu.$$

If

$$U_1 \sim \chi^2_r, \qquad U_2 \sim \chi^2_\nu,$$

and U_1 and U_2 are independent, then

$$\frac{(U_1/r)}{(U_2/\nu)} \sim F_{r,\nu}.$$

These are the basic building blocks for regression inference.

32 Week 4: ANOVA Decomposition, Overall F Test, and Nested Models

In this week, we study the ANOVA decomposition for linear regression, the overall significance test, and the comparison of nested models. These ideas connect the geometry of least squares with the inferential tools developed in previous weeks.

32.1 Learning Objectives

By the end of this week, students should be able to:

- define the total, regression, and error sums of squares;
- explain the ANOVA decomposition in regression with an intercept;
- interpret the degrees of freedom associated with SST, SSR, and SSE;
- perform the overall F test for regression;
- compare nested linear models using extra sums of squares;
- interpret ANOVA tables produced by statistical software.

32.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on analysis of variance in regression
  - sums of squares and decomposition of variability
  - tests for nested models
- Montgomery, Peck, and Vining:
  - sections on the ANOVA table
  - overall model significance
  - partial and sequential sums of squares

32.3 Review of the Linear Model

Recall the normal linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

When $\mathbf{X}$ has full column rank, the ordinary least squares estimator is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

and the fitted values are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

The residual vector is

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

From Week 2, we know that $\hat{\mathbf{Y}}$ and $\mathbf{e}$ are orthogonal. From Week 3, we know that this leads to useful distributional results for inference.

This week, we organize these ideas into the ANOVA framework.

32.4 Total, Explained, and Unexplained Variation

A central question in regression is:

How much of the variation in the response can be explained by the model?

To answer this, we decompose the total variation in $\mathbf{Y}$ into:

- variation explained by regression;
- variation left unexplained by the model.

When the model includes an intercept, this decomposition takes a particularly simple and important form.

32.5 Total Sum of Squares

Assume the model contains an intercept.

The total variation in the response is measured by the **total sum of squares**

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

where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

In vector form, if $\mathbf{1}$ denotes the vector of ones, then

$$\text{SST} = (\mathbf{Y} - \bar{Y}\mathbf{1})^\top (\mathbf{Y} - \bar{Y}\mathbf{1}).$$

This measures variation around the sample mean.

32.6 Error Sum of Squares

The error sum of squares is

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \mathbf{e}^\top \mathbf{e}.$$

This is the variation not explained by the fitted regression model.

It measures how far the observed responses are from the fitted values.

32.7 Regression Sum of Squares

The regression sum of squares is

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

This measures the part of the total variation explained by the regression model.

In vector form,

$$\text{SSR} = (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})^\top (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}).$$

32.8 The ANOVA Decomposition

When the model includes an intercept, we have the decomposition

$$\text{SST} = \text{SSR} + \text{SSE}.$$

This is one of the most important identities in regression.

It says that the total variation around the sample mean can be decomposed into:

- variation explained by the regression model;
- variation remaining in the residuals.

32.9 Why the Decomposition Holds

The key reason is orthogonality.

We can write

$$\mathbf{Y} - \bar{Y}\mathbf{1} = (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}) + (\mathbf{Y} - \hat{\mathbf{Y}}).$$

That is,

$$\mathbf{Y} - \bar{Y}\mathbf{1} = (\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}) + \mathbf{e}.$$

Because the model contains an intercept, the vector $\bar{Y}\mathbf{1}$ lies in the column space of $\mathbf{X}$. Hence both $\hat{\mathbf{Y}}$ and $\bar{Y}\mathbf{1}$ lie in the model space, so their difference also lies in the model space.

Since the residual vector $\mathbf{e}$ is orthogonal to the model space, we have

$$(\hat{\mathbf{Y}} - \bar{Y}\mathbf{1})^\top \mathbf{e} = 0.$$

Therefore, by the Pythagorean theorem,

$$\|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 = \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2 + \|\mathbf{e}\|^2,$$

which is exactly

$$\text{SST} = \text{SSR} + \text{SSE}.$$

32.10 Degrees of Freedom

The ANOVA decomposition is accompanied by a decomposition of degrees of freedom.

When the model includes an intercept and $\mathbf{X}$ has rank p, we have:

- total degrees of freedom: $n - 1$;
- regression degrees of freedom: $p - 1$;
- error degrees of freedom: $n - p$.

Thus,

$$n - 1 = (p - 1) + (n - p).$$

These match the sum of squares decomposition:

$$\text{SST} = \text{SSR} + \text{SSE}.$$

32.11 Mean Squares

To compare sums of squares on a common scale, we divide by their associated degrees of freedom.

The **mean square for regression** is

$$\text{MSR} = \frac{\text{SSR}}{p-1}.$$

The **mean square error** is

$$\text{MSE} = \frac{\text{SSE}}{n-p}.$$

From Week 3, we know that MSE is an unbiased estimator of σ^2.

32.12 The Overall F Test

A major inferential question is whether the regression model provides any explanatory power beyond the intercept-only model.

Suppose the model includes an intercept and $p-1$ additional predictors. The null hypothesis is

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_p = 0,$$

if the first coefficient corresponds to the intercept.

Equivalently, under H_0, the mean response does not depend on the predictors.

The alternative is that at least one non-intercept coefficient is nonzero.

32.13 Test Statistic

The overall F statistic is

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)}.$$

Under the null hypothesis,

$$F \sim F_{p-1,\, n-p}.$$

Large values of F provide evidence against H_0.

32.14 Interpretation of the Overall F Test

The numerator measures explained variation per regression degree of freedom.

The denominator measures unexplained variation per residual degree of freedom.

So the F statistic compares:

- how much signal the model explains;
- how much noise remains in the residuals.

If the predictors have no effect, then both quantities should be of similar size, and the ratio should not be unusually large.

If the predictors explain substantial variation, then the numerator should be much larger than the denominator.

32.15 Relationship to the Intercept-Only Model

The overall F test compares two models:

- the reduced model: intercept only;
- the full model: intercept plus predictors.

Thus the ANOVA decomposition provides the basis for formal model comparison.

This leads naturally to the idea of nested models.

32.16 Nested Models

Two models are **nested** if the reduced model is obtained by imposing constraints on the full model.

For example:

- reduced model:
$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i;$$
- full model:
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

The reduced model is nested within the full model because it is obtained by setting

$$\beta_2 = 0.$$

32.17 Extra Sum of Squares Principle

Suppose:

- the reduced model has error sum of squares SSE_R and rank p_R;
- the full model has error sum of squares SSE_F and rank p_F.

Since the full model has more flexibility, it cannot fit worse, so

$$\text{SSE}_F \leq \text{SSE}_R.$$

The quantity

$$\text{SSE}_R - \text{SSE}_F$$

measures the reduction in error due to adding the extra predictors.

This is called the **extra sum of squares** due to the added terms.

32.18 F Test for Nested Models

To test whether the extra predictors in the full model are needed, we use

$$F = \frac{(\text{SSE}_R - \text{SSE}_F)/(p_F - p_R)}{\text{SSE}_F/(n - p_F)}.$$

Under the null hypothesis that the additional parameters are unnecessary,

$$F \sim F_{p_F - p_R,\, n - p_F}.$$

Large values indicate that the full model provides a significantly better fit.

32.19 Connection with General Linear Hypotheses

This nested-model F test is equivalent to testing a general linear hypothesis of the form

$$H_0 : \mathbf{C}\beta = \mathbf{d}.$$

So the ANOVA comparison of nested models is another way of expressing the general F test from Week 3.

In practice, this is one of the most common uses of regression ANOVA tables.

32.20 Sequential and Partial Sums of Squares

In multiple regression, sums of squares can be defined in different ways depending on what is being adjusted for.

Two common ideas are:

- **sequential sums of squares**: terms are added in a specified order;
- **partial sums of squares**: each term is tested after adjusting for the others.

This distinction becomes important when predictors are correlated.

In this course, the main conceptual priority is to understand the extra sum of squares principle. Details of Type I, Type II, and Type III sums of squares can be introduced later if needed.

32.21 Coefficient of Determination

The coefficient of determination is

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

It measures the proportion of total variation explained by the regression model.

Its values lie between 0 and 1.

32.22 Interpretation of R Squared

- If R^2 is close to 1, the model explains a large proportion of the variation in the response.
- If R^2 is close to 0, the model explains little of the variation.

However, R^2 alone does not guarantee that the model is appropriate. A high R^2 does not ensure that assumptions are satisfied, and a low R^2 does not necessarily imply the model is useless.

32.23 Adjusted R Squared

Because R^2 never decreases when additional predictors are added, it can overstate improvement.

A commonly used adjustment is

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}.$$

Adjusted R^2 penalizes the inclusion of unnecessary predictors.

32.24 Worked Example by Hand

Consider the data

x_i	0	1	2	3
y_i	1	3	3	5

We already found that the fitted regression line is

$$\hat{Y} = 1.2 + 1.2x.$$

The observed response vector is

$$\mathbf{Y} = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 5 \end{bmatrix},$$

and the fitted values are

$$\hat{\mathbf{Y}} = \begin{bmatrix} 1.2 \\ 2.4 \\ 3.6 \\ 4.8 \end{bmatrix}.$$

The sample mean is

$$\bar{Y} = 3.$$

32.24.1 Compute SST

$$\text{SST} = (1-3)^2 + (3-3)^2 + (3-3)^2 + (5-3)^2 = 4 + 0 + 0 + 4 = 8.$$

32.24.2 Compute SSE

The residuals are

$$\mathbf{e} = \begin{bmatrix} -0.2 \\ 0.6 \\ -0.6 \\ 0.2 \end{bmatrix},$$

so

$$\text{SSE} = (-0.2)^2 + 0.6^2 + (-0.6)^2 + 0.2^2 = 0.8.$$

32.24.3 Compute SSR

Using $\text{SSR} = \text{SST} - \text{SSE}$,

$$\text{SSR} = 8 - 0.8 = 7.2.$$

32.24.4 Check the decomposition

$$8 = 7.2 + 0.8.$$

32.24.5 Degrees of freedom

Here $n = 4$ and $p = 2$, so:

- total df: $4 - 1 = 3$;
- regression df: $2 - 1 = 1$;
- error df: $4 - 2 = 2$.

32.24.6 Mean squares

$$\text{MSR} = \frac{7.2}{1} = 7.2, \qquad \text{MSE} = \frac{0.8}{2} = 0.4.$$

32.24.7 F statistic

$$F = \frac{7.2}{0.4} = 18.$$

This is the overall significance test for the regression.

32.25 ANOVA Table Structure

A standard regression ANOVA table has the following structure:

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	$p-1$	MSR	MSR/MSE
Error	SSE	$n-p$	MSE	
Total	SST	$n-1$		

Students should learn to move fluently between:

- formulas;
- geometric interpretation;
- software output.

32.26 R Demonstration

32.27 Fit a simple regression model

```r
x <- c(0, 1, 2, 3)
y <- c(1, 3, 3, 5)

fit <- lm(y ~ x)
summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
```

```
   1    2    3    4
-0.2  0.6 -0.6  0.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2000     0.5292   2.268   0.1515
x             1.2000     0.2828   4.243   0.0513 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6325 on 2 degrees of freedom
Multiple R-squared:    0.9, Adjusted R-squared:   0.85
F-statistic:    18 on 1 and 2 DF,  p-value: 0.05132
```

32.28 Obtain the ANOVA table

```r
anova(fit)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1    7.2     7.2      18 0.05132 .
Residuals  2    0.8     0.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

32.29 Verify sums of squares manually

```r
ybar <- mean(y)
yhat <- fitted(fit)
e <- resid(fit)

SST <- sum((y - ybar)^2)
SSE <- sum(e^2)
SSR <- sum((yhat - ybar)^2)

c(SST = SST, SSR = SSR, SSE = SSE)
```

```
SST SSR SSE
8.0 7.2 0.8
```

```
SST - SSR - SSE
```

```
[1] 2.220446e-16
```

32.30 Compute R squared manually

```
R2 <- SSR / SST
R2
```

```
[1] 0.9
```

```
summary(fit)$r.squared
```

```
[1] 0.9
```

```
summary(fit)$adj.r.squared
```

```
[1] 0.85
```

32.31 Compare nested models

```
dat <- data.frame(
  y = c(4, 5, 7, 10, 8, 12, 13, 14),
  x1 = c(1, 2, 3, 4, 5, 6, 7, 8),
  x2 = c(2, 1, 3, 2, 5, 4, 6, 5)
)

fit_reduced <- lm(y ~ x1, data = dat)
fit_full <- lm(y ~ x1 + x2, data = dat)

anova(fit_reduced, fit_full)
```

```
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      6 6.8214
2      5 4.6613  1    2.1601 2.3171 0.1884
```

32.32 Inspect the two fitted models

```
summary(fit_reduced)
```

```

Call:
lm(formula = y ~ x1, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85714 -0.30357  0.03571  0.33036  1.60714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5357     0.8308   3.052 0.022453 *
x1            1.4643     0.1645   8.900 0.000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.066 on 6 degrees of freedom
Multiple R-squared:  0.9296,    Adjusted R-squared:  0.9178
F-statistic: 79.21 on 1 and 6 DF,  p-value: 0.0001121
```

```
summary(fit_full)
```

```

Call:
lm(formula = y ~ x1 + x2, data = dat)

Residuals:
      1       2       3       4       5       6       7       8
```

```
 0.4032 -1.0403  0.3306  0.8871 -1.1371  0.4194  0.7903 -0.6532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9677     0.8041   3.691  0.01413 *
x1            1.8387     0.2876   6.394  0.00139 **
x2           -0.6048     0.3973  -1.522  0.18845
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9655 on 5 degrees of freedom
Multiple R-squared:  0.9519,    Adjusted R-squared:  0.9326
F-statistic: 49.46 on 2 and 5 DF,  p-value: 0.0005079
```

32.33 Interpretation of Software Output

For a fitted model in R:

- `anova(fit)` gives the ANOVA decomposition for a single model;
- `anova(fit_reduced, fit_full)` compares nested models;
- `summary(fit)` reports the overall F statistic, R^2, and adjusted R^2.

Students should understand that these outputs are not separate topics. They are all built from the same least squares geometry and distribution theory.

32.34 In-Class Discussion Questions

1. Why does the ANOVA decomposition require an intercept for the usual SST = SSR + SSE identity?
2. Why must $\text{SSE}_F \leq \text{SSE}_R$ for nested models?
3. What does the overall F test tell us that individual t tests do not?
4. Why can R^2 be misleading if used alone?

32.35 Practice Problems

32.36 Conceptual

1. Explain the meaning of SST, SSR, and SSE in words.

2. Explain why the regression degrees of freedom are $p-1$ when the model includes an intercept.
3. Explain the difference between the overall F test and a test for a single coefficient.

32.37 Computational

Suppose a regression model with intercept has:

- $n = 20$,
- $p = 4$,
- $\text{SST} = 100$,
- $\text{SSE} = 40$.

Compute:

1. SSR,
2. the degrees of freedom for regression and error,
3. MSR,
4. MSE,
5. the overall F statistic,
6. R^2.

32.38 Nested Model Problem

A reduced model has

$$\text{SSE}_R = 120$$

with $p_R = 3$, and a full model has

$$\text{SSE}_F = 90$$

with $p_F = 5$.

If $n = 30$, compute the nested-model F statistic.

32.39 Suggested Homework

Complete the following tasks:

- prove the decomposition $\text{SST} = \text{SSR} + \text{SSE}$ when the model includes an intercept;
- derive the overall F statistic from the ANOVA decomposition;
- fit a regression model in R and reproduce the ANOVA table by hand;
- compare two nested models using an extra sum of squares test;
- interpret both R^2 and adjusted R^2 for a chosen dataset.

32.40 Summary

In this week, we developed the ANOVA framework for linear regression.

We defined:

$$\text{SST}, \qquad \text{SSR}, \qquad \text{SSE},$$

and showed that, with an intercept,

$$\text{SST} = \text{SSR} + \text{SSE}.$$

This decomposition led to:

- the ANOVA table;
- the overall F test for regression;
- the comparison of nested models through extra sums of squares;
- the interpretation of R^2 and adjusted R^2.

Next week, a natural continuation is to study multiple regression in greater depth, including interpretation of partial regression coefficients and multicollinearity, or to move into matrix-based general linear hypotheses and estimability, depending on the course emphasis.

32.41 Appendix: Compact Formula Summary

With an intercept in the model,

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

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

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

and

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}.$$

Also,

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{p-1}, \qquad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-p}, \qquad F = \frac{\mathrm{MSR}}{\mathrm{MSE}},$$

and

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

33 Week 5: Multiple Regression, Partial Effects, and Categorical Predictors

In this week, we move from simple regression ideas to multiple regression. We study how regression coefficients are interpreted when several predictors are included in the model, how categorical predictors enter through indicator variables, and how interactions change the meaning of coefficients. The main goal is to help students read, build, and interpret regression models in realistic settings.

33.1 Learning Objectives

By the end of this week, students should be able to:

- interpret coefficients in a multiple regression model;
- explain the meaning of a partial regression coefficient;
- distinguish between marginal association and adjusted association;
- incorporate categorical predictors using indicator variables;
- interpret regression models with interactions;
- use software output to explain fitted multiple regression models.

33.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on multiple linear regression
  - interpretation of regression coefficients
  - indicator variables and model formulation
- Montgomery, Peck, and Vining:
  - sections on multiple regression
  - qualitative predictors
  - interaction terms and interpretation

33.3 Review of the Regression Framework

Recall the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

When the design matrix $\mathbf{X}$ has full column rank, the ordinary least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

In earlier weeks, we focused on estimation, inference, ANOVA decomposition, and the comparison of nested models. In this week, the emphasis shifts toward interpretation and modelling structure.

33.4 From Simple Regression to Multiple Regression

In simple linear regression, we write

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Here, the slope β_1 measures the expected change in the response associated with a one-unit increase in x.

In multiple regression, the model becomes

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

Now each coefficient must be interpreted while holding the other predictors fixed.

This is the key conceptual shift.

33.5 Why Multiple Regression Matters

Multiple regression is important because real data usually involve several explanatory variables.

Reasons for including multiple predictors include:

- improving prediction;
- adjusting for confounding variables;
- estimating the effect of one variable while controlling for others;
- allowing more realistic scientific interpretation.

A coefficient in multiple regression is therefore usually an adjusted effect, not a purely marginal one.

33.6 Interpreting the Intercept

In the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

the intercept β_0 is the expected value of the response when all predictors equal zero:

$$\mathbb{E}[Y_i \mid x_{i1}, \ldots, x_{ip}] = \beta_0 \quad \text{when } x_{i1} = \cdots = x_{ip} = 0.$$

This interpretation may or may not be scientifically meaningful.

Sometimes zero is a natural baseline. Sometimes it is outside the observed range, in which case the intercept is mainly a mathematical anchor for the model.

33.7 Interpreting Partial Regression Coefficients

Consider the model

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

Then:

- β_1 is the expected change in Y associated with a one-unit increase in x_1, holding x_2 fixed;
- β_2 is the expected change in Y associated with a one-unit increase in x_2, holding x_1 fixed.

This is called a **partial effect** or **adjusted effect**.

The phrase "holding other variables fixed" is essential and should always be stated clearly.

33.8 Marginal Association Versus Adjusted Association

Suppose x_1 and x_2 are correlated.

Then the relationship between Y and x_1 in a simple regression of Y on x_1 alone may differ from the coefficient of x_1 in a multiple regression including both x_1 and x_2.

This happens because:

- the simple regression coefficient describes a marginal association;
- the multiple regression coefficient describes an adjusted association.

These can differ substantially when predictors are related to each other.

33.9 Example of Adjusted Interpretation

Suppose we fit

$$\hat{Y} = 12.5 + 0.8\,x_1 - 1.2\,x_2.$$

Then:

- for each one-unit increase in x_1, the fitted mean response increases by 0.8 units, holding x_2 fixed;
- for each one-unit increase in x_2, the fitted mean response decreases by 1.2 units, holding x_1 fixed.

This interpretation is valid only within the modelling assumptions and over the range of data where the model is reasonable.

33.10 Matrix View of Multiple Regression

In multiple regression, the design matrix has the form

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

Each column corresponds to a predictor or model term.

This allows the same least squares and inference framework to handle:

- continuous predictors;
- categorical predictors represented by indicator variables;
- interactions;
- polynomial terms.

So the general linear model framework is highly flexible.

33.11 Categorical Predictors and Indicator Variables

Many real predictors are categorical rather than numeric.

Examples include:

- treatment group;
- sex;
- school type;
- region;
- machine type.

These enter a regression model through indicator variables, also called dummy variables.

33.12 Binary Predictor

Suppose z_i is a binary variable:

$$z_i = \begin{cases} 1, & \text{if observation } i \text{ is in group A}, \\ 0, & \text{if observation } i \text{ is in group B}. \end{cases}$$

Then the model

$$Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$$

has the following interpretation:

- when $z_i = 0$, the expected response is β_0;
- when $z_i = 1$, the expected response is $\beta_0 + \beta_1$.

Thus, β_1 measures the difference in group means.

33.13 More Than Two Categories

Suppose a categorical predictor has three levels:

- A,
- B,
- C.

We usually represent it with two indicator variables, for example:

$$z_{1i} = \begin{cases} 1, & \text{if level B,} \\ 0, & \text{otherwise,} \end{cases} \qquad z_{2i} = \begin{cases} 1, & \text{if level C,} \\ 0, & \text{otherwise.} \end{cases}$$

Then level A is the reference group, and the model is

$$Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i.$$

Interpretation:

- level A mean: β_0;
- level B mean: $\beta_0 + \beta_1$;
- level C mean: $\beta_0 + \beta_2$.

So coefficients are interpreted relative to the reference category.

33.14 Why We Do Not Include All Indicators with an Intercept

If a categorical variable has k levels, then with an intercept we include only $k-1$ indicator variables.

If we include all k indicators and also include an intercept, then the columns of the design matrix become linearly dependent. This causes rank deficiency.

This is sometimes called the dummy variable trap.

33.15 Continuous and Categorical Predictors Together

Regression models often mix continuous and categorical predictors.

For example,

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

may contain:

- a continuous predictor x_i;
- a binary indicator z_i.

Interpretation:

- β_1 is the slope in x holding group fixed;
- β_2 is the group difference holding x fixed.

This is one of the simplest forms of ANCOVA-style modelling.

33.16 Interaction Between Continuous Predictors

An interaction allows the effect of one predictor to depend on the value of another predictor.

For two continuous predictors, we may write

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i.$$

Now the effect of x_1 is no longer constant. Instead,

$$\frac{\partial \mathbb{E}[Y_i \mid x_{i1}, x_{i2}]}{\partial x_{i1}} = \beta_1 + \beta_3 x_{i2}.$$

So the slope for x_1 depends on the value of x_2.

Likewise, the slope for x_2 depends on x_1.

33.17 Interaction Between a Continuous and a Binary Predictor

Suppose we fit

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i,$$

where $z_i \in \{0, 1\}$.

Then:

- when $z_i = 0$,
$$\mathbb{E}[Y_i \mid x_i, z_i = 0] = \beta_0 + \beta_1 x_i;$$
- when $z_i = 1$,
$$\mathbb{E}[Y_i \mid x_i, z_i = 1] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i.$$

So:

- β_2 changes the intercept;
- β_3 changes the slope.

This allows the two groups to have different regression lines.

33.18 Main Effects in the Presence of Interaction

When an interaction is present, the meaning of the main effects changes.

For example, in

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i,$$

the coefficient β_1 is the slope for x only when $z = 0$, the reference group.

Similarly, β_2 is the group difference only when $x = 0$.

So students should be careful not to interpret main effects in isolation when interactions are included.

33.19 Centering Predictors for Interpretation

Sometimes the value zero is not meaningful for a continuous predictor.

In that case, it can be helpful to centre the predictor:

$$x_i^* = x_i - \bar{x}.$$

Then the intercept becomes the expected response at the average value of x, which may be much easier to interpret.

Centering can also improve interpretability in models with interactions.

33.20 Comparing Models With and Without Interaction

To assess whether an interaction is needed, we can compare nested models.

For example:

- reduced model:
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i;$$
- full model:
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i.$$

The reduced model is nested within the full model by setting

$$\beta_3 = 0.$$

So the interaction can be tested by a standard extra sum of squares F test.

33.21 Collinearity and Interpretation

In multiple regression, predictors may be strongly related to each other.

When this happens:

- coefficient estimates can become unstable;
- standard errors can become large;
- interpretation becomes more delicate.

Even when the overall model seems useful, individual coefficients may be hard to estimate precisely.

This issue is called **multicollinearity**.

We will study diagnostics for it more formally later, but students should already know that "holding other variables fixed" may become practically difficult if predictors tend to move together.

33.22 Multiple Regression as Conditional Mean Modelling

A good way to summarize multiple regression is:

$$\mathbb{E}[Y \mid X_1, \dots, X_p]$$

is being modelled as a linear function of predictors and model terms.

This viewpoint helps unify:

- continuous predictors;
- factors;
- interactions;
- transformed predictors.

It also helps students distinguish between the response itself and its conditional mean.

33.23 Worked Example With a Continuous and a Binary Predictor

Suppose we observe a response Y, a study-hours variable x, and a tutoring indicator z, where

- $z = 0$ means no tutoring;
- $z = 1$ means tutoring.

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i.$$

Suppose the fitted equation is

$$\hat{Y} = 50 + 3x + 8z.$$

Then:

- among students with the same tutoring status, one extra hour of study is associated with an increase of 3 points in the fitted mean score;
- among students with the same number of study hours, the tutoring group has a fitted mean score 8 points higher than the non-tutoring group.

If we add an interaction and obtain

$$\hat{Y} = 48 + 4x + 10z - 1.5xz,$$

then:

- for students without tutoring, the slope in study hours is 4;
- for students with tutoring, the slope in study hours is $4 - 1.5 = 2.5$.

Thus, the effect of study time depends on tutoring status.

33.24 R Demonstration With Multiple Regression

33.25 Fit a model with two continuous predictors

```
dat1 <- data.frame(
  y = c(12, 15, 14, 18, 20, 19, 23, 25),
  x1 = c(1, 2, 2, 3, 4, 4, 5, 6),
  x2 = c(5, 4, 6, 4, 3, 5, 2, 1)
)

fit1 <- lm(y ~ x1 + x2, data = dat1)
summary(fit1)
```

```
Call:
lm(formula = y ~ x1 + x2, data = dat1)

Residuals:
         1          2          3          4          5          6          7
-3.158e-01 -2.632e-02 -1.316e-01  7.105e-01  4.842e-16 -1.053e-01  2.895e-01
         8
-4.211e-01
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2895     1.1598  10.596 0.000129 ***
x1            2.2632     0.1681  13.464 4.05e-05 ***
x2           -0.4474     0.1697  -2.636 0.046186 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.423 on 5 degrees of freedom
Multiple R-squared:  0.9936,    Adjusted R-squared:  0.991
F-statistic: 387.3 on 2 and 5 DF,  p-value: 3.295e-06
```

33.26 Compare with a simple regression

```r
fit1_simple <- lm(y ~ x1, data = dat1)
summary(fit1_simple)
```

```
Call:
lm(formula = y ~ x1, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89308 -0.27201  0.05031  0.39308  0.73585

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3774     0.4988   18.80 1.46e-06 ***
x1            2.6289     0.1339   19.63 1.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.597 on 6 degrees of freedom
Multiple R-squared:  0.9847,    Adjusted R-squared:  0.9821
F-statistic: 385.4 on 1 and 6 DF,  p-value: 1.132e-06
```

```r
summary(fit1)
```

```
Call:
lm(formula = y ~ x1 + x2, data = dat1)

Residuals:
         1          2          3          4          5          6          7
-3.158e-01 -2.632e-02 -1.316e-01  7.105e-01  4.842e-16 -1.053e-01  2.895e-01
         8
-4.211e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2895     1.1598  10.596 0.000129 ***
x1            2.2632     0.1681  13.464 4.05e-05 ***
x2           -0.4474     0.1697  -2.636 0.046186 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.423 on 5 degrees of freedom
Multiple R-squared:  0.9936,    Adjusted R-squared:  0.991
F-statistic: 387.3 on 2 and 5 DF,  p-value: 3.295e-06
```

33.27 Fit a model with a categorical predictor

```r
dat2 <- data.frame(
  y = c(60, 62, 58, 71, 73, 70, 66, 68),
  hours = c(4, 5, 3, 4, 5, 6, 4, 5),
  group = factor(c("A", "A", "A", "B", "B", "B", "A", "B"))
)

fit2 <- lm(y ~ hours + group, data = dat2)
summary(fit2)
```

```
Call:
lm(formula = y ~ hours + group, data = dat2)

Residuals:
    1     2     3     4     5     6     7     8
-1.50 -0.25 -2.75  1.25  2.50 -1.25  4.50 -2.50
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   58.500      6.236   9.381 0.000232 ***
hours          0.750      1.512   0.496 0.641001
groupB         8.250      2.620   3.149 0.025399 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.025 on 5 degrees of freedom
Multiple R-squared:  0.7821,    Adjusted R-squared:  0.695
F-statistic: 8.975 on 2 and 5 DF,  p-value: 0.02215
```

```
model.matrix(fit2)
```

```
  (Intercept) hours groupB
1           1     4      0
2           1     5      0
3           1     3      0
4           1     4      1
5           1     5      1
6           1     6      1
7           1     4      0
8           1     5      1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

33.28 Fit a model with interaction

```
fit3 <- lm(y ~ hours * group, data = dat2)
summary(fit3)
```

```
Call:
lm(formula = y ~ hours * group, data = dat2)

Residuals:
```

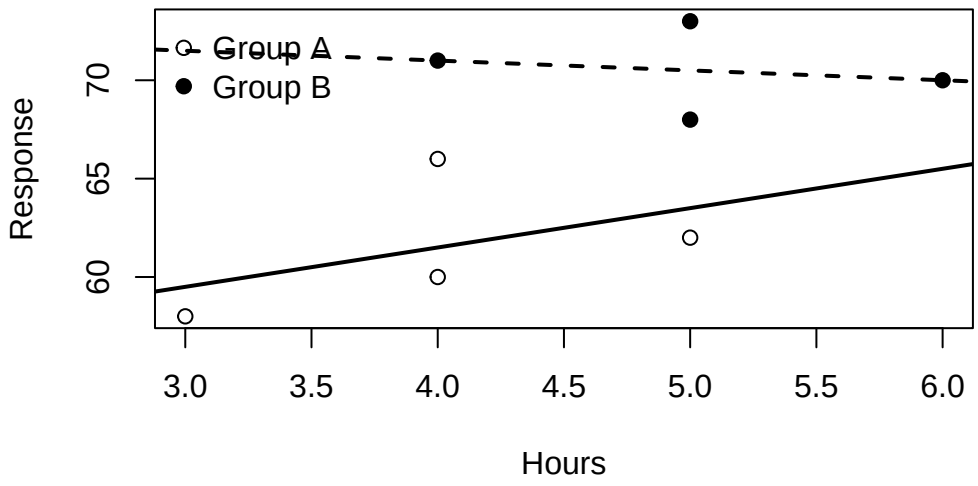
```
         1          2          3          4          5          6          7
-1.500e+00 -1.500e+00 -1.500e+00 -2.216e-15  2.500e+00 -7.730e-16  4.500e+00
         8
-2.500e+00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    53.500      9.026   5.927  0.00406 **
hours           2.000      2.222   0.900  0.41897
groupB         19.500     14.401   1.354  0.24715
hours:groupB   -2.500      3.142  -0.796  0.47083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.142 on 4 degrees of freedom
Multiple R-squared:  0.8119,    Adjusted R-squared:  0.6708
F-statistic: 5.755 on 3 and 4 DF,  p-value: 0.06202
```

```r
anova(fit2, fit3)
```

```
Analysis of Variance Table

Model 1: y ~ hours + group
Model 2: y ~ hours * group
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1      5 45.75
2      4 39.50  1      6.25 0.6329 0.4708
```

33.29 Plot group-specific regression lines

```r
plot(dat2$hours, dat2$y,
     pch = ifelse(dat2$group == "A", 1, 19),
     xlab = "Hours",
     ylab = "Response")
abline(a = coef(fit3)[1], b = coef(fit3)[2], lwd = 2)
abline(a = coef(fit3)[1] + coef(fit3)[3],
       b = coef(fit3)[2] + coef(fit3)[4],
       lwd = 2, lty = 2)
legend("topleft", legend = c("Group A", "Group B"),
       pch = c(1, 19), bty = "n")
```



33.30 Interpreting Software Output

In `summary(lm(...))`, each coefficient estimate answers a question about the conditional mean given the model terms included.

Students should always ask:

- what variables are being held fixed;
- what is the reference category;
- whether an interaction changes the meaning of the main effects;
- whether zero is a meaningful baseline for interpretation.

These questions matter more than memorizing formulas.

33.31 In-Class Discussion Questions

1. Why can the coefficient of a predictor change when a second predictor is added to the model?
2. Why do we need a reference category for categorical predictors?
3. How does an interaction change the interpretation of a main effect?
4. When might centring a predictor improve interpretation?

33.32 Practice Problems

33.33 Conceptual

1. Explain the meaning of a partial regression coefficient in your own words.
2. Explain the difference between a marginal association and an adjusted association.
3. Explain why a model with an interaction requires more careful interpretation than an additive model.

33.34 Computational

Suppose the fitted model is

$$\hat{Y} = 10 + 2x_1 - 3x_2.$$

1. Interpret the coefficient of x_1.
2. Interpret the coefficient of x_2.
3. Compute the fitted mean response when $x_1 = 4$ and $x_2 = 1$.

Now suppose the fitted model is

$$\hat{Y} = 20 + 5x + 7z - 2xz,$$

where z is binary.

1. Write the fitted mean function when $z = 0$.
2. Write the fitted mean function when $z = 1$.
3. Interpret the interaction coefficient.

33.35 Indicator Variable Problem

A factor has four levels: A, B, C, and D.

1. How many indicator variables are needed if the model includes an intercept?
2. If A is the reference group, write a regression model using indicators for B, C, and D.
3. State the expected response for each group.

33.36 Suggested Homework

Complete the following tasks:

- fit a multiple regression model with at least two continuous predictors and interpret all coefficients;
- fit a model with one continuous predictor and one categorical predictor, then identify the reference category and explain all coefficients;
- add an interaction term and compare the additive and interaction models;
- use `model.matrix()` in R to inspect the design matrix for a model with factors;
- write a short explanation of why coefficient interpretation changes when additional predictors are added.

33.37 Summary

In this week, we developed the interpretation of multiple regression models.

We emphasized that:

- regression coefficients in multiple regression are partial effects;
- categorical predictors are incorporated through indicator variables;
- interactions allow the effect of one predictor to depend on another;
- the meaning of a coefficient depends on the full model specification.

These ideas are essential for moving from formal least squares theory to practical statistical modelling.

Next week, a natural continuation is to study multicollinearity, variable selection, and model building, or to move into diagnostics and residual analysis, depending on the course emphasis.

33.38 Appendix: Compact Interpretation Guide

For the additive model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

- β_0: expected response when $x_1 = x_2 = 0$;
- β_1: expected change in response for a one-unit increase in x_1, holding x_2 fixed;
- β_2: expected change in response for a one-unit increase in x_2, holding x_1 fixed.

For the model with a binary predictor

$$Y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon,$$

- β_0: mean for the reference group when $x = 0$;
- β_1: slope in x for fixed group;
- β_2: group difference for fixed x.

For the interaction model

$$Y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon,$$

- slope in x when $z = 0$: β_1;
- slope in x when $z = 1$: $\beta_1 + \beta_3$;
- group difference when $x = 0$: β_2.

34 Week 6: Residual Analysis, Diagnostics, and Model Adequacy

In this week, we study how to assess whether a fitted linear regression model is adequate. After learning estimation, inference, ANOVA, and interpretation, we now turn to model checking. The main tools are residuals, diagnostic plots, and numerical measures that help identify nonlinearity, unequal variance, outliers, and influential observations.

34.1 Learning Objectives

By the end of this week, students should be able to:

- explain why model diagnostics are necessary after fitting a regression model;
- define raw, standardized, and studentized residuals;
- interpret residual plots for common types of model failure;
- distinguish between outliers in the response direction and influential observations in the design space;
- use leverage, Cook's distance, and related diagnostics;
- assess model adequacy using graphical and numerical tools.

34.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on residual analysis
  - diagnostics for regression models
  - influence and unusual observations
- Montgomery, Peck, and Vining:
  - sections on residuals
  - diagnostic plots
  - outliers, leverage, and influence

34.3 Why Diagnostics Matter

A fitted regression model may look statistically significant and still be inappropriate.

For example:

- the relationship may not be linear;
- the error variance may not be constant;
- the error distribution may be strongly non-normal;
- a small number of unusual observations may dominate the fit.

So regression analysis does not end when we obtain estimates and p-values. We must also ask whether the model assumptions are reasonable and whether the fit is being driven by problematic observations.

34.4 Review of the Linear Model Assumptions

Recall the classical linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with assumptions such as

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

Under the normal linear model, we also assume

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

These assumptions support:

- unbiasedness of OLS;
- standard error formulas;
- t and F inference;
- prediction intervals.

Diagnostics help us investigate whether these assumptions seem plausible for the observed data.

34.5 Residuals

The fitted values are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta},$$

and the residual vector is

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

The ith residual is

$$e_i = Y_i - \hat{Y}_i.$$

A residual measures how far an observed response is from the fitted value at that observation.

Residuals are the basic raw material of regression diagnostics.

34.6 Important Warning About Residuals

Residuals are not the same as the true errors.

The true model error is

$$\varepsilon_i = Y_i - \mathbb{E}[Y_i \mid \mathbf{x}_i],$$

whereas the residual is

$$e_i = Y_i - \hat{Y}_i.$$

Residuals are observable, but errors are not. Diagnostics therefore use residuals as proxies for model errors.

34.7 Properties of Residuals

From least squares geometry, we know that

$$\mathbf{X}^\top \mathbf{e} = \mathbf{0}.$$

If the model contains an intercept, then

$$\sum_{i=1}^{n} e_i = 0.$$

Thus residuals are constrained and are not independent. This is one reason why it is useful to standardize them before interpretation.

34.8 Residual Variance and Leverage

Recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top.$$

Its diagonal entries

$$h_{ii}$$

are called **leverages**.

Under the standard model,

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii}).$$

So residuals do not all have the same variance. Observations with high leverage tend to have smaller residual variance.

This is why raw residuals alone can be misleading.

34.9 Standardized Residuals

A common adjustment is to divide each residual by its estimated standard deviation.

The **standardized residual** is

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

This puts residuals on a comparable scale across observations.

Large absolute values of r_i may indicate unusual observations in the response direction.

34.10 Studentized Residuals

A more refined version is the **studentized residual**.

One version uses the variance estimate computed from the full model. Another version uses the variance estimate obtained after deleting the ith observation.

The externally studentized residual is often written as

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}},$$

where $\hat{\sigma}^2_{(i)}$ is the residual variance estimate from the model fit without observation i.

These are especially useful for outlier detection.

34.11 Fitted Values Versus Residuals Plot

One of the most important diagnostic plots is the plot of residuals versus fitted values.

This plot helps detect:

- nonlinearity;
- unequal variance;
- outliers;
- missing structure.

A good residual plot usually shows points randomly scattered around zero with roughly constant spread.

Patterns are warning signs.

34.12 Interpreting Common Patterns

If the residual plot shows a curved pattern, that suggests the mean function may not be adequately linear.

If the spread of residuals increases or decreases with the fitted values, that suggests heteroscedasticity, meaning nonconstant error variance.

If a few points are isolated far from the rest, they may be outliers or influential observations.

Thus the residual plot is often the first and most important diagnostic tool.

34.13 Residuals Versus Individual Predictors

It is often helpful to plot residuals against each predictor separately.

A residual-versus-predictor plot can reveal:

- nonlinearity in that predictor;
- different spread across ranges of the predictor;
- group-specific patterns;
- possible interactions not included in the model.

These plots are often more informative than a single omnibus diagnostic.

34.14 Normal Q-Q Plot

The normal Q-Q plot is used to assess whether the residual distribution is approximately normal.

If the normality assumption is reasonable, the residual points should fall approximately along a straight line.

Departures from linearity suggest:

- skewness;
- heavy tails;
- light tails;
- extreme outliers.

Normality matters most when sample size is small and exact t and F inference is important.

34.15 Histogram of Residuals

A histogram of residuals can also be useful, though it is usually less informative than a Q-Q plot.

It may reveal:

- skewness;
- multimodality;
- heavy tails;
- extreme asymmetry.

However, the histogram depends strongly on bin choices, so it is usually used as a supplementary plot rather than the main normality diagnostic.

34.16 Scale-Location Plot

A scale-location plot often displays

$$\sqrt{|r_i|}$$

against fitted values.

This is another way to assess whether the variance is approximately constant across the fitted range.

A roughly horizontal band is desirable. A systematic increase or decrease suggests heteroscedasticity.

34.17 Outliers

An **outlier** is an observation whose response value is unusual relative to the fitted model.

In regression, an observation can be outlying in the response direction even if its predictor values are not unusual.

Large residuals or large studentized residuals often indicate potential outliers.

However, not every outlier is influential.

34.18 Leverage

Leverage measures how unusual an observation is in the predictor space.

The leverage of observation i is

$$h_{ii}.$$

High leverage points are far from the center of the predictor cloud in a geometric sense.

These points have greater potential to affect the fitted regression line or plane.

A rough rule of thumb is that leverage values substantially larger than

$$\frac{2p}{n} \quad \text{or sometimes} \quad \frac{3p}{n}$$

may deserve attention, where p is the number of parameters including the intercept.

These are only rough guidelines, not formal cutoffs.

34.19 Outlier Versus High-Leverage Point

It is important to distinguish between:

- a point with a large residual;
- a point with high leverage.

A point can have:

- low leverage and large residual;
- high leverage and small residual;
- both high leverage and large residual.

The last case is often the most concerning because such a point may strongly influence the fitted model.

34.20 Influence

An observation is **influential** if removing it changes the fitted model noticeably.

Influence depends on both:

- how unusual the response is;
- how unusual the predictor values are.

So influential points often combine high leverage with a sizable residual.

Influence is not the same as being an outlier.

34.21 Cook's Distance

One of the most widely used influence measures is **Cook's distance**.

Cook's distance for observation i measures how much the fitted values change when observation i is removed.

A common formula is

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}.$$

Large values of D_i indicate potentially influential observations.

Plots of Cook's distance help identify observations that deserve closer inspection.

34.22 DFFITS and DFBETAS

Other influence measures include:

- **DFFITS**, which measures the effect of deleting an observation on its fitted value;
- **DFBETAS**, which measure the effect of deleting an observation on each estimated coefficient.

These are useful when we want to know not only whether a point is influential, but how it changes the model.

For an introductory treatment, Cook's distance and leverage usually provide a strong starting point.

34.23 Added-Variable and Partial Residual Plots

When there are several predictors, ordinary residual plots may not fully reveal whether one variable needs a nonlinear term or whether its effect remains after adjustment.

Useful advanced plots include:

- **added-variable plots**, which assess the contribution of one predictor after adjusting for others;
- **partial residual plots**, which help visualize possible nonlinearity for a specific predictor.

These are especially useful in multiple regression, though they are often introduced after students become comfortable with basic residual plots.

34.24 Diagnosing Nonlinearity

If the mean function is nonlinear but we fit a linear model, residual plots may show curvature.

Possible remedies include:

- adding polynomial terms;
- applying transformations;
- including interactions;
- using a different modelling framework.

Diagnostics should not be viewed only as fault-finding. They also guide model improvement.

34.25 Diagnosing Heteroscedasticity

If the error variance is not constant, residual plots may show a funnel shape or other changing spread.

Possible remedies include:

- transforming the response;
- weighted least squares;
- modelling the variance structure explicitly;
- using heteroscedasticity-robust standard errors in some contexts.

At this stage, the main goal is to recognize the pattern and understand its consequences.

34.26 Diagnosing Non-Normality

If residuals are strongly non-normal, this may affect exact small-sample inference.

Possible causes include:

- skewed responses;
- heavy-tailed errors;
- outliers;
- omitted structure.

Possible remedies include:

- transformation;
- alternative modelling assumptions;
- robust procedures;
- careful interpretation if sample size is large and inference is approximately stable.

34.27 Diagnostics Are Contextual

There is no single diagnostic that automatically declares a model valid or invalid.

Instead, diagnostics require judgment.

Students should ask:

- Is the pattern strong or mild?
- Is it scientifically meaningful?
- Does one point dominate the fit?
- Would conclusions change if the model were modified?
- Is the issue important for the goal of the analysis: explanation, prediction, or inference?

34.28 Worked Example With an Outlying Observation

Consider the data

$$x = (1, 2, 3, 4, 5, 6, 7, 8),$$

and

$$y = (2, 4, 5, 8, 10, 11, 13, 25).$$

The final observation may look unusual because the response jumps upward relative to the earlier trend.

If we fit a simple linear regression, we should examine:

- the scatterplot with fitted line;
- residuals versus fitted values;
- studentized residuals;
- leverage;
- Cook's distance.

This is a good example for discussing the difference between an outlier and an influential point.

34.29 R Demonstration With Basic Diagnostic Plots

34.30 Fit a simple model

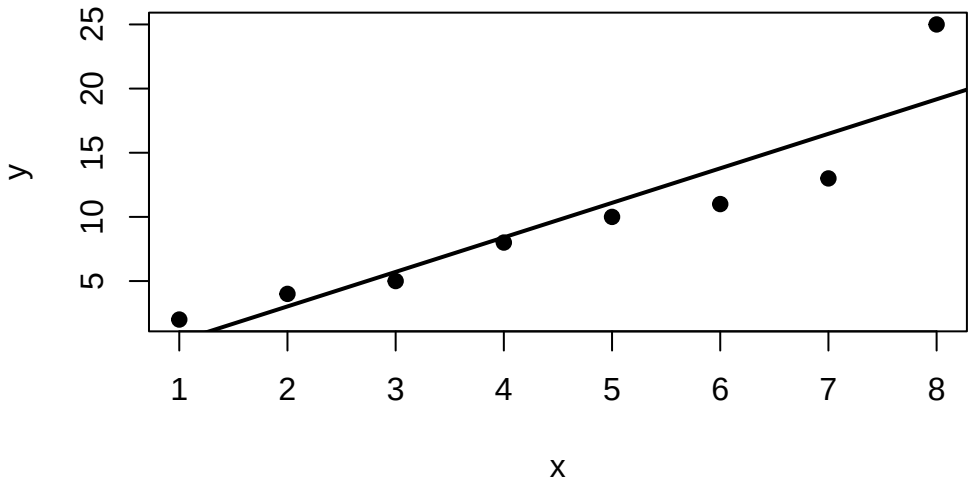
```r
x <- 1:8
y <- c(2, 4, 5, 8, 10, 11, 13, 25)

dat <- data.frame(x = x, y = y)
fit <- lm(y ~ x, data = dat)
summary(fit)
```

```
Call:
lm(formula = y ~ x, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4762 -1.5179 -0.5595  1.1488  5.8333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3571     2.4532  -0.961  0.37374
x             2.6905     0.4858   5.538  0.00146 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.148 on 6 degrees of freedom
Multiple R-squared:  0.8364,    Adjusted R-squared:  0.8091
F-statistic: 30.67 on 1 and 6 DF,  p-value: 0.001462
```

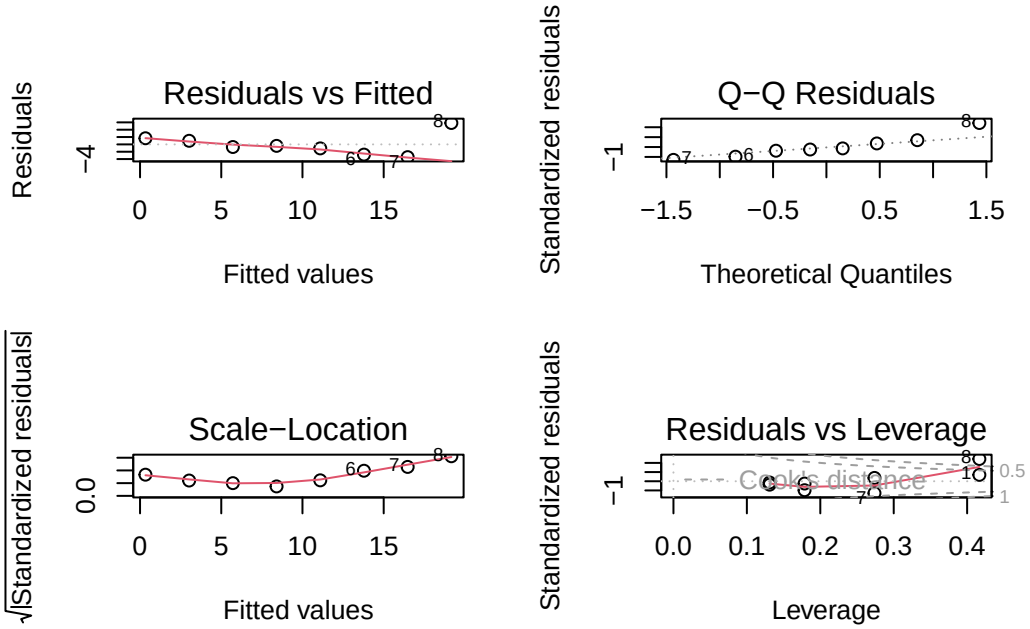
34.31 Scatterplot with fitted line

```r
plot(dat$x, dat$y, pch = 19, xlab = "x", ylab = "y")
abline(fit, lwd = 2)
```



34.32 Residual plots from base R

```r
par(mfrow = c(2, 2))
plot(fit)
```




```
par(mfrow = c(1, 1))
```

34.33 Extract basic diagnostics numerically

```
data.frame(
  fitted = fitted(fit),
  residual = resid(fit),
  std_resid = rstandard(fit),
  stud_resid = rstudent(fit),
  leverage = hatvalues(fit),
  cooks_d = cooks.distance(fit)
)
```

```
      fitted   residual  std_resid stud_resid  leverage     cooks_d
1  0.3333333  1.6666667  0.6930976  0.6596673 0.4166667 0.171565824
2  3.0238095  0.9761905  0.3638424  0.3358671 0.2738095 0.024957133
3  5.7142857 -0.7142857 -0.2503174 -0.2297101 0.1785714 0.006810742
4  8.4047619 -0.4047619 -0.1379056 -0.1260900 0.1309524 0.001432860
5 11.0952381 -1.0952381 -0.3731563 -0.3446665 0.1309524 0.010491110
6 13.7857143 -2.7857143 -0.9762380 -0.9716857 0.1785714 0.103591380
```

```
7 16.4761905 -3.4761905 -1.2956340 -1.3936654 0.2738095 0.316470107
8 19.1666667  5.8333333  2.4258417 15.9752413 0.4166667 2.101681345
```

34.34 Identify potentially unusual observations

```r
which(abs(rstudent(fit)) > 2)
```

```
8
8
```

```r
which(hatvalues(fit) > 2 * length(coef(fit)) / nrow(dat))
```

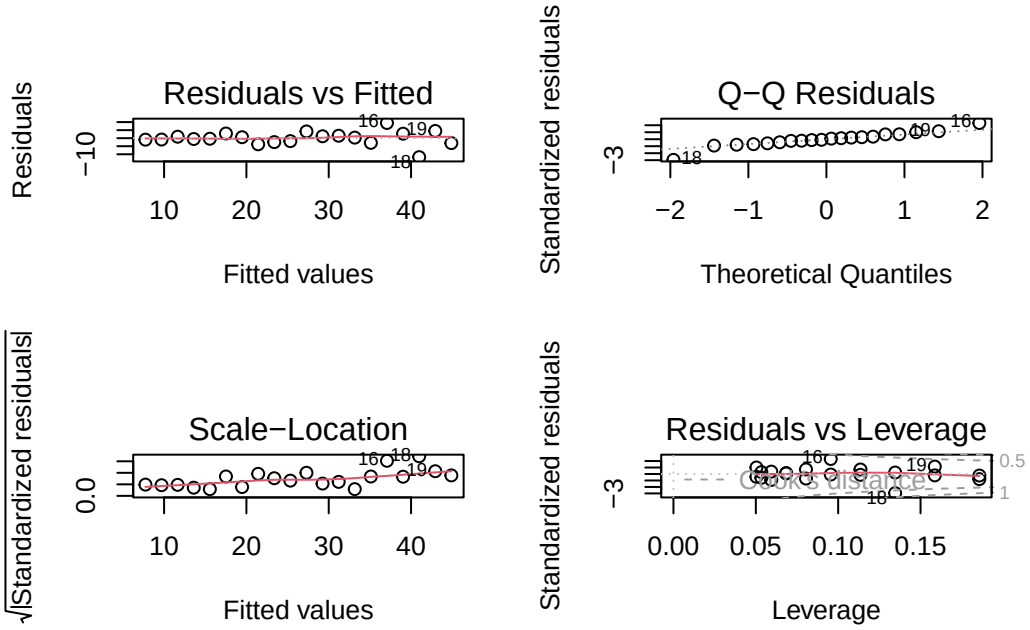
```
named integer(0)
```

```r
which(cooks.distance(fit) > 4 / nrow(dat))
```

```
8
8
```

34.35 Example With Heteroscedasticity-Like Pattern

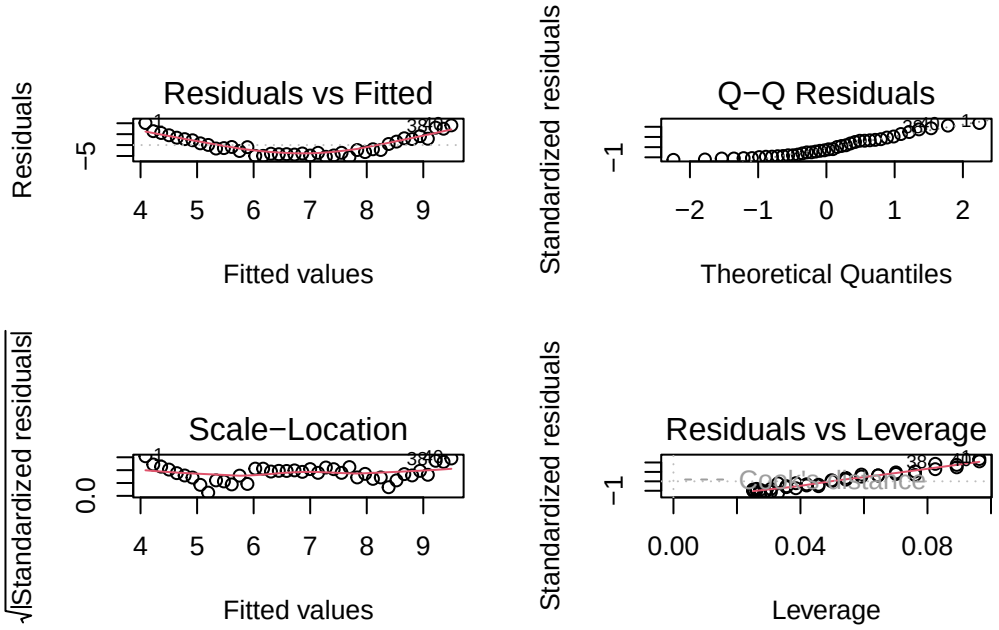
```r
set.seed(123)
x2 <- seq(1, 20, by = 1)
y2 <- 5 + 2 * x2 + rnorm(length(x2), sd = x2 / 3)

dat2 <- data.frame(x = x2, y = y2)
fit2 <- lm(y ~ x, data = dat2)

par(mfrow = c(2, 2))
plot(fit2)
```




```
par(mfrow = c(1, 1))
```

34.36 Example With Curvature

```
set.seed(321)
x3 <- seq(-3, 3, length.out = 40)
y3 <- 2 + x3 + 1.5 * x3^2 + rnorm(length(x3), sd = 1)

dat3 <- data.frame(x = x3, y = y3)
fit3 <- lm(y ~ x, data = dat3)

par(mfrow = c(2, 2))
plot(fit3)
```




```
par(mfrow = c(1, 1))
```

34.37 Comparing a Linear and Quadratic Fit

```
fit3_quad <- lm(y ~ x + I(x^2), data = dat3)
anova(fit3, fit3_quad)
```

```
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     38 765.56
2     37  37.17  1    728.39 725.14 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit3_quad)
```

```
Call:
lm(formula = y ~ x + I(x^2), data = dat3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26296 -0.67453  0.07213  0.50424  2.31450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01622    0.23783   8.478 3.38e-10 ***
x            0.90164    0.08923  10.104 3.45e-12 ***
I(x^2)       1.51417    0.05623  26.928  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 37 degrees of freedom
Multiple R-squared:  0.9572,    Adjusted R-squared:  0.9549
F-statistic: 413.6 on 2 and 37 DF,  p-value: < 2.2e-16
```

34.38 Interpreting Software Output

In practice, useful commands in R include:

- `plot(fit)` for the standard diagnostic panel;
- `resid(fit)` for residuals;
- `rstandard(fit)` for standardized residuals;
- `rstudent(fit)` for studentized residuals;
- `hatvalues(fit)` for leverages;
- `cooks.distance(fit)` for Cook's distances.

Students should learn to connect each numerical output to a concrete modelling question.

34.39 A Practical Diagnostic Workflow

A sensible basic workflow is:

- inspect the scatterplot and fitted line;
- examine the residuals-versus-fitted plot;
- check the normal Q-Q plot;
- inspect leverage and Cook's distance;

- investigate any unusual observations directly in the data;
- decide whether model revision is needed.

This sequence often works well for both simple and multiple regression.

34.40 What To Do After Finding a Problem

Finding a diagnostic issue does not mean we automatically delete observations.

Instead, possible next steps include:

- checking for data entry or measurement errors;
- understanding whether the point represents a different regime;
- refitting with and without the point for sensitivity analysis;
- revising the model form;
- transforming variables;
- reporting the issue transparently.

Model criticism should be thoughtful, not mechanical.

34.41 In-Class Discussion Questions

1. Why are raw residuals not enough for identifying unusual observations?
2. Why can a high-leverage point have a small residual and still matter?
3. Why is Cook's distance more about influence than outlyingness?
4. What kinds of model failure are easiest to detect from a residual-versus-fitted plot?

34.42 Practice Problems

34.43 Conceptual

1. Explain the difference between an outlier, a high-leverage observation, and an influential observation.
2. Explain why residuals have unequal variance.
3. Explain why a normal Q-Q plot is useful even though residuals are not independent.

34.44 Computational

Suppose a regression model has $n = 25$ observations and $p = 4$ parameters.

1. Compute the rough leverage benchmark $2p/n$.
2. If one observation has $h_{ii} = 0.45$, explain whether this seems unusually large.
3. Suppose an observation has a large studentized residual but very small leverage. What kind of problem does this suggest?
4. Suppose another observation has large leverage but a very small residual. Why might it still deserve attention?

34.45 Model-Criticism Problem

A residual-versus-fitted plot shows a clear U-shape.

1. What model assumption is likely failing?
2. What changes to the model might you consider?
3. Why would simply reporting coefficient p-values be insufficient here?

34.46 Suggested Homework

Complete the following tasks:

- fit a regression model in R and produce the standard diagnostic plots;
- identify any observations with large studentized residuals, high leverage, or large Cook's distance;
- write a short interpretation of each diagnostic plot;
- modify a model to address one detected issue, such as curvature or unequal variance;
- compare the original and revised models.

34.47 Summary

In this week, we studied model diagnostics for linear regression.

We focused on:

- residuals and standardized residuals;
- fitted-versus-residual plots;
- normal Q-Q plots;
- leverage and influence;

- Cook's distance and related ideas;
- practical judgment in assessing model adequacy.

These ideas are essential because regression analysis is not complete until the fitted model has been critically examined.

Next week, a natural continuation is to study transformations, remedies for nonconstant variance, and weighted least squares, or to move into multicollinearity and model selection, depending on the course emphasis.

34.48 Appendix: Compact Diagnostic Summary

For observation i:

- residual:
$$e_i = Y_i - \hat{Y}_i;$$
- leverage:
$$h_{ii} = \text{the } i\text{th diagonal entry of } \mathbf{H};$$
- standardized residual:
$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}};$$
- Cook's distance:
$$D_i = \frac{e_i^2}{p\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}.$$

Typical diagnostic questions are:

- Is the mean structure adequate?
- Is the variance roughly constant?
- Are the residuals approximately normal?
- Are any observations unusually influential?

35 Week 7: Transformations, Weighted Least Squares, and Remedial Measures

In this week, we study what to do when the assumptions of the ordinary linear model are not adequate. Building on residual analysis and diagnostics from the previous week, we now consider practical remedies for nonlinearity, nonconstant variance, and other violations of model assumptions. The emphasis is on understanding when transformations help, how weighted least squares works, and how to think systematically about model improvement.

35.1 Learning Objectives

By the end of this week, students should be able to:

- explain why transformations are used in regression analysis;
- distinguish between transforming the response and transforming predictors;
- interpret common response transformations such as logarithms and square roots;
- explain the basic idea of weighted least squares;
- derive the weighted least squares estimator;
- recognize situations where weighted least squares is appropriate;
- compare model improvement strategies based on diagnostics and scientific context.

35.2 Reading

Recommended reading for this week:

- Seber and Lee (2003) :
  - sections on transformations
  - weighted least squares
  - remedial measures for model inadequacy
- Montgomery et al. (2021) :
  - sections on transformations and model adequacy
  - weighted least squares
  - variance stabilization and practical remedies

35.3 Why Remedies Are Needed

Diagnostics often reveal that a fitted linear model is inadequate.

Common problems include:

- curvature in the mean structure;
- nonconstant error variance;
- skewed response distributions;
- influential observations;
- omitted interactions or nonlinear effects.

When this happens, we should not stop at saying the model is flawed. We should also ask how the model might be improved.

Some common remedies are:

- transforming the response;
- transforming predictors;
- adding polynomial or interaction terms;
- using weighted least squares;
- reconsidering the scope of the scientific question.

35.4 Review of the Ordinary Linear Model

The ordinary linear regression model is

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon] = \mathbf{0}, \qquad \operatorname{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

The ordinary least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

provided that $\mathbf{X}$ has full column rank.

When the variance is not constant, or when the mean structure is poorly represented by a linear form, OLS may still be computable, but its interpretation and inferential properties can become unsatisfactory.

35.5 Transformations in Regression

A transformation changes the scale on which a variable is analysed.

For example:

- logarithm;
- square root;
- reciprocal;
- power transformation.

Transformations are often used for one or more of the following reasons:

- to improve linearity;
- to stabilize variance;
- to reduce skewness;
- to make the model more scientifically interpretable;
- to reduce the impact of extreme values.

Transformations should not be viewed as purely mechanical. They should be guided by diagnostics and by substantive understanding of the problem.

35.6 Transforming the Response

Suppose the original response is Y, but instead of modelling Y directly, we model

$$g(Y)$$

for some transformation g.

Then the model becomes

$$g(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

This can help if the original relationship between the mean and the predictors is nonlinear, or if the variability of Y changes with its level.

35.7 Log Transformation of the Response

A very common choice is the logarithm:

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This is often useful when:

- the response is positive;
- the variance increases with the mean;
- the relationship is multiplicative rather than additive;
- the response distribution is right-skewed.

In this model, the interpretation of coefficients changes.

If x increases by one unit, then the expected log response changes by β_1. On the original scale, this corresponds approximately to a multiplicative change.

More precisely, if

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

then increasing x by one unit multiplies the fitted median response approximately by

$$e^{\beta_1}.$$

35.8 Square-Root Transformation

Another common response transformation is the square root:

$$\sqrt{Y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This is often used for count-like responses or responses whose variance tends to increase with the mean.

Compared with the log transformation, the square-root transformation is milder and can still be used when the response includes zeros.

35.9 Reciprocal and Other Power Transformations

In some settings, transformations such as

$$\frac{1}{Y}, \qquad Y^{\lambda}$$

or more general Box-Cox type power transformations may be useful.

These transformations can sometimes help linearize relationships or stabilize variance, but they may make interpretation more difficult. So one should balance statistical convenience and interpretability.

35.10 Transforming Predictors

Instead of transforming the response, we may transform one or more predictors.

For example, if the relationship between Y and x is curved, we might consider

- $\log(x)$;
- $\sqrt{x}$;
- x^2 or higher-order polynomial terms.

A model such as

$$Y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$$

is still linear in the parameters, even though it is nonlinear in the original predictor.

This is an important distinction: the model remains a linear model as long as it is linear in the coefficients.

35.11 Polynomial Terms as a Remedy

Suppose residual plots suggest curvature in the relationship between Y and x.

A common remedy is to add a quadratic term:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

This is still a linear model because it is linear in β_0, β_1, and β_2.

Polynomial terms often provide a more interpretable remedy than transforming the response, depending on the scientific context.

35.12 Choosing Between Transformations and Added Terms

Suppose diagnostics show nonlinearity.

Then several remedies may be reasonable:

- transform the predictor;
- transform the response;
- add polynomial terms;
- include an interaction;
- restrict attention to a smaller range of the data.

There is often no single automatic answer. Choice depends on:

- what shape is suggested by diagnostics;
- which form is scientifically meaningful;
- how easy the resulting model is to explain;
- whether inference or prediction is the main goal.

35.13 Heteroscedasticity and Variance Stabilization

A common problem in regression is heteroscedasticity, meaning that

$$\mathrm{Var}(\varepsilon_i)$$

is not constant across observations.

Residual plots may reveal a funnel shape or changing spread.

Sometimes a response transformation can reduce this problem.

For example:

- a log transformation often helps when the standard deviation is roughly proportional to the mean;
- a square-root transformation often helps for count-like data.

But in other cases, it is better to model the unequal variance directly. This leads to weighted least squares.

35.14 Weighted Least Squares

Suppose the observations still satisfy

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta,$$

but now the variance is

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{V},$$

where $\mathbf{V}$ is not the identity matrix.

A particularly common case is when the errors are uncorrelated but have unequal variances:

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 v_i.$$

Then observations have different levels of reliability.

Ordinary least squares treats all observations equally. Weighted least squares gives more weight to observations with smaller variance.

35.15 Basic Idea of Weights

If an observation has high variance, it contains less precise information about the mean.

If an observation has low variance, it contains more precise information.

So a sensible idea is to weight observations inversely to their variance.

If

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 v_i,$$

then we often use weights

$$w_i = \frac{1}{v_i}.$$

Larger weights correspond to more precise observations.

35.16 Weighted Least Squares Criterion

In weighted least squares, we minimize

$$Q(\beta) = \sum_{i=1}^{n} w_i (Y_i - x_i^\top \beta)^2.$$

In matrix form, if

$$\mathbf{W} = \operatorname{diag}(w_1, \dots, w_n),$$

then the criterion is

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta).$$

This is the weighted analogue of the ordinary residual sum of squares.

35.17 Derivation of the Weighted Least Squares Estimator

Differentiate the weighted criterion with respect to β and set the derivative equal to zero.

The weighted normal equations are

$$\mathbf{X}^\top \mathbf{W} \mathbf{X} \, \widehat{\beta}_{WLS} = \mathbf{X}^\top \mathbf{W} \mathbf{Y}.$$

Assuming invertibility, the weighted least squares estimator is

$$\widehat{\beta}_{WLS} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y}.$$

This has exactly the same structural form as OLS, but with $\mathbf{W}$ inserted to reflect unequal precision.

35.18 Transformation View of Weighted Least Squares

Weighted least squares can also be understood as transforming the model.

If $\mathbf{W}^{1/2}$ is the diagonal matrix with entries $\sqrt{w_i}$, then multiplying the model by $\mathbf{W}^{1/2}$ gives

$$\mathbf{W}^{1/2}\mathbf{Y} = \mathbf{W}^{1/2}\mathbf{X}\beta + \mathbf{W}^{1/2}\varepsilon.$$

If the weights are chosen appropriately, the transformed errors have constant variance, and ordinary least squares on the transformed system becomes appropriate.

This is a very useful conceptual link between WLS and OLS.

35.19 Interpreting Weighted Least Squares

Weighted least squares does not change the target mean model

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

Instead, it changes how observations are used in estimation.

Observations with smaller variance have more influence on the fitted coefficients.

Thus WLS is often best viewed as a remedy for unequal precision rather than a new mean model.

35.20 When Weighted Least Squares Is Appropriate

Weighted least squares is particularly useful when:

- variance is known up to a proportional constant;
- variance can be reasonably modelled as a function of the mean or a predictor;
- observations are averages based on different sample sizes;
- some observations are measured more precisely than others.

For example, if each response is an average based on m_i repeated measurements, then the variance may be proportional to $1/m_i$, so weights proportional to m_i are natural.

35.21 Feasible Weighted Least Squares

In many applications, the variances are not known exactly.

Instead, we estimate the variance pattern from the data and then use estimated weights. This is often called **feasible weighted least squares**.

Typical workflow:

- fit an initial OLS model;
- inspect residuals to understand how variance changes;
- propose a variance model;
- compute estimated weights;
- refit using WLS.

This approach is practical, though it introduces additional modelling decisions.

35.22 Example of a Mean-Variance Relationship

Suppose residual plots suggest

$$\mathrm{Var}(Y_i) \propto x_i^2.$$

Then it may be appropriate to use weights

$$w_i = \frac{1}{x_i^2}.$$

Alternatively, dividing the model through by x_i may also lead to a transformed model with approximately constant variance.

This illustrates the close connection between variance modelling and transformation.

35.23 Response Transformation Versus WLS

Both response transformation and WLS can address heteroscedasticity, but they do so differently.

A response transformation changes the scale of the response and often changes interpretation.

Weighted least squares keeps the response on its original scale, but changes the estimation procedure.

Which is better depends on the context.

Students should compare:

- interpretability;
- adequacy of the residual plots after refitting;
- scientific plausibility of the variance model.

35.24 Practical Caveats

No remedy should be applied blindly.

Questions to ask include:

- Does the transformed model make scientific sense?
- Does the transformed scale improve diagnostics meaningfully?
- Are the chosen weights justified?
- Does a more flexible mean model explain the apparent variance pattern?
- Would a different modelling framework be more appropriate?

Model repair is not merely technical. It should remain connected to the original scientific problem.

35.25 Worked Example With a Log Transformation

Suppose the response increases rapidly with the predictor and residual plots show larger spread for larger fitted values.

A model for the original scale may look like

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

but the diagnostics suggest increasing variance and curvature.

We might instead fit

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

If the transformed residual plots look more stable and more linear, then the log model may be preferable.

Interpretation then becomes multiplicative rather than additive.

35.26 Worked Example With Weighted Least Squares

Suppose we observe responses with variance increasing as the predictor grows.

If residual analysis suggests

$$\mathrm{Var}(Y_i) \propto x_i^2,$$

then we may fit a weighted model with weights

$$w_i = \frac{1}{x_i^2}.$$

This gives lower weight to high-variance observations and can produce more stable coefficient estimates and more appropriate standard errors.

35.27 R Demonstration With a Log Transformation

35.28 Generate heteroscedastic data

```
set.seed(123)
x <- seq(1, 20, by = 1)
y <- exp(1 + 0.12 * x + rnorm(length(x), sd = 0.25))

dat <- data.frame(x = x, y = y)

fit_raw <- lm(y ~ x, data = dat)
fit_log <- lm(log(y) ~ x, data = dat)
```

35.29 Compare the fitted models

```
summary(fit_raw)
```

```
Call:
lm(formula = y ~ x, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0995 -2.6491  0.1043  2.0486  9.1089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2531     1.9582  -0.640     0.53
x             1.3205     0.1635   8.078 2.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.215 on 18 degrees of freedom
Multiple R-squared:  0.7838,    Adjusted R-squared:  0.7718
F-statistic: 65.25 on 1 and 18 DF,  p-value: 2.134e-07
```

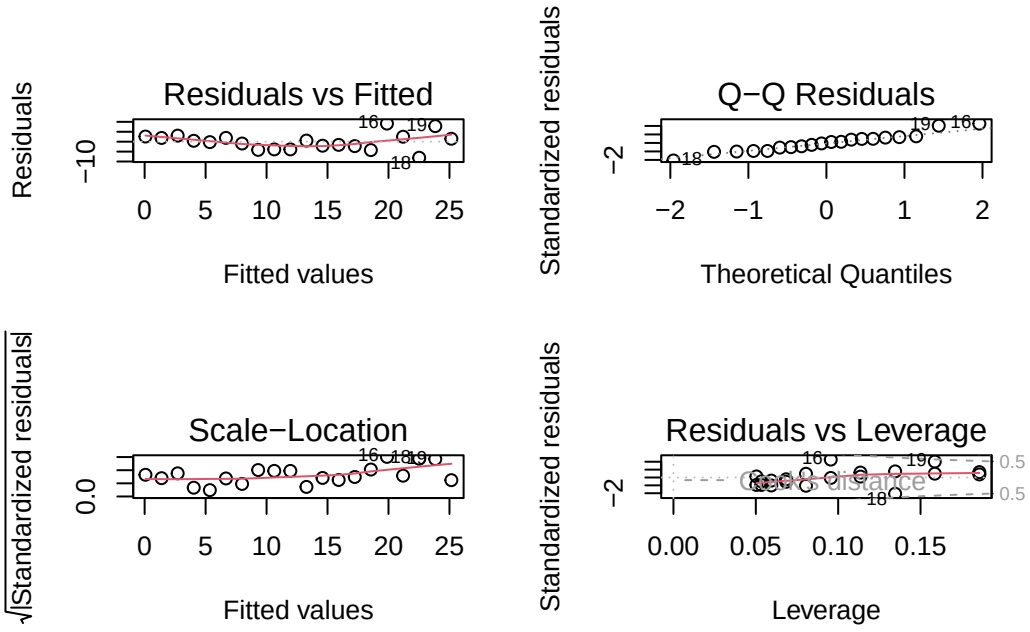
```
summary(fit_log)
```

```
Call:
lm(formula = log(y) ~ x, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49698 -0.15070 -0.00942  0.12985  0.43338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.077523   0.115500   9.329 2.57e-08 ***
x           0.115989   0.009642  12.030 4.85e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2486 on 18 degrees of freedom
Multiple R-squared:  0.8894,    Adjusted R-squared:  0.8832
F-statistic: 144.7 on 1 and 18 DF,  p-value: 4.848e-10
```

35.30 Diagnostic plots for the untransformed model

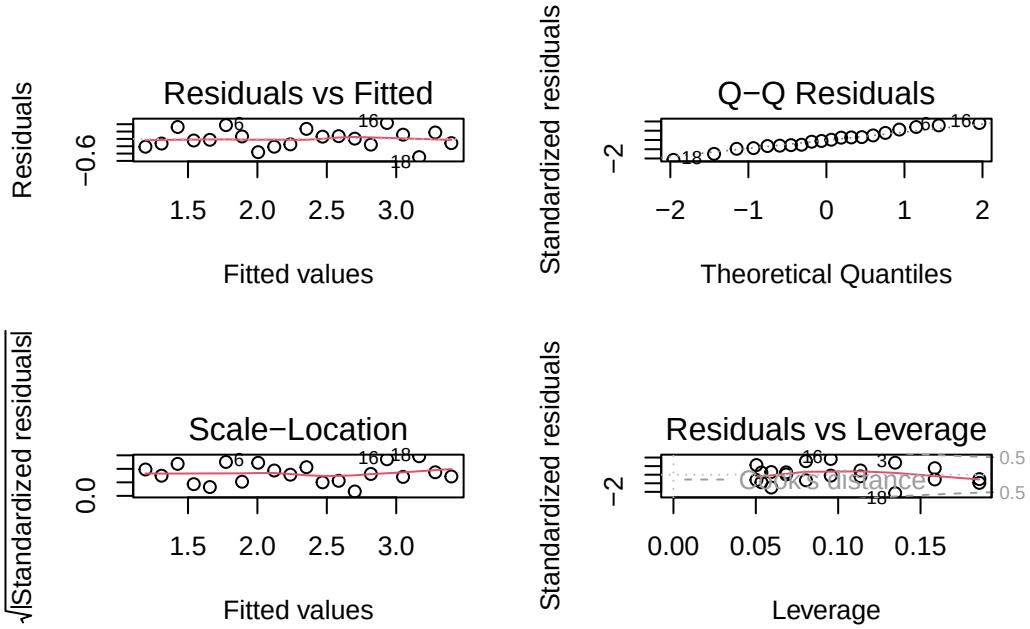
```
par(mfrow = c(2, 2))
plot(fit_raw)
```




```
par(mfrow = c(1, 1))
```

35.31 Diagnostic plots for the log-transformed model

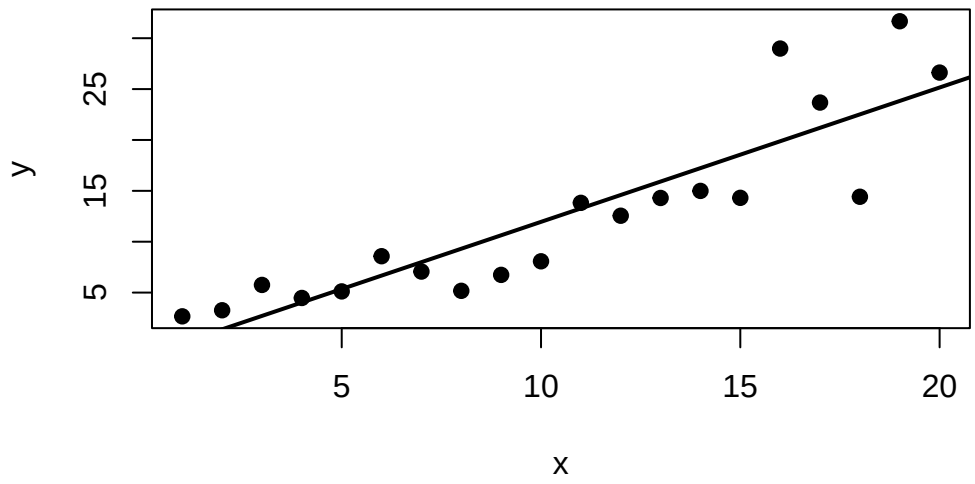
```
par(mfrow = c(2, 2))
plot(fit_log)
```



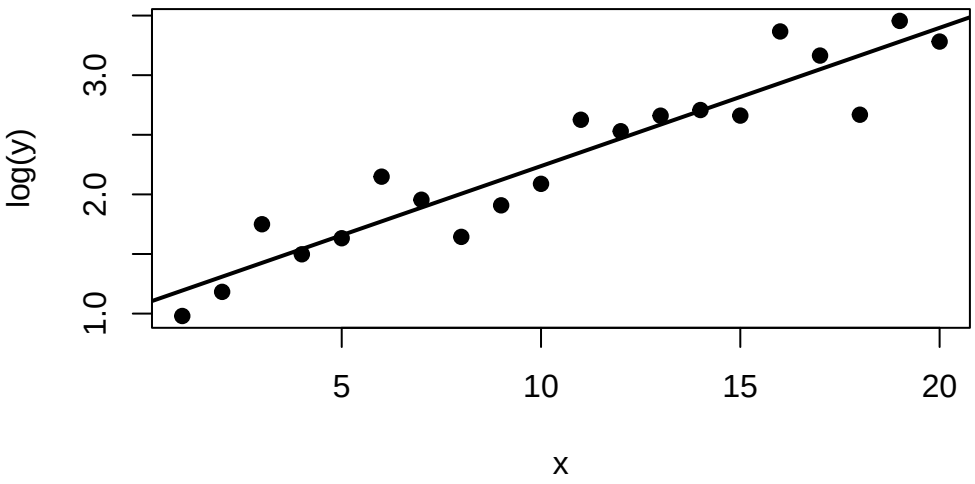

```
par(mfrow = c(1, 1))
```

35.32 Plot data on original and transformed scales

```
plot(dat$x, dat$y, pch = 19, xlab = "x", ylab = "y")
abline(fit_raw, lwd = 2)
```




```r
plot(dat$x, log(dat$y), pch = 19, xlab = "x", ylab = "log(y)")
abline(fit_log, lwd = 2)
```



35.33 R Demonstration With Weighted Least Squares

35.34 Generate data with variance increasing in x

```
set.seed(456)
x2 <- seq(1, 15, by = 1)
y2 <- 5 + 2 * x2 + rnorm(length(x2), sd = 0.6 * x2)

dat2 <- data.frame(x = x2, y = y2)

fit_ols <- lm(y ~ x, data = dat2)
fit_wls <- lm(y ~ x, data = dat2, weights = 1 / x^2)
```

35.35 Compare summaries

```
summary(fit_ols)
```

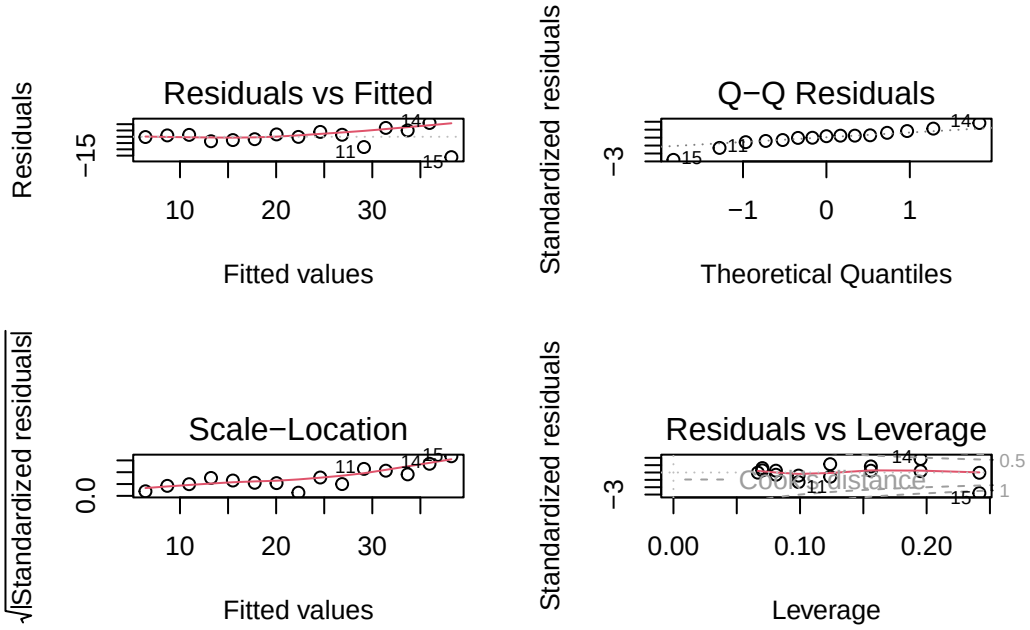
```
Call:
lm(formula = y ~ x, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.190  -2.291   1.063   2.853  10.942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.1385     3.6015   1.149    0.271
x             2.2723     0.3961   5.736 6.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.628 on 13 degrees of freedom
Multiple R-squared:  0.7168,    Adjusted R-squared:  0.695
F-statistic: 32.91 on 1 and 13 DF,  p-value: 6.862e-05
```

```r
summary(fit_wls)
```

```
Call:
lm(formula = y ~ x, data = dat2, weights = 1/x^2)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.0753 -0.4167  0.1633  0.4726  0.7858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0940     0.6638   6.167 3.39e-05 ***
x             2.2713     0.2155  10.541 9.73e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6107 on 13 degrees of freedom
Multiple R-squared:  0.8953,    Adjusted R-squared:  0.8872
F-statistic: 111.1 on 1 and 13 DF,  p-value: 9.732e-08
```

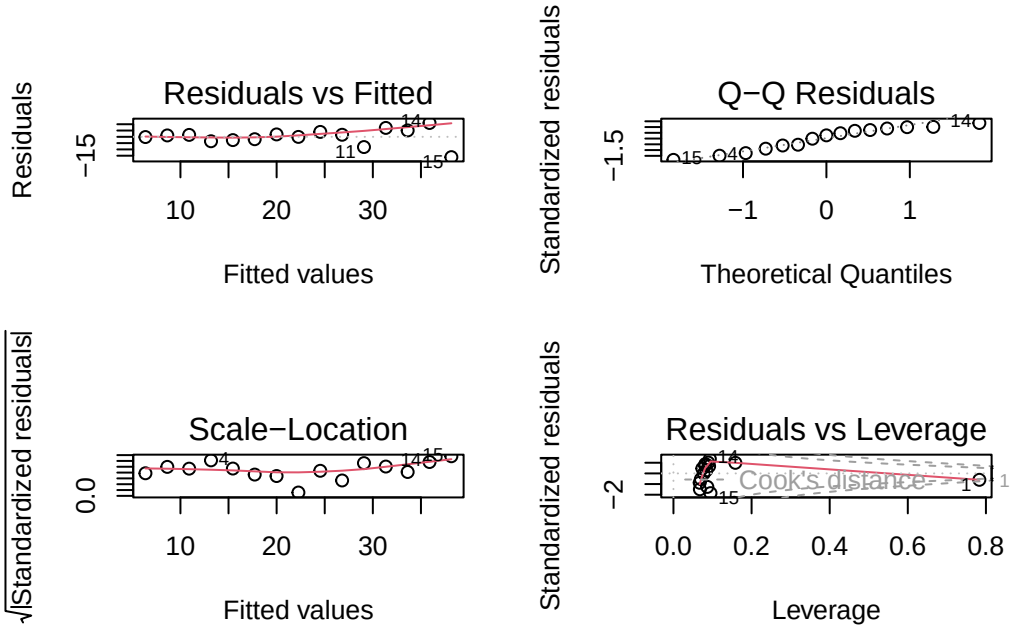
35.36 Compare diagnostic plots

```r
par(mfrow = c(2, 2))
plot(fit_ols)
```




```
par(mfrow = c(1, 1))
```

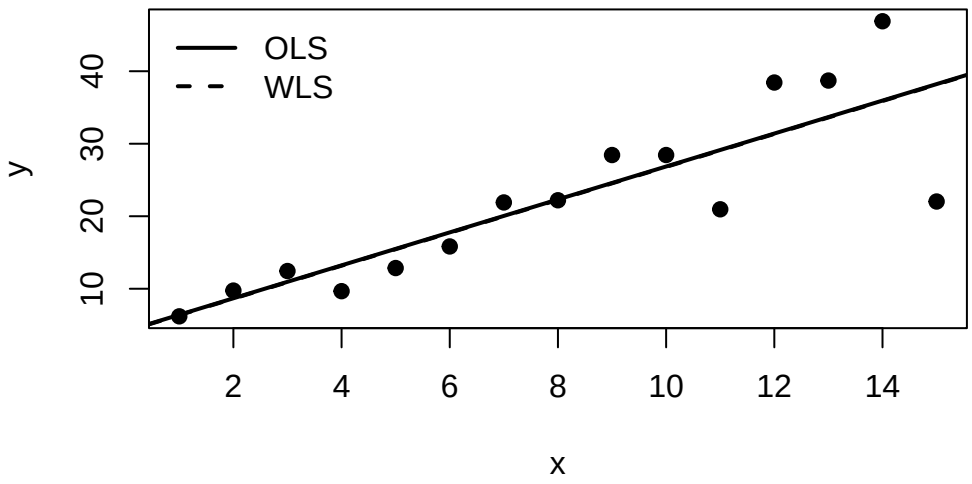
```
par(mfrow = c(2, 2))
plot(fit_wls)
```




```
par(mfrow = c(1, 1))
```

35.37 Plot fitted lines

```
plot(dat2$x, dat2$y, pch = 19, xlab = "x", ylab = "y")
abline(fit_ols, lwd = 2)
abline(fit_wls, lwd = 2, lty = 2)
legend("topleft",
       legend = c("OLS", "WLS"),
       lty = c(1, 2),
       lwd = 2,
       bty = "n")
```



35.38 Example With a Quadratic Remedy for Curvature

```
set.seed(789)
x3 <- seq(-3, 3, length.out = 50)
y3 <- 3 + 2 * x3 + 1.8 * x3^2 + rnorm(length(x3), sd = 1)

dat3 <- data.frame(x = x3, y = y3)

fit_lin <- lm(y ~ x, data = dat3)
fit_quad <- lm(y ~ x + I(x^2), data = dat3)

summary(fit_lin)
```

```
Call:
lm(formula = y ~ x, data = dat3)

Residuals:
   Min     1Q Median     3Q    Max
-6.891 -4.391 -1.388  3.994 11.626
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.4756     0.7498   11.30 3.94e-15 ***
x             2.1258     0.4243    5.01 7.79e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.302 on 48 degrees of freedom
Multiple R-squared:  0.3434,    Adjusted R-squared:  0.3297
F-statistic:  25.1 on 1 and 48 DF,  p-value: 7.79e-06
```

```
summary(fit_quad)
```

```
Call:
lm(formula = y ~ x + I(x^2), data = dat3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4044 -0.4148 -0.0022  0.5909  1.6173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74823    0.19245   14.28   <2e-16 ***
x            2.12581    0.07258   29.29   <2e-16 ***
I(x^2)       1.83425    0.04595   39.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9069 on 47 degrees of freedom
Multiple R-squared:  0.9812,    Adjusted R-squared:  0.9804
F-statistic:  1226 on 2 and 47 DF,  p-value: < 2.2e-16
```

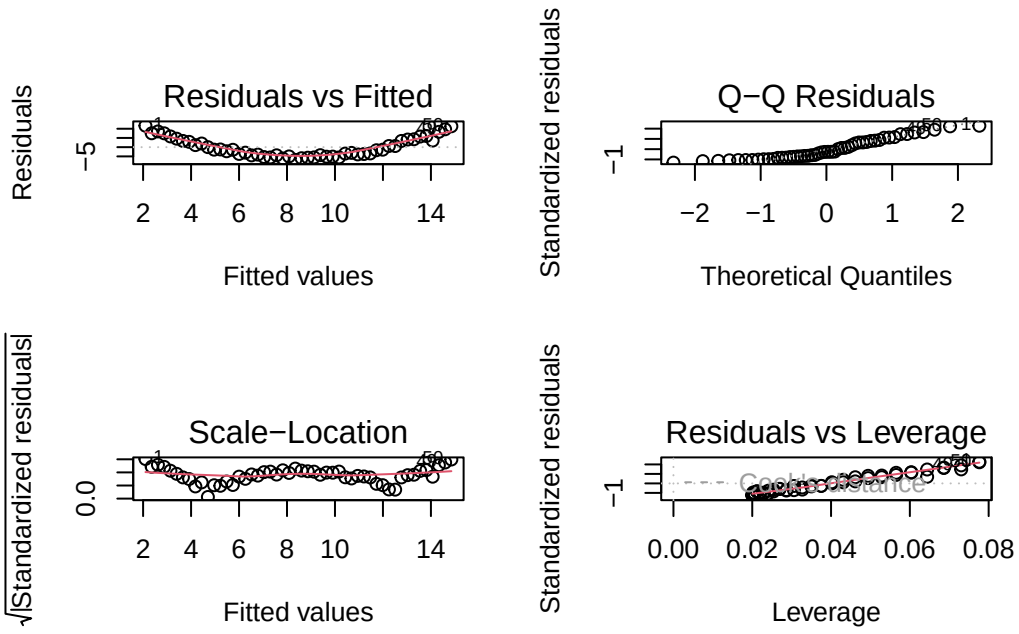
```
anova(fit_lin, fit_quad)
```

```
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
1     48 1349.18
2     47   38.66  1    1310.5 1593.4 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

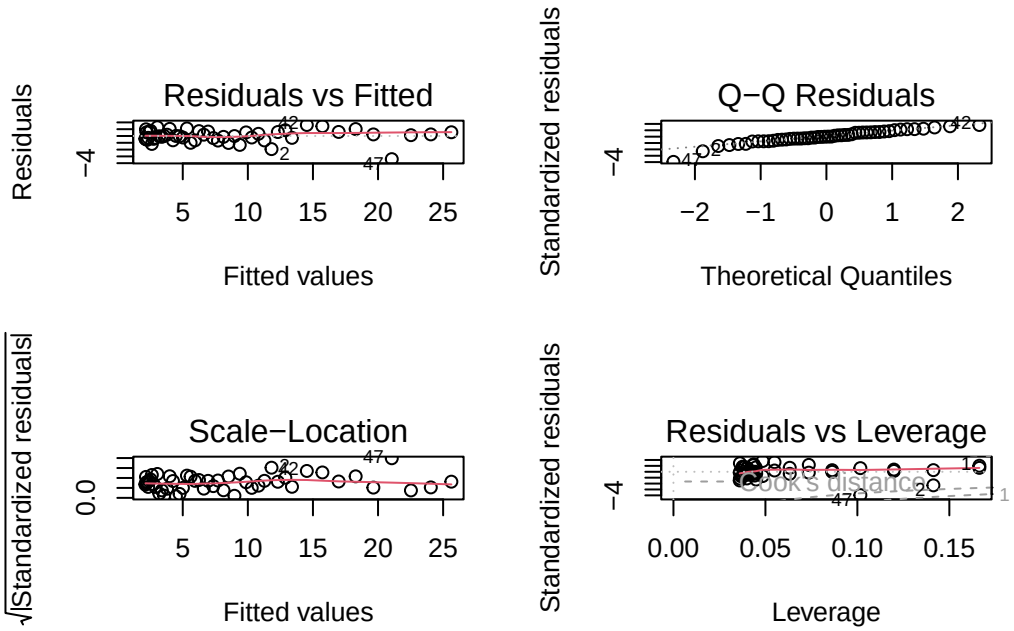
35.39 Compare diagnostic plots for linear and quadratic fits

```
par(mfrow = c(2, 2))
plot(fit_lin)
```




```
par(mfrow = c(1, 1))
```

```
par(mfrow = c(2, 2))
plot(fit_quad)
```




```
par(mfrow = c(1, 1))
```

35.40 Interpreting Software Output

Useful commands in R include:

- `lm(..., weights = ...)` for weighted least squares;
- `plot(fit)` for standard diagnostic plots;
- `anova(fit1, fit2)` for comparing nested mean models;
- `predict()` for fitted values on the chosen modelling scale.

Students should always keep track of the scale on which the model is fitted. This is especially important when the response is transformed.

35.41 A Practical Remedy Workflow

A useful workflow after diagnostics is:

- identify the main problem from residual analysis;
- decide whether the issue concerns the mean structure, the variance structure, or both;
- try a scientifically reasonable remedy;

- refit the model;
- compare diagnostics before and after the change;
- interpret the new model on the correct scale.

This encourages disciplined model improvement rather than ad hoc trial and error.

35.42 In-Class Discussion Questions

1. When is a response transformation preferable to adding polynomial terms?
2. How does the interpretation of coefficients change under a log transformation?
3. Why does weighted least squares downweight high-variance observations?
4. Why should model remedies be guided by both diagnostics and subject-matter knowledge?

35.43 Practice Problems

35.44 Conceptual

1. Explain the difference between transforming the response and transforming a predictor.
2. Explain why a log transformation may help when the variance grows with the mean.
3. Explain why weighted least squares can be viewed as ordinary least squares on a transformed system.

35.45 Computational

Suppose a regression model has

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2.$$

1. What weights would be natural for weighted least squares?
2. Which observations receive the largest weights?
3. Why do these weights make sense?

Now suppose the fitted model is

$$\log(Y_i) = 1.5 + 0.2x_i.$$

1. What is the fitted log response when $x_i = 3$?

2. What is the fitted value on the original response scale if you exponentiate the fitted mean?
3. Why should interpretation on the original scale be made carefully?

35.46 Model-Improvement Problem

A residual-versus-fitted plot shows both a curved pattern and increasing spread.

1. What kinds of model inadequacy does this suggest?
2. Name two possible remedies.
3. How would you decide which remedy is more appropriate?

35.47 Suggested Homework

Complete the following tasks:

- fit a regression model that shows heteroscedasticity or curvature;
- apply one response transformation and assess whether the diagnostics improve;
- fit a weighted least squares model with a justified set of weights;
- compare OLS and WLS fits both numerically and graphically;
- write a short discussion explaining which remedy you prefer and why.

35.48 Summary

In this week, we studied practical remedies for model inadequacy in linear regression.

We focused on:

- transforming the response or predictors;
- using polynomial terms to address curvature;
- stabilizing variance through transformation;
- using weighted least squares when error variances differ across observations;
- comparing alternative remedies using diagnostics and interpretation.

These ideas help students move from model criticism to model improvement.

Next week, a natural continuation is to study multicollinearity, variable selection, and model-building strategies, or to move into generalized least squares and correlated errors, depending on the course emphasis.

35.49 Appendix: Compact Formula Summary

Response transformation model:

$$g(Y_i) = x_i^\top \beta + \varepsilon_i.$$

Weighted least squares criterion:

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{W}(\mathbf{Y} - \mathbf{X}\beta).$$

Weighted least squares estimator:

$$\widehat{\beta}_{WLS} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{Y}.$$

Typical diagnostic questions after a remedy:

- Is the mean structure more adequate?
- Is the variance more stable?
- Is interpretation still meaningful?
- Does the new model answer the scientific question well?

36 Week 8: Multicollinearity, Variable Selection, and Model Building

In this week, we study how regression models behave when predictors are strongly related to one another, how this affects interpretation and inference, and how to think carefully about selecting variables for a useful final model. The emphasis is on understanding multicollinearity, model-building principles, and the strengths and limitations of common selection procedures.

36.1 Learning Objectives

By the end of this week, students should be able to:

- explain what multicollinearity is and why it matters in multiple regression;
- distinguish between good prediction and stable coefficient interpretation;
- diagnose multicollinearity using correlations, variance inflation factors, and related tools;
- explain the ideas behind forward selection, backward elimination, and stepwise procedures;
- compare model selection based on adjusted R^2, AIC, BIC, and cross-validation style thinking;
- discuss principled strategies for building and comparing regression models.

36.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on collinearity
  - model-building considerations
  - variable selection and related issues
- Montgomery, Peck, and Vining:
  - sections on multicollinearity
  - variable selection methods
  - model-building strategies and cautions

36.3 Why Model Building Is Difficult

In multiple regression, it is often easy to write down many candidate predictors, transformations, and interactions.

The harder questions are:

- Which variables should be included?
- Which effects are scientifically meaningful?
- Are some predictors redundant?
- Can we trust the estimated coefficients?
- Are we building a model for explanation, inference, or prediction?

A good model is rarely defined only by having the largest possible R^2. We also care about interpretability, stability, scientific plausibility, and predictive usefulness.

36.4 Review of the Multiple Regression Model

Recall the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

or in matrix form,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

When predictors are moderately distinct and the design matrix is well behaved, the least squares estimator

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

can be interpreted in the usual way.

But when predictors are strongly related to one another, estimation becomes more delicate.

36.5 What Is Multicollinearity

Multicollinearity means that one predictor is highly linearly related to one or more of the other predictors.

At an extreme, one predictor may be an exact linear combination of others. Then the design matrix loses full rank, and the OLS estimator is not uniquely defined.

More commonly, the relationship is not exact, but is still strong enough to cause instability. This is often called **near multicollinearity**.

36.6 Why Multicollinearity Matters

Multicollinearity does not necessarily harm the fitted values very much. In fact, a model can still predict reasonably well.

The main problems are:

- coefficient estimates can become unstable;
- standard errors can become large;
- signs and magnitudes of coefficients can become counterintuitive;
- individual t tests can become weak even when the overall model is useful;
- small changes in the data can lead to large changes in estimated coefficients.

Thus multicollinearity is especially important when interpretation and inference on individual coefficients matter.

36.7 A Simple Intuition

Suppose two predictors measure almost the same underlying quantity.

Then the model has difficulty deciding how much of the fitted effect should be attributed to one predictor and how much to the other.

The sum of their contributions may be estimated fairly well, but the individual coefficients may not be.

This is why high collinearity often affects coefficient interpretation more than overall model fit.

36.8 Exact Collinearity

If one column of $\mathbf{X}$ is an exact linear combination of the others, then

$$\operatorname{rank}(\mathbf{X}) < p,$$

and

$$\mathbf{X}^\top \mathbf{X}$$

is singular.

In that case, ordinary least squares does not produce a unique coefficient vector without imposing additional constraints.

This can happen, for example, if we include all indicator variables for a factor together with an intercept.

36.9 Near Collinearity

In many applications, the problem is not exact singularity but near singularity.

Then

$$\mathbf{X}^\top \mathbf{X}$$

is invertible, but poorly conditioned.

This makes

$$(\mathbf{X}^\top \mathbf{X})^{-1}$$

numerically and statistically unstable, which inflates the variance of coefficient estimates.

36.10 Multicollinearity and Variance

Recall that under the standard linear model,

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}.$$

So when the columns of $\mathbf{X}$ are highly correlated, the inverse matrix tends to have large diagonal entries.

This leads to large coefficient variances and hence large standard errors.

That is the algebraic reason multicollinearity makes inference unstable.

36.11 Pairwise Correlations

A simple first check is to inspect the pairwise correlations among predictors.

Large pairwise correlations may suggest multicollinearity.

However, this is not a complete diagnostic, because a predictor can be strongly related to a combination of several others even if no single pairwise correlation is extreme.

So pairwise correlations are useful, but not sufficient.

36.12 Variance Inflation Factor

One of the most common diagnostics is the **variance inflation factor**, or VIF.

For predictor x_j, the VIF is

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

where R_j^2 is the coefficient of determination obtained by regressing x_j on the remaining predictors.

Interpretation:

- if x_j is almost unrelated to the others, then R_j^2 is small and the VIF is close to 1;
- if x_j is highly explained by the others, then R_j^2 is close to 1 and the VIF is large.

So the VIF measures how much the variance of $\widehat{\beta}_j$ is inflated by collinearity.

36.13 Interpreting VIF Values

There is no universal cutoff, but common informal rules are:

- VIF near 1: little concern;
- VIF above 5: possible concern;
- VIF above 10: serious concern in many applications.

These are only rough guidelines. The seriousness depends on the context, the sample size, and the goal of the analysis.

36.14 Tolerance

A related quantity is **tolerance**, defined as

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

Small tolerance indicates that the predictor is largely explained by the others.

Some software reports tolerance instead of VIF.

36.15 Consequences for Hypothesis Tests

A common symptom of multicollinearity is the following:

- the overall F test is significant;
- individual t tests are weak or nonsignificant.

This can happen because the predictors collectively explain the response well, but it is difficult to estimate the separate contribution of each one.

Students often find this confusing at first, but it is a standard effect of collinearity.

36.16 Condition Number and Eigenvalue Thinking

Another way to view multicollinearity is through the eigenstructure of the predictor matrix or the correlation matrix of predictors.

If the design is nearly singular, then some directions in predictor space are weakly supported by the data.

This is often summarized through condition numbers or condition indices.

A large condition number indicates that the model matrix is close to singular in some direction.

For an introductory course, VIFs and predictor correlations are often enough, but it is useful to mention the geometric idea.

36.17 Remedies for Multicollinearity

Possible responses to multicollinearity include:

- removing one of several redundant predictors;
- combining related predictors into a single summary variable;
- centring variables, especially when polynomial or interaction terms are present;
- collecting more data, if possible;
- focusing on prediction rather than individual coefficient interpretation;
- using regularization methods such as ridge regression, if the course later extends in that direction.

The best remedy depends on the scientific objective.

36.18 Centering and Polynomial Terms

When polynomial terms such as x and x^2 are both included, the terms can be highly correlated.

A common remedy is to centre the predictor first:

$$x_i^* = x_i - \bar{x}.$$

Then the model may be written using x_i^* and $(x_i^*)^2$.

This often improves numerical stability and makes the intercept more interpretable, though it does not solve all collinearity issues.

36.19 Variable Selection as a Modelling Problem

Beyond collinearity, regression analysts must often decide which predictors belong in the model.

This is called **variable selection** or **model selection**.

The central challenge is that adding variables can improve apparent fit in the sample, but may produce a model that is unstable, hard to interpret, or overly tailored to the data.

Thus variable selection is not just a computational problem. It is a statistical and scientific problem.

36.20 Goals of Variable Selection

Variable selection may be done for different reasons:

- to improve prediction;
- to simplify interpretation;
- to reduce cost of measurement;
- to remove irrelevant or redundant predictors;
- to identify a scientifically meaningful parsimonious model.

Because these goals differ, there is no single best selection rule for every application.

36.21 Forward Selection

In **forward selection**, we begin with a small model, often the intercept-only model, and add predictors one at a time.

At each step, we add the variable that gives the best improvement according to some criterion, such as:

- partial F test;
- smallest p-value;
- largest drop in AIC;
- largest increase in adjusted R^2.

This continues until no candidate addition meets the chosen rule.

36.22 Backward Elimination

In **backward elimination**, we begin with a larger model and remove predictors one at a time.

At each step, we remove the least useful variable according to a chosen rule.

This continues until all remaining variables satisfy the stopping criterion.

Backward elimination requires that the initial model be estimable, and it may be sensitive to collinearity and hierarchical structure.

36.23 Stepwise Selection

Stepwise selection combines forward and backward ideas.

A variable may enter at one stage and later be removed if its contribution becomes weak after other variables enter the model.

This procedure is popular in software because it is automated, but it should be used cautiously.

Automatic procedures can be unstable and may encourage overly mechanical modelling.

36.24 Problems With Automatic Selection

Automatic variable selection can create several difficulties:

- it ignores model uncertainty;
- repeated searching inflates the chance of false discoveries;
- selected coefficients and p-values may look more certain than they really are;
- different but similar datasets may lead to different selected models;
- scientific structure may be lost if the procedure is used blindly.

So automated methods should be treated as exploratory tools, not as final arbiters of truth.

36.25 Hierarchical Principle

When interactions or polynomial terms are included, the **hierarchical principle** is often recommended.

For example, if the model contains

$$x_1 x_2,$$

then it is usually sensible to keep the corresponding main effects x_1 and x_2 in the model as well.

Similarly, if a quadratic term x^2 is included, then it is usually sensible to keep the linear term x.

This helps preserve interpretability and avoids awkward models.

36.26 Criteria for Comparing Models

Several criteria are commonly used in model comparison.

36.26.1 Adjusted R Squared

Adjusted R^2 is

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}.$$

Unlike ordinary R^2, adjusted R^2 can decrease when unnecessary predictors are added.

It rewards fit but penalizes unnecessary complexity in a simple way.

36.26.2 AIC

The Akaike information criterion is typically written as

$$\text{AIC} = -2 \log L + 2k,$$

where L is the fitted likelihood and k is the number of estimated parameters.

Smaller AIC indicates a better tradeoff between fit and complexity.

AIC is often more prediction-oriented than strict parsimony-oriented.

36.26.3 BIC

The Bayesian information criterion is

$$\text{BIC} = -2 \log L + k \log n.$$

Because the penalty term is stronger than that of AIC when n is moderate or large, BIC tends to favor smaller models.

36.26.4 Mallows' Cp

Another classical criterion is Mallows' C_p, which compares model bias and variance relative to a fuller model.

It is less frequently emphasized in introductory software workflows today, but it remains conceptually important in regression theory.

36.27 Prediction-Oriented Thinking

If the main goal is prediction rather than interpretation, then the evaluation should focus more on performance on new data.

This leads naturally to validation ideas such as:

- training and test set comparison;
- cross-validation;
- prediction error rather than coefficient significance.

Even if a course stays within classical regression, it is valuable for students to know that in-sample fit is not the same as out-of-sample performance.

36.28 Overfitting

A model is **overfit** when it captures not only real structure but also noise specific to the sample.

Symptoms include:

- excellent fit on the observed data;
- unstable coefficients;
- poor performance on new data;
- excessive complexity relative to the available sample size.

Variable selection is closely tied to the problem of overfitting.

36.29 Parsimony

A guiding principle in model building is **parsimony**.

A parsimonious model is one that is no more complicated than necessary for the purpose at hand.

This does not mean the smallest possible model. It means a model that balances:

- adequacy of fit;
- interpretability;
- stability;
- scientific usefulness.

36.30 Subject-Matter Knowledge

No model-building strategy should rely only on algorithmic output.

Subject-matter knowledge can help determine:

- which variables are essential controls;
- which interactions are scientifically plausible;
- which terms should remain in the model even if their p-values are not small;
- whether a selected model is substantively reasonable.

This is especially important when the goal is explanation or causal interpretation.

36.31 A Practical Model-Building Strategy

A reasonable workflow is:

- start from the scientific question;
- define a set of plausible predictors and model terms;
- fit an initial model;
- assess collinearity and diagnostics;
- simplify or revise where appropriate;
- compare a small number of meaningful candidate models;
- justify the final choice in words, not just by one number.

This approach is often more reliable than indiscriminate stepwise searching.

36.32 Worked Example With Strongly Correlated Predictors

Suppose we fit a model with two predictors, x_1 and x_2, that are highly correlated.

We may find that:

- both variables together give a strong overall fit;
- the overall F test is significant;
- one or both individual coefficient tests are weak;
- the signs of coefficients may look unstable or surprising.

This is a classic signature of multicollinearity.

36.33 Worked Example With Competing Models

Suppose we have candidate models:

- a small model with two predictors;
- a medium model with four predictors;
- a larger model with six predictors and interactions.

The best choice may differ depending on whether we care most about:

- interpretability;
- inference on key coefficients;
- predictive performance;
- scientific completeness.

This is why model-building decisions should be tied to the purpose of the analysis.

36.34 R Demonstration With Correlated Predictors

36.35 Generate data with multicollinearity

```r
set.seed(123)
n <- 80
x1 <- rnorm(n)
x2 <- 0.92 * x1 + rnorm(n, sd = 0.25)
x3 <- rnorm(n)
y <- 3 + 2 * x1 - 1.5 * x2 + 1.2 * x3 + rnorm(n, sd = 1)

dat <- data.frame(y = y, x1 = x1, x2 = x2, x3 = x3)

fit_full <- lm(y ~ x1 + x2 + x3, data = dat)
summary(fit_full)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1143 -0.6551 -0.1352  0.6916  2.2288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1453     0.1122  28.042  < 2e-16 ***
x1            1.4916     0.4759   3.134  0.00245 **
x2           -0.9448     0.4834  -1.954  0.05433 .
x3            0.9778     0.1190   8.213  4.3e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 76 degrees of freedom
Multiple R-squared:  0.5447,    Adjusted R-squared:  0.5267
F-statistic: 30.31 on 3 and 76 DF,  p-value: 5.434e-13
```

36.36 Inspect predictor correlations

```
round(cor(dat[, c("x1", "x2", "x3")]), 3)
```

```
       x1    x2     x3
x1  1.000 0.966 -0.012
x2  0.966 1.000  0.021
x3 -0.012 0.021  1.000
```

36.37 Compute VIF values

```
vif_manual <- function(model) {
  X <- model.matrix(model)[, -1, drop = FALSE]
  out <- numeric(ncol(X))
  names(out) <- colnames(X)
  for (j in seq_len(ncol(X))) {
    fit_j <- lm(X[, j] ~ X[, -j, drop = FALSE])
    r2_j <- summary(fit_j)$r.squared
    out[j] <- 1 / (1 - r2_j)
  }
  out
}

vif_manual(fit_full)
```

```
       x1        x2        x3
15.229322 15.233765  1.016988
```

36.38 Compare with simpler models

```
fit_small <- lm(y ~ x1 + x3, data = dat)
fit_alt <- lm(y ~ x2 + x3, data = dat)

summary(fit_small)
```

```
Call:
lm(formula = y ~ x1 + x3, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95252 -0.63902 -0.09662  0.57804  2.41530

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1579     0.1140  27.698  < 2e-16 ***
x1            0.5925     0.1242   4.772 8.51e-06 ***
x3            0.9478     0.1202   7.885 1.69e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.019 on 77 degrees of freedom
Multiple R-squared:  0.5218,    Adjusted R-squared:  0.5094
F-statistic: 42.01 on 2 and 77 DF,  p-value: 4.623e-13
```

```r
summary(fit_alt)
```

```
Call:
lm(formula = y ~ x2 + x3, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.05628 -0.65766 -0.08436  0.55012  2.57271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1667     0.1182  26.792  < 2e-16 ***
x2            0.5197     0.1308   3.974 0.000158 ***
x3            0.9302     0.1247   7.462  1.1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.057 on 77 degrees of freedom
Multiple R-squared:  0.4858,    Adjusted R-squared:  0.4725
F-statistic: 36.38 on 2 and 77 DF,  p-value: 7.541e-12
```

```r
anova(fit_small, fit_full)
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     77 80.019
2     76 76.189  1    3.8293 3.8198 0.05433 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

36.39 R Demonstration With Automatic Selection

36.40 Use AIC-based stepwise selection

```r
fit_null <- lm(y ~ 1, data = dat)
fit_scope <- lm(y ~ x1 + x2 + x3, data = dat)

step_forward <- step(fit_null,
                     scope = list(lower = fit_null, upper = fit_scope),
                     direction = "forward",
                     trace = 0)

step_backward <- step(fit_scope, direction = "backward", trace = 0)

summary(step_forward)
```

```
Call:
lm(formula = y ~ x3 + x1 + x2, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1143 -0.6551 -0.1352  0.6916  2.2288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   3.1453     0.1122  28.042  < 2e-16 ***
x3            0.9778     0.1190   8.213  4.3e-12 ***
x1            1.4916     0.4759   3.134  0.00245 **
x2           -0.9448     0.4834  -1.954  0.05433 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 76 degrees of freedom
Multiple R-squared:  0.5447,    Adjusted R-squared:  0.5267
F-statistic: 30.31 on 3 and 76 DF,  p-value: 5.434e-13
```

```r
summary(step_backward)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1143 -0.6551 -0.1352  0.6916  2.2288

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1453     0.1122  28.042  < 2e-16 ***
x1            1.4916     0.4759   3.134  0.00245 **
x2           -0.9448     0.4834  -1.954  0.05433 .
x3            0.9778     0.1190   8.213  4.3e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.001 on 76 degrees of freedom
Multiple R-squared:  0.5447,    Adjusted R-squared:  0.5267
F-statistic: 30.31 on 3 and 76 DF,  p-value: 5.434e-13
```

36.41 Compare AIC, BIC, and adjusted R squared

```r
model_summary <- function(model) {
  c(
    AIC = AIC(model),
```

```
    BIC = BIC(model),
    adj_R2 = summary(model)$adj.r.squared
  )
}

rbind(
  small = model_summary(fit_small),
  alt = model_summary(fit_alt),
  full = model_summary(fit_full),
  step_forward = model_summary(step_forward),
  step_backward = model_summary(step_backward)
)
```

```
                  AIC      BIC    adj_R2
small         235.0488 244.5769 0.5093799
alt           240.8499 250.3780 0.4724811
full          233.1258 245.0359 0.5267118
step_forward  233.1258 245.0359 0.5267118
step_backward 233.1258 245.0359 0.5267118
```

36.42 Example With Polynomial Terms and Centering

```
set.seed(456)
x <- seq(1, 20, length.out = 60)
y2 <- 5 + 1.2 * x - 0.05 * x^2 + rnorm(length(x), sd = 2)

dat2 <- data.frame(y = y2, x = x, xc = x - mean(x))

fit_poly_raw <- lm(y ~ x + I(x^2), data = dat2)
fit_poly_center <- lm(y ~ xc + I(xc^2), data = dat2)

summary(fit_poly_raw)
```

```
Call:
lm(formula = y ~ x + I(x^2), data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-4.2389 -1.1964  0.1053  1.0663  4.3306

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.618794   0.933843   6.017 1.35e-07 ***
x            1.106743   0.203829   5.430 1.21e-06 ***
I(x^2)      -0.044706   0.009444  -4.734 1.50e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.034 on 57 degrees of freedom
Multiple R-squared:  0.3813,    Adjusted R-squared:  0.3596
F-statistic: 17.56 on 2 and 57 DF,  p-value: 1.141e-06
```

```
summary(fit_poly_center)
```

```

Call:
lm(formula = y ~ xc + I(xc^2), data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2389 -1.1964  0.1053  1.0663  4.3306

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.310735   0.394001  31.245  < 2e-16 ***
xc           0.167912   0.047087   3.566 0.000742 ***
I(xc^2)     -0.044706   0.009444  -4.734  1.5e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.034 on 57 degrees of freedom
Multiple R-squared:  0.3813,    Adjusted R-squared:  0.3596
F-statistic: 17.56 on 2 and 57 DF,  p-value: 1.141e-06
```

36.43 Compare collinearity before and after centering

```r
cor(cbind(dat2$x, dat2$x^2))
```

```
          [,1]      [,2]
[1,] 1.0000000 0.9729506
[2,] 0.9729506 1.0000000
```

```r
cor(cbind(dat2$xc, dat2$xc^2))
```

```
              [,1]          [,2]
[1,]  1.000000e+00 -1.214134e-16
[2,] -1.214134e-16  1.000000e+00
```

36.44 Interpreting Software Output

Useful commands in R include:

- `cor()` for predictor correlations;
- `summary(lm(...))` for coefficient estimates and overall fit;
- `AIC()` and `BIC()` for model comparison;
- `step()` for automated exploratory selection;
- `model.matrix()` for checking the design matrix.

Students should remember that software can rank candidate models, but interpretation and final justification must still come from statistical reasoning.

36.45 A Practical Collinearity and Selection Workflow

A sensible workflow is:

- inspect the scientific role of each predictor;
- examine predictor correlations and VIFs;
- identify redundant or unstable terms;
- compare a small set of plausible models;
- avoid blind selection when interactions or scientific controls are important;
- interpret the final model in light of both diagnostics and the original research question.

36.46 In-Class Discussion Questions

1. Why can a model have a significant overall F test but weak individual t tests?
2. Why does high collinearity affect coefficient interpretation more than fitted values?
3. What are the dangers of relying entirely on stepwise selection?
4. In what situations is a larger model worth keeping even if some terms are not individually significant?

36.47 Practice Problems

36.48 Conceptual

1. Explain multicollinearity in your own words.
2. Explain why VIF is connected to regressing one predictor on the others.
3. Explain the difference between a model selected for interpretation and a model selected for prediction.

36.49 Computational

Suppose a predictor x_j has

$$R_j^2 = 0.90$$

when regressed on the remaining predictors.

1. Compute the VIF.
2. Compute the tolerance.
3. Explain what these values imply.

Now suppose a model has

- AIC = 210 for Model A,
- AIC = 205 for Model B,
- BIC = 220 for Model A,
- BIC = 225 for Model B.

1. Which model is preferred by AIC?
2. Which model is preferred by BIC?
3. What does this tell you about the fit-complexity tradeoff?

36.50 Model-Building Problem

You are comparing two models:

- Model 1 contains age, income, and education;
- Model 2 contains age, income, education, and an age-by-income interaction.

1. Why should the interaction model usually retain the main effects?
2. What statistical and scientific questions would guide whether the interaction should remain?
3. Why is it not enough to choose only by the smallest p-value?

36.51 Suggested Homework

Complete the following tasks:

- fit a multiple regression model with at least four predictors;
- compute predictor correlations and VIFs;
- identify any signs of multicollinearity and explain their consequences;
- compare several candidate models using adjusted R^2, AIC, and BIC;
- write a short justification for your preferred final model, explicitly discussing both statistical and subject-matter considerations.

36.52 Summary

In this week, we studied multicollinearity, variable selection, and model building in multiple regression.

We emphasized that:

- collinearity can make coefficients unstable even when overall fit is good;
- VIFs and related tools help diagnose this problem;
- automatic selection procedures are useful but limited;
- model comparison criteria such as adjusted R^2, AIC, and BIC serve different goals;
- good model building requires both statistical judgment and scientific context.

Next week, a natural continuation is to study formal general linear hypotheses, estimability, and matrix-based inference in more depth, or to move toward regularization methods such as ridge regression and lasso if the course is oriented toward modern regression.

36.53 Appendix: Compact Formula Summary

Variance inflation factor:

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

Tolerance:

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

Adjusted R^2:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)}.$$

AIC:

$$\mathrm{AIC} = -2\log L + 2k.$$

BIC:

$$\mathrm{BIC} = -2\log L + k\log n.$$

Typical model-building questions:

- Are the predictors interpretable?
- Are some predictors redundant?
- Is collinearity harming coefficient stability?
- Does the model answer the scientific question?
- Will the model likely generalize beyond the observed data?

37 Week 9: General Linear Hypotheses, Contrasts, and Estimability

In this week, we bring together many earlier ideas into the general framework of linear inference in regression models. The focus is on linear functions of parameters, contrasts, general linear hypotheses, and the important issue of estimability. This week helps students move from coefficient-by-coefficient thinking to a broader matrix-based understanding of inference in linear models.

37.1 Learning Objectives

By the end of this week, students should be able to:

- define a linear function of regression parameters;
- explain what a contrast is and why contrasts are useful;
- formulate general linear hypotheses in matrix form;
- carry out inference for linear combinations of parameters;
- explain the meaning of estimability in linear models;
- distinguish between full-rank and rank-deficient settings;
- interpret software output for linear hypothesis tests and contrasts.

37.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on linear functions of parameters
  - general linear hypotheses
  - estimability and rank-deficient models
- Montgomery, Peck, and Vining:
  - sections on tests for combinations of parameters
  - qualitative predictors and comparisons among means
  - extra sum of squares and related inference

37.3 Why This Week Matters

In earlier weeks, we tested individual coefficients and compared nested models.

But many important questions in regression are not of the form:

- is one coefficient equal to zero?

Instead, they are questions such as:

- are two slopes equal?
- is the average of two treatment effects equal to a third?
- are all group means the same?
- is a certain interaction effect absent?
- does a linear combination of parameters equal a specified value?

These are all questions about linear functions of the parameter vector.

This week provides the general language for expressing and testing such questions.

37.4 Review of the Linear Model

Recall the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta, \qquad \mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

Under the normal linear model,

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n).$$

When $\mathbf{X}$ has full column rank, the ordinary least squares estimator is

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

We also know that

$$\widehat{\beta} \sim N_p(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$

This makes linear combinations of $\widehat{\beta}$ especially important.

37.5 Linear Functions of Parameters

A **linear function** of the parameter vector is any quantity of the form

$$a^\top \beta,$$

where a is a fixed vector.

Examples include:

- a single coefficient, such as β_2;
- the sum $\beta_1 + \beta_2$;
- the difference $\beta_3 - \beta_4$;
- the average $\frac{1}{2}(\beta_2 + \beta_3)$.

These are all linear in β.

The corresponding estimator is

$$a^\top \widehat{\beta}.$$

37.6 Distribution of a Linear Function

Since $\widehat{\beta}$ is multivariate normal, any linear function of it is normal.

Thus,

$$a^\top \widehat{\beta} \sim N(a^\top \beta, \sigma^2 a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a)\,.$$

If σ^2 is estimated by

$$\widehat{\sigma}^2 = \frac{\text{SSE}}{n-p},$$

then

$$\frac{a^\top \widehat{\beta} - a^\top \beta}{\widehat{\sigma}\sqrt{a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a}} \sim t_{n-p}.$$

So the familiar single-coefficient t test is just a special case of inference for a linear combination.

37.7 Contrasts

A **contrast** is a special linear combination whose coefficients sum to zero.

That is, a linear function

$$\sum_{j=1}^{k} c_j \mu_j$$

is a contrast if

$$\sum_{j=1}^{k} c_j = 0.$$

Contrasts are especially important when comparing treatment means or group means.

Examples:

- $\mu_1 - \mu_2$;
- $\mu_1 - \frac{\mu_2+\mu_3}{2}$;
- $\mu_1 + \mu_2 - \mu_3 - \mu_4$.

Contrasts measure relative differences rather than overall level.

37.8 Why Contrasts Are Useful

Suppose we have several group means.

Questions like the following are naturally expressed as contrasts:

- is treatment A different from treatment B?
- is the control mean equal to the average of two treatment means?
- are two pairs of means equally separated?

Contrasts give a flexible and interpretable way to express these comparisons.

They are central in ANOVA and regression models with categorical predictors.

37.9 Contrasts in a Regression Framework

Suppose a factor with three levels is coded with an intercept and two indicator variables.

Then the regression coefficients determine the group means, and comparisons among means can be written as linear functions of β.

So the contrast framework and the regression framework are fully connected.

This is one of the reasons the general linear model is so powerful.

37.10 Confidence Intervals for Linear Functions

For a linear function $a^\top \beta$, a confidence interval is

$$a^\top \widehat{\beta} \pm t_{1-\alpha/2,\, n-p} \widehat{\sigma} \sqrt{a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a}.$$

This formula allows us to make inference for any estimable linear combination, not just a single coefficient.

37.11 Testing a Single Linear Function

To test

$$H_0 : a^\top \beta = c,$$

we use the test statistic

$$T = \frac{a^\top \widehat{\beta} - c}{\widehat{\sigma} \sqrt{a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a}}.$$

Under H_0,

$$T \sim t_{n-p}.$$

Again, this is just the general version of the usual coefficient t test.

37.12 General Linear Hypotheses

A **general linear hypothesis** has the form

$$H_0 : \mathbf{C}\beta = \mathbf{d},$$

where:

- $\mathbf{C}$ is an $r \times p$ matrix;
- $\mathbf{d}$ is an $r \times 1$ vector;
- r is the number of linear restrictions.

This framework includes many important hypothesis tests as special cases.

37.13 Examples of General Linear Hypotheses

Examples include:

- testing one coefficient:
$$H_0 : \beta_2 = 0;$$
- testing equality of two coefficients:
$$H_0 : \beta_2 - \beta_3 = 0;$$
- testing two restrictions simultaneously:
$$H_0 : \begin{cases} \beta_2 = 0, \\ \beta_3 = 0; \end{cases}$$
- testing whether three group means are equal.

All of these can be written in the form

$$\mathbf{C}\beta = \mathbf{d}.$$

37.14 F Test for a General Linear Hypothesis

Under the normal linear model, the hypothesis

$$H_0 : \mathbf{C}\beta = \mathbf{d}$$

is tested by

$$F = \frac{(\mathbf{C}\widehat{\beta} - \mathbf{d})^\top \left[\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top\right]^{-1} (\mathbf{C}\widehat{\beta} - \mathbf{d})/r}{\widehat{\sigma}^2}.$$

Under H_0,

$$F \sim F_{r,\, n-p}.$$

This is the general matrix form of the regression F test.

37.15 Connection With Nested Models

The general linear hypothesis test is equivalent to comparing a reduced model and a full model when the reduced model is obtained by imposing linear restrictions.

So the extra sum of squares F test from Week 4 is a special case of the general linear hypothesis test.

This is an important unifying idea.

37.16 When r = 1

If there is only one restriction, then the general F test reduces to the square of a t test.

That is, when $r = 1$,

$$F = T^2.$$

So single-parameter inference and multi-parameter inference are part of the same framework.

37.17 Matrix Formulation of Contrasts

If we are interested in several contrasts at once, we can stack them into a matrix.

For example, if we want to test

$$\beta_2 - \beta_3 = 0 \qquad \text{and} \qquad \beta_3 - \beta_4 = 0,$$

then we may write

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \qquad \mathbf{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This turns several related coefficient comparisons into one unified hypothesis.

37.18 Estimability

So far, we have mostly assumed that $\mathbf{X}$ has full column rank.

But in some important cases, the design matrix is rank-deficient. Then the parameter vector itself is not uniquely identifiable.

In such settings, we ask a more refined question:

Which linear functions of β can still be estimated uniquely from the model?

This leads to the concept of **estimability**.

37.19 Why Estimability Is Needed

Suppose two different parameter vectors, say β and β^*, produce the same mean vector:

$$\mathbf{X}\beta = \mathbf{X}\beta^*.$$

Then the data cannot distinguish between these two parameter vectors.

So a particular coefficient may not be uniquely meaningful.

However, some combinations of coefficients may still be uniquely determined by the mean structure. Those combinations are estimable.

37.20 Definition of Estimability

A linear function

$$a^\top \beta$$

is **estimable** if there exists a vector t such that

$$a^\top = t^\top \mathbf{X}.$$

Equivalently, a must lie in the row space of $\mathbf{X}$.

This condition ensures that the target quantity depends only on the mean vector $\mathbf{X}\beta$, and not on the particular parameterization used.

37.21 Interpretation of Estimability

Estimability means that the quantity is determined by the model's observable mean structure.

If a linear function is not estimable, then different parameter vectors producing the same fitted mean can give different values of that function.

So no unbiased linear estimator can uniquely recover it.

37.22 Example With a Factor Model

Suppose we write a one-way mean model as

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

for groups $i = 1, \dots, g$.

If we include all group indicators together with an intercept, then the parameters are not uniquely identified, because adding a constant to all τ_i and subtracting it from μ gives the same mean structure.

In this case:

- μ alone is not uniquely defined;
- τ_i alone is not uniquely defined;
- but differences such as $\tau_i - \tau_j$ are estimable.

This is a classic example.

37.23 Estimable Functions in Rank-Deficient Models

Even when the coefficient vector is not unique, fitted values are still unique, provided we project onto the column space of $\mathbf{X}$.

Likewise, every estimable linear function has a unique value determined by the model.

So regression analysis in rank-deficient settings often focuses on estimable functions rather than on individual raw coefficients.

37.24 Parameterization Matters for Coefficients, but Not for Estimable Functions

A factor can be parameterized in several ways:

- treatment coding;
- sum-to-zero coding;
- cell-means coding.

The individual coefficients change across parameterizations.

But meaningful estimable comparisons, such as differences between group means, do not depend on the coding scheme.

This is a key conceptual lesson.

37.25 Contrasts and Estimability

In many ANOVA-type models, contrasts among means are estimable even when the raw parameter vector is overparameterized.

This is one reason contrasts are so central: they often represent the scientifically meaningful and estimable quantities.

37.26 Least Squares in Rank-Deficient Models

When $\mathbf{X}$ is rank-deficient, the normal equations do not yield a unique coefficient vector.

Different generalized inverse solutions may produce different coefficient vectors.

However:

- the fitted values are unique;
- the residual sum of squares is unique;
- estimable linear functions have unique least squares estimates.

So the inferential target should be framed carefully.

37.27 Generalized Inverse View

A generalized inverse of $\mathbf{X}^\top\mathbf{X}$ can be used to write one least squares solution.

This leads to expressions similar to the full-rank case, but students should remember that not every coefficient itself is uniquely meaningful.

What matters most is whether the function of interest is estimable.

37.28 Software and Estimability

Modern software often handles rank deficiency automatically.

It may:

- drop aliased columns;
- report coefficients as not estimable;
- use a default parameterization that makes the fit identifiable.

Students should not interpret every reported coefficient mechanically. They should understand the underlying estimable structure.

37.29 Worked Example With Equality of Slopes

Suppose we fit a regression model with two predictors and want to test whether their coefficients are equal:

$$H_0 : \beta_1 = \beta_2.$$

This can be written as

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

so

$$a^\top = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}.$$

The corresponding estimate is

$$\hat{\beta}_1 - \hat{\beta}_2,$$

and inference follows from the general linear function formula.

37.30 Worked Example With Group Means

Suppose three group means are

$$\mu_A, \quad \mu_B, \quad \mu_C.$$

If we want to compare group A with the average of groups B and C, we consider the contrast

$$\mu_A - \frac{1}{2}\mu_B - \frac{1}{2}\mu_C.$$

The coefficients sum to zero, so this is a contrast.

This question arises naturally in designed experiments and treatment comparisons.

37.31 R Demonstration With a Linear Hypothesis

37.32 Fit a multiple regression model

```r
set.seed(123)
n <- 60
x1 <- rnorm(n)
x2 <- rnorm(n)
x3 <- rnorm(n)
y <- 2 + 1.5 * x1 + 1.5 * x2 - 0.8 * x3 + rnorm(n, sd = 1)

dat <- data.frame(y = y, x1 = x1, x2 = x2, x3 = x3)
fit <- lm(y ~ x1 + x2 + x3, data = dat)
summary(fit)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3089 -0.6843 -0.1389  0.5141  2.1748

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9643     0.1200  16.362  < 2e-16 ***
x1            1.5324     0.1363  11.243 5.53e-16 ***
x2            1.5535     0.1380  11.260 5.21e-16 ***
x3           -0.6660     0.1171  -5.685 4.91e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9259 on 56 degrees of freedom
Multiple R-squared:  0.8465,    Adjusted R-squared:  0.8382
F-statistic: 102.9 on 3 and 56 DF,  p-value: < 2.2e-16
```

37.33 Test whether two coefficients are equal

```
b <- coef(fit)
V <- vcov(fit)

a <- c(0, 1, -1, 0)
est <- sum(a * b)
se <- sqrt(t(a) %*% V %*% a)
t_stat <- est / se
df <- df.residual(fit)
p_value <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)

c(estimate = est, se = se, t = t_stat, p_value = p_value)
```

```
   estimate          se           t     p_value
-0.02112118  0.20603078 -0.10251470  0.91871436
```

37.34 Confidence interval for a linear combination

```
tcrit <- qt(0.975, df = df)
c(lower = est - tcrit * se,
  estimate = est,
  upper = est + tcrit * se)
```

```
      lower    estimate       upper
-0.43385043 -0.02112118  0.39160806
```

37.35 Test two restrictions simultaneously

```
Cmat <- rbind(
  c(0, 1, -1, 0),
  c(0, 0, 1, 1)
)
dvec <- c(0, 0)
```

```r
Cb_minus_d <- Cmat %*% b - dvec
middle <- solve(Cmat %*% V %*% t(Cmat))
r <- nrow(Cmat)

F_stat <- as.numeric(t(Cb_minus_d) %*% middle %*% Cb_minus_d / r)
p_F <- pf(F_stat, df1 = r, df2 = df, lower.tail = FALSE)

c(F_stat = F_stat, p_value = p_F)
```

```
      F_stat      p_value
1.967284e+01 3.379376e-07
```

37.36 Demonstration With a Factor and Contrasts

```r
set.seed(456)
group <- factor(rep(c("A", "B", "C"), each = 12))
mu <- c(A = 10, B = 13, C = 15)
y2 <- mu[group] + rnorm(length(group), sd = 2)

dat2 <- data.frame(y = y2, group = group)
fit_group <- lm(y ~ group, data = dat2)
summary(fit_group)
```

```
Call:
lm(formula = y ~ group, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3815 -1.6167  0.1576  1.9721  3.6541

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.0948     0.6795  14.857 3.56e-16 ***
groupB        3.8520     0.9609   4.009 0.000328 ***
groupC        4.8824     0.9609   5.081 1.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.354 on 33 degrees of freedom
Multiple R-squared:  0.4651,    Adjusted R-squared:  0.4326
F-statistic: 14.34 on 2 and 33 DF,  p-value: 3.288e-05
```

```r
model.matrix(fit_group)[1:8, ]
```

```
  (Intercept) groupB groupC
1           1      0      0
2           1      0      0
3           1      0      0
4           1      0      0
5           1      0      0
6           1      0      0
7           1      0      0
8           1      0      0
```

37.37 Estimate the contrast A minus average of B and C

```r
b2 <- coef(fit_group)
V2 <- vcov(fit_group)

# Under treatment coding:
# mean_A = beta0
# mean_B = beta0 + beta_B
# mean_C = beta0 + beta_C
# contrast = mean_A - (mean_B + mean_C)/2 = -(beta_B + beta_C)/2

a2 <- c(0, -1/2, -1/2)
est2 <- sum(a2 * b2)
se2 <- sqrt(t(a2) %*% V2 %*% a2)
t2 <- est2 / se2
df2 <- df.residual(fit_group)
p2 <- 2 * pt(abs(t2), df = df2, lower.tail = FALSE)

c(estimate = est2, se = se2, t = t2, p_value = p2)
```

```
     estimate            se             t       p_value
-4.367214e+00  8.321888e-01 -5.247864e+00  8.879243e-06
```

37.38 Example of rank deficiency

```
X_bad <- cbind(1, diag(3))
X_bad <- rbind(X_bad, X_bad)
colnames(X_bad) <- c("Intercept", "G1", "G2", "G3")

qr(X_bad)$rank
```

```
[1] 3
```

```
ncol(X_bad)
```

```
[1] 4
```

```
X_bad
```

```
     Intercept G1 G2 G3
[1,]         1  1  0  0
[2,]         1  0  1  0
[3,]         1  0  0  1
[4,]         1  1  0  0
[5,]         1  0  1  0
[6,]         1  0  0  1
```

37.39 Interpreting Software Output

Useful commands in R include:

- `coef()` for estimated coefficients;
- `vcov()` for the covariance matrix of coefficient estimates;
- `model.matrix()` for inspecting the design matrix;
- `anova()` for nested-model versions of linear hypothesis tests.

Even when software provides default hypothesis tests, students should learn to express the hypothesis itself in matrix form. That is often the most important conceptual step.

37.40 A Practical Workflow for Linear Hypotheses

A useful workflow is:

- identify the scientific question;
- express the target as a linear function or a set of linear restrictions;
- write down the vector a or matrix $\mathbf{C}$;
- compute the estimate and its standard error;
- interpret the result on the original scientific scale.

This approach makes regression inference more flexible and more transparent.

37.41 In-Class Discussion Questions

1. Why is a contrast defined by coefficients summing to zero?
2. Why are some functions estimable even when the full parameter vector is not uniquely identifiable?
3. Why is the general linear hypothesis framework more powerful than testing coefficients one by one?
4. Why should interpretation focus on estimable functions rather than arbitrary parameterizations?

37.42 Practice Problems

37.43 Conceptual

1. Explain the difference between a coefficient and a general linear function of coefficients.
2. Explain why treatment comparisons are often naturally expressed as contrasts.
3. Explain estimability in your own words.

37.44 Computational

Suppose

$$\widehat{\beta} = \begin{bmatrix} 4 \\ 1 \\ -2 \end{bmatrix}, \qquad \widehat{\mathrm{Var}}(\widehat{\beta}) = \begin{bmatrix} 0.5 & 0.1 & 0.0 \\ 0.1 & 0.4 & 0.2 \\ 0.0 & 0.2 & 0.6 \end{bmatrix}.$$

Let

$$a = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

1. Compute the estimate of $a^\top \beta$.
2. Compute its estimated variance.
3. Write the corresponding t statistic for testing whether $a^\top \beta = 0$.

37.45 Hypothesis-Matrix Problem

Write the matrix $\mathbf{C}$ and vector $\mathbf{d}$ for each hypothesis:

1. $H_0 : \beta_2 = \beta_3$;
2. $H_0 : \beta_2 = 0$ and $\beta_4 = 0$;
3. $H_0 : \beta_2 + \beta_3 - 2\beta_4 = 1$.

37.46 Suggested Homework

Complete the following tasks:

- fit a regression model and test at least two nontrivial linear combinations of coefficients;
- write each scientific question first in words and then in matrix form;
- compute a confidence interval for one contrast of interest;
- fit a model with a factor and interpret at least two group comparisons as contrasts;
- write a short reflection explaining why estimability matters in categorical models.

37.47 Summary

In this week, we studied the general framework of linear inference in regression.

We emphasized that:

- many meaningful questions involve linear combinations of parameters rather than single coefficients;
- contrasts are important special cases of linear functions;
- general linear hypotheses unify many common tests;
- estimability determines which parameter functions are uniquely learnable from the model;

- meaningful inference should focus on estimable functions, especially in rank-deficient settings.

Next week, a natural continuation is to move into analysis of covariance, one-way and two-way ANOVA as special cases of the linear model, or to extend toward generalized least squares and correlated errors, depending on the course emphasis.

37.48 Appendix: Compact Formula Summary

Linear function of parameters:

$$a^\top \beta.$$

Estimated variance of its estimator:

$$\widehat{\mathrm{Var}}(a^\top \hat{\beta}) = \hat{\sigma}^2 a^\top (\mathbf{X}^\top \mathbf{X})^{-1} a.$$

General linear hypothesis:

$$H_0 : \mathbf{C}\beta = \mathbf{d}.$$

General F statistic:

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

Estimability condition:

$$a^\top \text{ is estimable if } a^\top = t^\top \mathbf{X} \text{ for some } t.$$

38 Week 10: One-Way ANOVA, Two-Way ANOVA, and ANCOVA in the Linear Model Framework

In this week, we study one-way ANOVA, two-way ANOVA, and analysis of covariance as special cases of the general linear model. The main goal is to show that these topics are not separate from regression, but are natural modelling frameworks within the same matrix-based system. This helps students unify mean comparisons, factor effects, covariate adjustment, and interaction analysis under one common language.

38.1 Learning Objectives

By the end of this week, students should be able to:

- explain how one-way ANOVA is a special case of the linear model;
- formulate and interpret two-way ANOVA models with and without interaction;
- explain the role of a covariate in ANCOVA;
- distinguish between main effects and interaction effects in factorial models;
- interpret ANOVA and ANCOVA models using regression notation and software output;
- compare group means appropriately after adjusting for covariates.

38.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on analysis of variance
  - analysis of covariance
  - linear model treatment of factor and covariate effects
- Montgomery, Peck, and Vining:
  - sections on qualitative predictors

- ANOVA-style linear models
- ANCOVA and adjusted comparisons

38.3 Why These Topics Belong Together

Students often first encounter ANOVA and regression as if they were separate techniques.

But in fact, they are part of the same linear modelling framework.

For example:

- one-way ANOVA compares group means;
- two-way ANOVA studies the effects of two factors and possibly their interaction;
- ANCOVA compares groups while adjusting for a continuous covariate.

All of these can be written as linear models using an appropriate design matrix.

This week emphasizes that unification.

38.4 Review of the General Linear Model

Recall the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta, \qquad \mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n.$$

The content of the model depends on how we construct $\mathbf{X}$.

If $\mathbf{X}$ contains continuous predictors, we obtain regression models.

If $\mathbf{X}$ contains indicator variables for factor levels, we obtain ANOVA-type models.

If $\mathbf{X}$ contains both factor indicators and continuous covariates, we obtain ANCOVA-type models.

38.5 One-Way ANOVA

Suppose observations are divided into g groups, and we want to compare their means.

One-way ANOVA can be written as

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1, \dots, g,$$

where:

- μ_i is the mean of group i;
- ε_{ij} are independent errors with mean zero and common variance σ^2.

This is sometimes called the cell-means model.

38.6 Alternative Parameterization of One-Way ANOVA

Another common parameterization is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, \dots, g,$$

where:

- μ is an overall baseline level;
- τ_i is the effect of group i.

To identify the model, we usually impose a constraint such as

$$\sum_{i=1}^{g} \tau_i = 0.$$

Different coding schemes lead to different coefficient interpretations, but the fitted group means are the same.

38.7 Null Hypothesis in One-Way ANOVA

The usual one-way ANOVA null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g.$$

Equivalently, in the effect parameterization, this is

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$

subject to the model constraints.

This is a general linear hypothesis and can be tested by an F test.

38.8 One-Way ANOVA as Regression With Indicators

Suppose there are three groups: A, B, and C.

Using treatment coding with group A as the reference group, we may write

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

where:

- $z_{1i} = 1$ if observation i is in group B and 0 otherwise;
- $z_{2i} = 1$ if observation i is in group C and 0 otherwise.

Then:

- mean of group A: β_0;
- mean of group B: $\beta_0 + \beta_1$;
- mean of group C: $\beta_0 + \beta_2$.

Thus one-way ANOVA is simply a regression model with categorical predictors.

38.9 ANOVA Table Interpretation

For one-way ANOVA, the total variation is decomposed into:

- variation between groups;
- variation within groups.

The standard ANOVA table compares the mean square for groups to the mean square for error.

This is an F test for whether the group means are all equal.

Within the linear model framework, this is just a comparison of a reduced intercept-only model and a fuller model that includes group indicators.

38.10 Two-Way ANOVA

Now suppose there are two factors:

- factor A with levels $i = 1, \dots, a$;
- factor B with levels $j = 1, \dots, b$.

A general two-way ANOVA model with interaction is

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

Here:

- 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 for the combination of levels (i, j).

38.11 Main Effects in Two-Way ANOVA

A **main effect** describes how the mean response changes across levels of one factor, averaging over the levels of the other factor, when the interaction is absent or appropriately interpreted.

For example:

- factor A main effect asks whether changing the level of A shifts the mean response;
- factor B main effect asks the analogous question for B.

However, if interaction is strong, main effects must be interpreted with care.

38.12 Interaction in Two-Way ANOVA

Interaction means that the effect of one factor depends on the level of the other factor.

If there is no interaction, the mean structure is additive:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

If interaction is present, the difference among levels of factor A changes across levels of factor B, or vice versa.

Graphically, in an interaction plot, no interaction corresponds roughly to parallel profiles.

38.13 Null Hypotheses in Two-Way ANOVA

Common hypotheses include:

- no main effect of factor A;
- no main effect of factor B;
- no interaction between A and B.

Each of these is a general linear hypothesis and can be tested with an F statistic by comparing appropriate reduced and full models.

38.14 Importance of Testing Interaction First

In practice, it is often wise to assess the interaction term before making strong statements about main effects.

Why?

Because if interaction is present, the effect of factor A is not the same at all levels of factor B, and vice versa.

So a marginal main-effect interpretation may be misleading.

38.15 Balanced and Unbalanced Designs

A design is **balanced** if each factor combination has the same number of observations.

Balanced designs have many nice algebraic properties:

- sums of squares decompose cleanly;
- effects are more orthogonal;
- interpretations are often simpler.

In unbalanced designs, sums of squares and hypothesis tests depend more strongly on the model specification and coding choices.

Students should be aware that real data are often unbalanced.

38.16 Two-Way ANOVA as Regression

Just as in one-way ANOVA, two-way ANOVA can be represented using indicator variables.

The design matrix can include:

- indicators for levels of factor A;
- indicators for levels of factor B;
- products of indicators for interaction terms.

So two-way ANOVA is also just a regression model with categorical predictors and possible interactions.

38.17 Analysis of Covariance

Analysis of covariance, or **ANCOVA**, combines:

- one or more categorical predictors;
- one or more continuous covariates.

A simple ANCOVA model is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

where:

- x_i is a continuous covariate;
- z_i is a group indicator.

This model compares groups while adjusting for the covariate.

38.18 Why ANCOVA Is Useful

Suppose two groups differ in an outcome, but also differ in a background variable related to the outcome.

If we compare the raw group means, the comparison may be confounded by the background variable.

ANCOVA adjusts for the covariate so that the group comparison is made at a common covariate level.

This often improves both fairness of comparison and precision of inference.

38.19 Interpreting an ANCOVA Model

Suppose

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

with $z_i \in \{0, 1\}$.

Then:

- β_1 is the change in the mean response associated with a one-unit increase in x, holding group fixed;
- β_2 is the difference between groups, holding the covariate fixed.

Thus ANCOVA is a regression model in which one of the predictors is a factor.

38.20 Parallel Slopes Assumption

A standard ANCOVA model without interaction assumes the relationship between the response and the covariate has the same slope in all groups.

That is, if

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

then both groups share the same slope β_1.

This is often called the **parallel slopes assumption**.

38.21 ANCOVA With Interaction

If we want to allow the covariate effect to differ by group, we add an interaction:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i.$$

Now the slope in x depends on the group.

This model should be considered when the relationship between the response and the covariate appears different across groups.

38.22 Why the Parallel Slopes Assumption Matters

If the slopes are not truly parallel but we fit a common-slope ANCOVA model, then the adjusted group comparison may be misleading.

So a practical workflow is often:

- fit the interaction model first;
- assess whether the interaction is needed;
- if the interaction is not important, use the simpler common-slope ANCOVA model.

38.23 Adjusted Means

A major idea in ANCOVA is the comparison of **adjusted means**.

These are group means adjusted to a common covariate value, often the overall mean of the covariate.

Adjusted means are useful because they allow group comparison after accounting for systematic covariate differences.

38.24 One-Way ANOVA, Two-Way ANOVA, and ANCOVA as Nested Models

These topics can all be understood through model comparison.

Examples:

- one-way ANOVA compares the intercept-only model to a model with group effects;
- two-way ANOVA compares models with and without main effects or interaction;
- ANCOVA compares models with and without the covariate, group effect, or interaction.

This nested-model view links directly back to the general linear hypothesis framework.

38.25 Post Hoc Comparisons

After a significant ANOVA result, we may wish to compare specific groups.

Examples include:

- all pairwise comparisons;
- treatment versus control;
- preplanned contrasts.

These are again linear functions of model parameters.

So post hoc or planned comparisons fit naturally into the contrast framework from the previous week.

38.26 Interpretation and Caution

In factor models, coefficient interpretation depends on coding.

For example, treatment coding, sum-to-zero coding, and cell-means coding all yield different raw coefficients.

Therefore students should focus on:

- fitted means;
- differences among means;
- interactions;
- clearly defined hypotheses.

These are more stable and meaningful than memorizing one coding-specific coefficient interpretation.

38.27 Worked Example With One-Way ANOVA

Suppose three teaching methods are compared.

A one-way ANOVA model asks whether the mean outcome differs across the three methods.

This can be written either as a group-means model or as a regression with two indicators and an intercept.

The overall F test asks whether all three group means are equal.

If the null is rejected, further contrasts can be used to compare specific methods.

38.28 Worked Example With Two-Way ANOVA

Suppose yield is measured under:

- three fertilizer types;
- two irrigation levels.

A two-way ANOVA model can assess:

- whether fertilizer matters;
- whether irrigation matters;
- whether the effect of fertilizer depends on irrigation.

This is a direct example of main effects and interaction in a factorial design.

38.29 Worked Example With ANCOVA

Suppose two treatment groups are compared on a final score, and baseline score is available as a covariate.

A simple ANCOVA model adjusts the treatment comparison for baseline.

This often provides a more precise and fairer comparison than comparing final means alone.

38.30 R Demonstration With One-Way ANOVA

38.31 Simulate one-way ANOVA data

```r
set.seed(123)
group <- factor(rep(c("A", "B", "C"), each = 12))
mu <- c(A = 10, B = 13, C = 15)
y <- mu[group] + rnorm(length(group), sd = 2)

dat1 <- data.frame(y = y, group = group)
fit1 <- lm(y ~ group, data = dat1)

summary(fit1)
```

```
Call:
lm(formula = y ~ group, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7417 -1.3371 -0.0372  1.3274  3.9969

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3884     0.5465  19.008  < 2e-16 ***
groupB        2.1886     0.7729   2.832  0.00783 **
groupC        4.9800     0.7729   6.443 2.63e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.893 on 33 degrees of freedom
Multiple R-squared:  0.5583,    Adjusted R-squared:  0.5316
F-statistic: 20.86 on 2 and 33 DF,  p-value: 1.393e-06
```

```r
anova(fit1)
```

```
Analysis of Variance Table

Response: y
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
group      2 149.53  74.764  20.858 1.393e-06 ***
Residuals 33 118.28   3.584
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

38.32 Group means and model matrix

```
tapply(dat1$y, dat1$group, mean)
```

```
       A        B        C
10.38836 12.57694 15.36833
```

```
model.matrix(fit1)[1:10, ]
```

```
   (Intercept) groupB groupC
1            1      0      0
2            1      0      0
3            1      0      0
4            1      0      0
5            1      0      0
6            1      0      0
7            1      0      0
8            1      0      0
9            1      0      0
10           1      0      0
```

38.33 Pairwise comparisons through linear contrasts

```
b <- coef(fit1)
V <- vcov(fit1)

# Under treatment coding with A as reference:
# mean_A = beta0
# mean_B = beta0 + beta_groupB
# mean_C = beta0 + beta_groupC
```

```r
# Compare B and C
a <- c(0, 1, -1)
est <- sum(a * b)
se <- sqrt(t(a) %*% V %*% a)
t_stat <- est / se
df <- df.residual(fit1)
p_val <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)

c(estimate = est, se = se, t = t_stat, p_value = p_val)
```

```
     estimate            se             t       p_value
-2.7913915175  0.7729112696 -3.6115290685  0.0009983035
```

38.34 R Demonstration With Two-Way ANOVA

38.35 Simulate factorial data

```r
set.seed(456)
A <- factor(rep(c("Low", "Medium", "High"), each = 16))
B <- factor(rep(rep(c("I1", "I2"), each = 8), times = 3))

mean_table <- matrix(c(
  10, 12,
  13, 15,
  16, 20
), nrow = 3, byrow = TRUE)

mu2 <- numeric(length(A))
for (i in seq_along(mu2)) {
  mu2[i] <- mean_table[which(levels(A) == A[i]), which(levels(B) == B[i])]
}

y2 <- mu2 + rnorm(length(mu2), sd = 1.8)
dat2 <- data.frame(y = y2, A = A, B = B)
```

38.36 Fit additive and interaction models

```
fit2_add <- lm(y ~ A + B, data = dat2)
fit2_int <- lm(y ~ A * B, data = dat2)

summary(fit2_add)
```

```

Call:
lm(formula = y ~ A + B, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7012 -0.9352 -0.0968  1.3177  3.6363

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6318     0.5460  17.641  < 2e-16 ***
ALow          3.1589     0.6687   4.724 2.39e-05 ***
AMedium       6.9782     0.6687  10.435 1.76e-13 ***
BI2           3.2551     0.5460   5.962 3.84e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.891 on 44 degrees of freedom
Multiple R-squared:  0.7669,    Adjusted R-squared:  0.751
F-statistic: 48.26 on 3 and 44 DF,  p-value: 5.777e-14
```

```
summary(fit2_int)
```

```

Call:
lm(formula = y ~ A * B, data = dat2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7466 -0.9687 -0.0766  1.2641  3.5344

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   9.7795     0.6832  14.314  < 2e-16 ***
ALow          2.9039     0.9662   3.005  0.00446 **
AMedium       6.7902     0.9662   7.028 1.33e-08 ***
BI2           2.9598     0.9662   3.063  0.00381 **
ALow:BI2      0.5099     1.3664   0.373  0.71091
AMedium:BI2   0.3760     1.3664   0.275  0.78456
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.932 on 42 degrees of freedom
Multiple R-squared:  0.7677,    Adjusted R-squared:  0.7401
F-statistic: 27.77 on 5 and 42 DF,  p-value: 2.628e-12
```

```
anova(fit2_add, fit2_int)
```

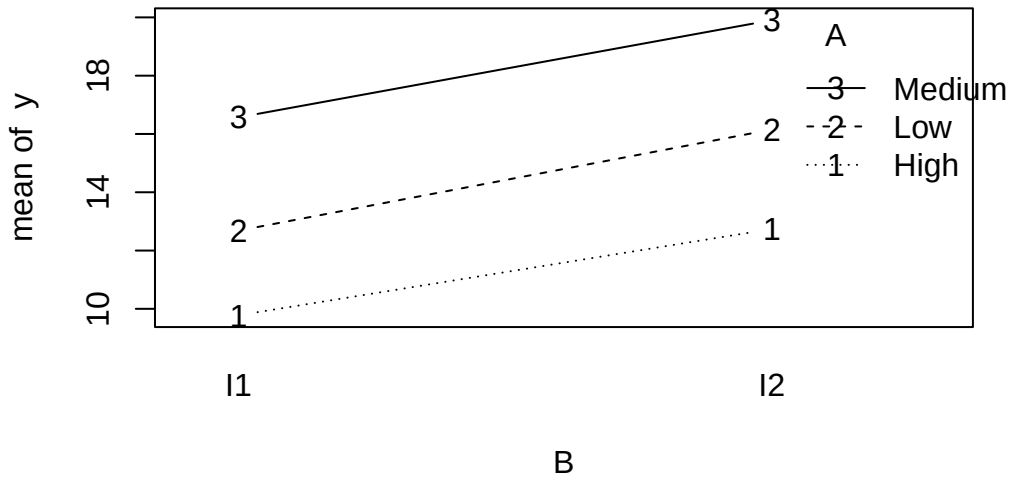
```
Analysis of Variance Table

Model 1: y ~ A + B
Model 2: y ~ A * B
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 157.40
2     42 156.84  2   0.55902 0.0749  0.928
```

```
anova(fit2_int)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          2 390.72 195.360 52.3157 3.958e-12 ***
B          1 127.15 127.149 34.0494 6.855e-07 ***
A:B        2   0.56   0.280  0.0749     0.928
Residuals 42 156.84   3.734
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

38.37 Interaction plot

```r
with(dat2, interaction.plot(x.factor = B, trace.factor = A, response = y,
                            fun = mean, type = "b", legend = TRUE))
```



38.38 R Demonstration With ANCOVA

38.39 Simulate ANCOVA-style data

```r
set.seed(789)
n <- 80
group3 <- factor(rep(c("Control", "Treatment"), each = n / 2))
x <- rnorm(n, mean = 50, sd = 10)
y3 <- 20 + 0.7 * x + ifelse(group3 == "Treatment", 4, 0) + rnorm(n, sd = 4)

dat3 <- data.frame(y = y3, x = x, group = group3)
fit3 <- lm(y ~ x + group, data = dat3)
fit3_int <- lm(y ~ x * group, data = dat3)

summary(fit3)
```

```
Call:
lm(formula = y ~ x + group, data = dat3)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2460  -2.4058   0.4175   2.2633  12.6605

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    21.61302    2.49417   8.665 5.30e-13 ***
x               0.65937    0.05125  12.865  < 2e-16 ***
groupTreatment  4.46782    0.95667   4.670 1.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.094 on 77 degrees of freedom
Multiple R-squared:  0.7591,    Adjusted R-squared:  0.7528
F-statistic: 121.3 on 2 and 77 DF,  p-value: < 2.2e-16
```

```
summary(fit3_int)
```

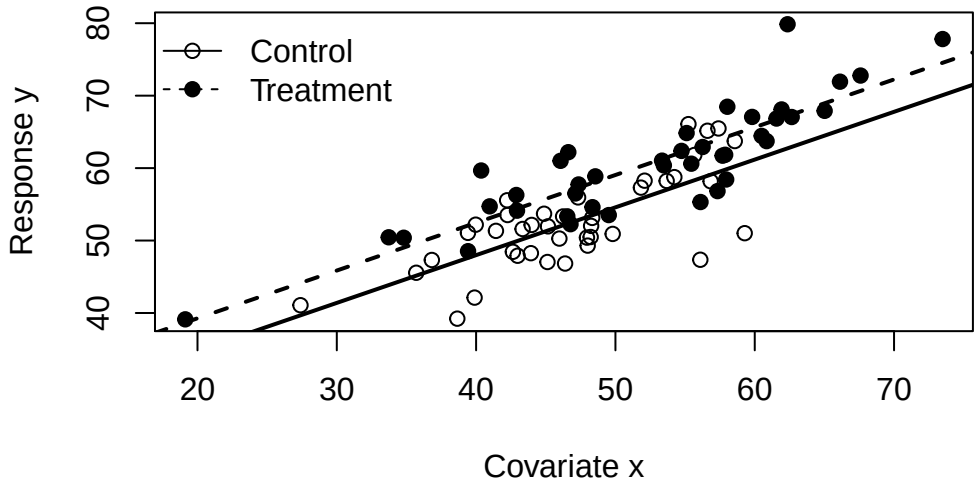
```
Call:
lm(formula = y ~ x * group, data = dat3)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2681  -2.4000   0.4132   2.2713  12.6713

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.499220   4.414169   4.871 5.92e-06 ***
x                 0.661792   0.092900   7.124 5.13e-10 ***
groupTreatment    4.638217   5.520886   0.840    0.403
x:groupTreatment -0.003501   0.111709  -0.031    0.975
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.121 on 76 degrees of freedom
Multiple R-squared:  0.7591,    Adjusted R-squared:  0.7496
F-statistic: 79.81 on 3 and 76 DF,  p-value: < 2.2e-16
```

```r
anova(fit3, fit3_int)
```

```
Analysis of Variance Table

Model 1: y ~ x + group
Model 2: y ~ x * group
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     77 1290.8
2     76 1290.8  1  0.016686 0.001 0.9751
```

38.40 Plot ANCOVA fit

```r
plot(dat3$x, dat3$y,
     pch = ifelse(dat3$group == "Control", 1, 19),
     xlab = "Covariate x",
     ylab = "Response y")
abline(a = coef(fit3)[1], b = coef(fit3)[2], lwd = 2)
abline(a = coef(fit3)[1] + coef(fit3)[3], b = coef(fit3)[2], lwd = 2, lty = 2)
legend("topleft",
       legend = c("Control", "Treatment"),
       pch = c(1, 19),
       lty = c(1, 2),
       bty = "n")
```



38.41 Interpreting Software Output

In R:

- `lm(y ~ group)` fits a one-way ANOVA model;
- `lm(y ~ A * B)` fits a two-way ANOVA model with interaction;
- `lm(y ~ x + group)` fits a common-slope ANCOVA model;
- `lm(y ~ x * group)` fits an ANCOVA model with group-specific slopes.

Students should recognize that these are all regression models differing only in the structure of the design matrix.

38.42 A Practical Workflow

A useful modelling workflow is:

- identify whether predictors are factors, covariates, or both;
- decide whether interactions are scientifically plausible;
- fit the more complete model when appropriate;
- test whether interaction is needed;
- interpret fitted means or adjusted means rather than raw coding-specific coefficients;

- use contrasts for focused follow-up questions.

38.43 In-Class Discussion Questions

1. Why is one-way ANOVA just a regression model with indicator variables?
2. Why should interaction be considered before interpreting main effects in two-way ANOVA?
3. What does ANCOVA adjust for that ordinary group-mean comparison does not?
4. Why can coefficient interpretation depend on coding while fitted means do not?

38.44 Practice Problems

38.45 Conceptual

1. Explain the difference between a one-way ANOVA model and an ANCOVA model.
2. Explain what interaction means in a two-way ANOVA setting.
3. Explain why adjusted means are useful in ANCOVA.

38.46 Computational

Suppose there are three groups A, B, and C, and treatment coding is used with A as the reference group. The fitted model is

$$\hat{Y} = 8 + 2z_1 + 5z_2.$$

1. What is the fitted mean for group A?
2. What is the fitted mean for group B?
3. What is the fitted mean for group C?
4. What is the estimated difference between groups B and C?

Now consider the ANCOVA model

$$\hat{Y} = 12 + 0.6x + 3z,$$

where $z = 1$ for treatment and $z = 0$ for control.

1. Interpret the coefficient of x.
2. Interpret the coefficient of z.
3. Compute the fitted mean when $x = 10$ for control and treatment.

38.47 Model-Comparison Problem

You fit the following two models:

- Model 1: `y ~ x + group`
- Model 2: `y ~ x * group`

1. What extra term is in Model 2?
2. What scientific question does the comparison test?
3. Why would a significant interaction affect the interpretation of the group effect?

38.48 Suggested Homework

Complete the following tasks:

- fit a one-way ANOVA model and interpret the overall F test;
- carry out at least two follow-up contrasts among group means;
- fit a two-way ANOVA model with interaction and interpret the interaction term;
- fit an ANCOVA model and explain the meaning of adjusted group comparison;
- compare an ANCOVA model with and without interaction and explain which one you prefer.

38.49 Summary

In this week, we studied one-way ANOVA, two-way ANOVA, and ANCOVA within the general linear model framework.

We emphasized that:

- ANOVA models are regression models with categorical predictors;
- two-way ANOVA introduces main effects and interaction;
- ANCOVA combines factor effects with continuous covariate adjustment;
- fitted means, adjusted means, and contrasts are often more meaningful than raw coding-specific coefficients;
- model comparison through general linear hypotheses unifies all these settings.

Next week, a natural continuation is to study generalized least squares and correlated errors, or to move toward repeated-measures and mixed-model ideas, depending on the course emphasis.

38.50 Appendix: Compact Formula Summary

One-way ANOVA cell-means model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}.$$

Two-way ANOVA with interaction:

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

Simple ANCOVA model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i.$$

ANCOVA with interaction:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i.$$

Unifying principle:

- factors enter through indicator variables;
- interactions enter through products of model terms;
- ANOVA, ANCOVA, and regression are all linear models.

39 Week 11: Generalized Least Squares, Correlated Errors, and Beyond Ordinary Least Squares

In this week, we study what happens when the error terms in a linear model are no longer independent with common variance. This leads to generalized least squares, a natural extension of ordinary least squares that accounts for nonidentity covariance structure. The goal is to help students understand how the linear model changes when observations are correlated or have unequal precision in a more general matrix form.

39.1 Learning Objectives

By the end of this week, students should be able to:

- explain why ordinary least squares is not always appropriate when errors are correlated or heteroscedastic;
- state the generalized linear model covariance assumption for the error vector in a linear model context;
- derive the generalized least squares estimator;
- explain how generalized least squares reduces to ordinary least squares and weighted least squares in special cases;
- interpret the transformed-model view of generalized least squares;
- recognize practical examples of correlated errors and structured covariance models.

39.2 Reading

Recommended reading for this week:

- Seber and Lee:
  - sections on generalized least squares
  - correlated observations and covariance structure
  - extensions of least squares methods

- Montgomery, Peck, and Vining:
  - sections on departures from ordinary linear model assumptions
  - weighted and generalized least squares ideas
  - practical modelling considerations for dependence and unequal variance

39.3 Why Ordinary Least Squares Can Fail

So far, much of the course has relied on the classical assumption

$$\mathrm{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

This means:

- all observations have the same error variance;
- errors are uncorrelated.

But many real datasets do not satisfy this assumption.

Examples include:

- repeated observations on the same unit;
- measurements taken over time;
- clustered or grouped data;
- observations with known unequal precision;
- spatial or serial dependence.

In such cases, ordinary least squares may still be unbiased for the mean model under suitable assumptions, but it is no longer the most efficient linear estimator, and the usual standard error formulas become incorrect.

39.4 Review of the Linear Model

Recall the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with

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

Under ordinary least squares, we assumed

$$\mathrm{Var}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

This gave the estimator

$$\widehat{\beta}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Now we allow a more general covariance structure.

39.5 A More General Covariance Model

Suppose instead that

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

where $\mathbf{V}$ is a known positive definite matrix.

Then

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2 \mathbf{V}.$$

The matrix $\mathbf{V}$ may represent:

- unequal variances;
- correlations among observations;
- both at once.

This is the setting of generalized least squares.

39.6 Why the Covariance Matrix Matters

If two observations are highly correlated, then together they contain less independent information than two unrelated observations.

If one observation has much larger variance than another, it should typically receive less weight in estimation.

Thus the covariance structure affects how much information each part of the data contributes.

Ordinary least squares ignores this structure. Generalized least squares incorporates it directly.

39.7 The Generalized Least Squares Criterion

In generalized least squares, we minimize the quadratic form

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta).$$

This is the natural analogue of the ordinary least squares criterion

$$(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta),$$

but now weighted by the inverse covariance structure.

The matrix $\mathbf{V}^{-1}$ downweights directions in the data that are more variable or more redundant.

39.8 Derivation of the Generalized Least Squares Estimator

Differentiate the generalized least squares criterion with respect to β and set the derivative equal to zero.

This gives the generalized normal equations

$$\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\, \widehat{\beta}_{GLS} = \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}.$$

Assuming invertibility, the generalized least squares estimator is

$$\widehat{\beta}_{GLS} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}.$$

This is the central formula for the week.

39.9 Interpretation of the GLS Estimator

The GLS estimator has the same structural form as OLS and WLS, but with the covariance matrix inserted.

It can be interpreted as an estimator that gives appropriate weight to observations according to the joint covariance structure.

Compared with OLS:

- observations with larger variance are effectively downweighted;

- correlated observations are not treated as if they were independent;
- the estimator uses the information in the data more efficiently when $\mathbf{V}$ is correctly specified.

39.10 Special Case: Ordinary Least Squares

If

$$\mathbf{V} = \mathbf{I}_n,$$

then GLS becomes

$$\widehat{\beta}_{GLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

which is exactly the ordinary least squares estimator.

So OLS is a special case of GLS.

39.11 Special Case: Weighted Least Squares

If the covariance matrix is diagonal,

$$\mathbf{V} = \text{diag}(v_1, \dots, v_n),$$

then

$$\mathbf{V}^{-1} = \text{diag}(1/v_1, \dots, 1/v_n).$$

This is the weighted least squares setting from the previous week, with weights proportional to inverse variances.

So WLS is also a special case of GLS.

39.12 The Transformed-Model View

A very important way to understand GLS is through a transformation.

Because $\mathbf{V}$ is positive definite, there exists a matrix $\mathbf{A}$ such that

$$\mathbf{A}^\top \mathbf{A} = \mathbf{V}^{-1}.$$

For example, we may take $\mathbf{A} = \mathbf{V}^{-1/2}$.

Multiplying the model by $\mathbf{A}$ gives

$$\mathbf{A}\mathbf{Y} = \mathbf{A}\mathbf{X}\beta + \mathbf{A}\varepsilon.$$

Now the transformed error has variance

$$\mathrm{Var}(\mathbf{A}\varepsilon) = \mathbf{A}(\sigma^2\mathbf{V})\mathbf{A}^\top = \sigma^2\mathbf{I}_n.$$

So GLS can be viewed as ordinary least squares applied to a transformed model with spherical errors.

This is one of the most important conceptual ideas in the course.

39.13 Distribution of the GLS Estimator

If the covariance structure is correctly specified and

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{V}),$$

then the GLS estimator is normally distributed:

$$\widehat{\beta}_{GLS} \sim N_p(\beta, \sigma^2(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X})^{-1}).$$

This parallels the OLS distribution theory, but with $\mathbf{V}^{-1}$ appearing throughout.

39.14 Gauss-Markov Interpretation

Under the covariance model

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

the generalized least squares estimator is the best linear unbiased estimator among all linear unbiased estimators.

So GLS is the natural Gauss-Markov extension of OLS to nonidentity covariance structures.

This is sometimes phrased by saying that OLS is BLUE under spherical errors, while GLS is BLUE under the more general covariance model.

39.15 Estimating Sigma Squared Under GLS

In the transformed model, residual sums of squares can be defined using the GLS criterion.

The generalized residual sum of squares is

$$SSE_{GLS} = (\mathbf{Y} - \mathbf{X}\widehat{\beta}_{GLS})^\top \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\widehat{\beta}_{GLS}).$$

Under the normal model with known $\mathbf{V}$, this plays the same role as the ordinary SSE after transformation.

An estimator of σ^2 is then

$$\widehat{\sigma}^2_{GLS} = \frac{SSE_{GLS}}{n - p}.$$

39.16 Why Known V Is Rare

In practice, the covariance matrix $\mathbf{V}$ is usually not known exactly.

Instead, we may have:

- scientific knowledge suggesting a covariance pattern;
- repeated-measures structure suggesting dependence;
- known weights from design considerations;
- an estimated covariance model.

When $\mathbf{V}$ is replaced by an estimate, the resulting estimator is often called **feasible GLS**.

39.17 Feasible Generalized Least Squares

The basic idea of feasible GLS is:

- propose a covariance model for $\mathbf{V}$;
- estimate the unknown covariance parameters from the data;
- plug the estimated covariance matrix into the GLS formula.

This is practical, but it introduces an extra modelling step and relies on the covariance model being approximately correct.

39.18 Examples of Covariance Structures

Several structured covariance matrices appear often in applications.

39.18.1 Unequal Variances Only

If observations are independent but have different variances, then $\mathbf{V}$ is diagonal.

This is the weighted least squares case.

39.18.2 Compound Symmetry

In grouped or repeated-measures settings, one may assume a common variance and a common correlation within group.

This gives a covariance pattern often called compound symmetry.

39.18.3 Autoregressive Structure

For time-ordered observations, nearby errors may be more strongly correlated than distant ones.

A simple model is the AR(1) structure, where correlations decay geometrically with lag.

39.18.4 Block-Diagonal Structure

If observations are grouped into independent clusters, then $\mathbf{V}$ may be block diagonal, with each block describing within-cluster dependence.

These examples help students see that covariance structure is part of model formulation, not just a technical afterthought.

39.19 Correlated Errors in Time-Ordered Data

Suppose we observe data over time:

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

If the errors satisfy

$$\operatorname{Corr}(\varepsilon_t, \varepsilon_{t-1}) \neq 0,$$

then ordinary least squares may still estimate the mean trend, but standard errors based on independence are not trustworthy.

This is one of the classic motivations for generalized least squares.

39.20 Clustered and Repeated Observations

Suppose several observations come from the same individual, school, hospital, or geographic region.

Then responses within the same cluster may be correlated.

In this case, the assumption of independent errors is not appropriate, and a structured covariance model or a mixed-model approach may be better suited.

GLS provides an important stepping stone toward these more advanced frameworks.

39.21 Relationship to Robust Standard Errors

One alternative to full covariance modelling is to keep the OLS mean estimator but adjust the standard errors to be robust to heteroscedasticity or dependence.

This is a different strategy from GLS.

- GLS changes the estimator itself using a covariance model;
- robust standard errors keep the estimator but correct inference approximately.

Both are useful, but they serve different purposes.

This distinction is valuable for students to understand even if robust methods are treated only briefly.

39.22 When GLS Is Worth Using

GLS is especially attractive when:

- there is a defensible covariance model;
- the dependence or heteroscedasticity is substantial;
- efficient estimation matters;
- the transformed model remains interpretable.

If the covariance model is poorly specified, the gains from GLS may be limited, and robustness or alternative modelling strategies may be preferable.

39.23 Diagnostics for Correlated Errors

Symptoms that may suggest correlated errors include:

- residuals that display runs or systematic temporal patterns;
- residual plots against time showing clustering;
- repeated-measures structure built into the design;
- known grouping or spatial dependence in the data collection process.

The key lesson is that covariance structure should be motivated both by diagnostics and by how the data were collected.

39.24 Worked Example With Known Unequal Precision

Suppose each response is an average of m_i repeated measurements.

Then the variance of the average may satisfy

$$\mathrm{Var}(Y_i) = \frac{\sigma^2}{m_i}.$$

In that case, a natural choice is

$$v_i = \frac{1}{m_i}, \qquad w_i = m_i.$$

So observations based on more repeated measurements get more weight.

This is a very intuitive example of GLS through weighted least squares.

39.25 Worked Example With Correlated Pairs

Suppose observations come in natural pairs, and within each pair the errors are positively correlated.

Then the covariance matrix has a block structure, with a 2×2 covariance block for each pair.

GLS accounts for the fact that the two observations in a pair do not contribute as much independent information as two unrelated observations would.

39.26 R Demonstration With Weighted Least Squares as GLS

39.27 Simulate data with unequal variances

```r
set.seed(123)
n <- 40
x <- seq(1, 20, length.out = n)
v <- 0.3 + 0.08 * x^2
y <- 5 + 1.2 * x + rnorm(n, sd = sqrt(v))

dat <- data.frame(y = y, x = x, v = v)
```

```r
fit_ols <- lm(y ~ x, data = dat)
fit_wls <- lm(y ~ x, data = dat, weights = 1 / v)

summary(fit_ols)
```

```
Call:
lm(formula = y ~ x, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5979 -1.6024  0.2916  1.9490  5.0754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.63093    0.93057   4.976 1.43e-05 ***
x            1.24643    0.07813  15.954  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.779 on 38 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8667
F-statistic: 254.5 on 1 and 38 DF,  p-value: < 2.2e-16
```

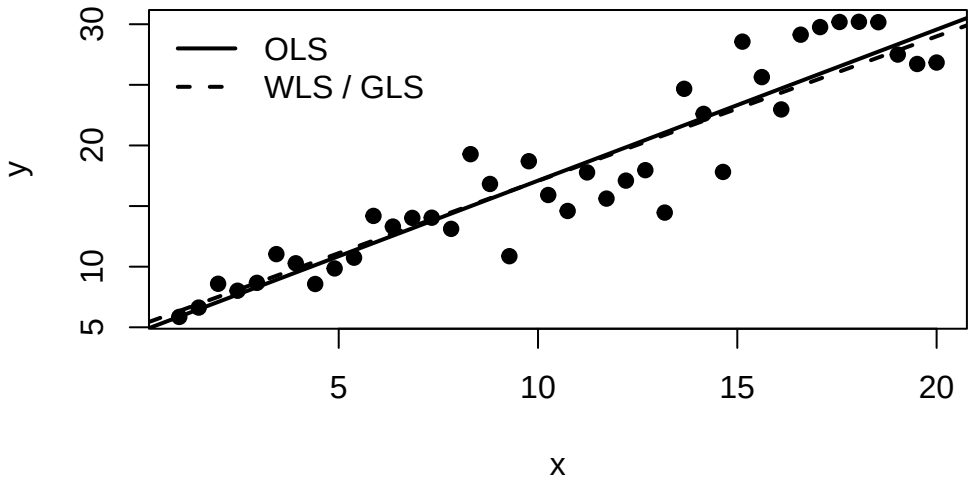
```r
summary(fit_wls)
```

```
Call:
lm(formula = y ~ x, data = dat, weights = 1/v)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-1.9970 -0.6075 -0.0324  0.6769  1.7500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.15391    0.34558   14.91   <2e-16 ***
x            1.19219    0.06256   19.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9073 on 38 degrees of freedom
Multiple R-squared:  0.9053,    Adjusted R-squared:  0.9028
F-statistic: 363.1 on 1 and 38 DF,  p-value: < 2.2e-16
```

39.28 Compare fitted lines

```r
plot(dat$x, dat$y, pch = 19, xlab = "x", ylab = "y")
abline(fit_ols, lwd = 2)
abline(fit_wls, lwd = 2, lty = 2)
legend("topleft",
       legend = c("OLS", "WLS / GLS"),
       lty = c(1, 2),
       lwd = 2,
       bty = "n")
```



39.29 R Demonstration With a Hand-Built GLS Computation

39.30 Construct a covariance matrix and compute GLS directly

```
set.seed(456)
n2 <- 8
x2 <- seq(1, n2)
X <- cbind(1, x2)

rho <- 0.6
V <- outer(1:n2, 1:n2, function(i, j) rho^abs(i - j))
beta_true <- c(2, 1.5)

eps <- t(chol(V)) %*% rnorm(n2)
y2 <- as.vector(X %*% beta_true + eps)

Y <- matrix(y2, ncol = 1)

beta_gls <- solve(t(X) %*% solve(V) %*% X) %*% t(X) %*% solve(V) %*% Y
beta_ols <- solve(t(X) %*% X) %*% t(X) %*% Y

beta_gls
```

```
        [,1]
   0.7277586
x2 1.6715885
```

```
beta_ols
```

```
       [,1]
   1.089930
x2 1.596804
```

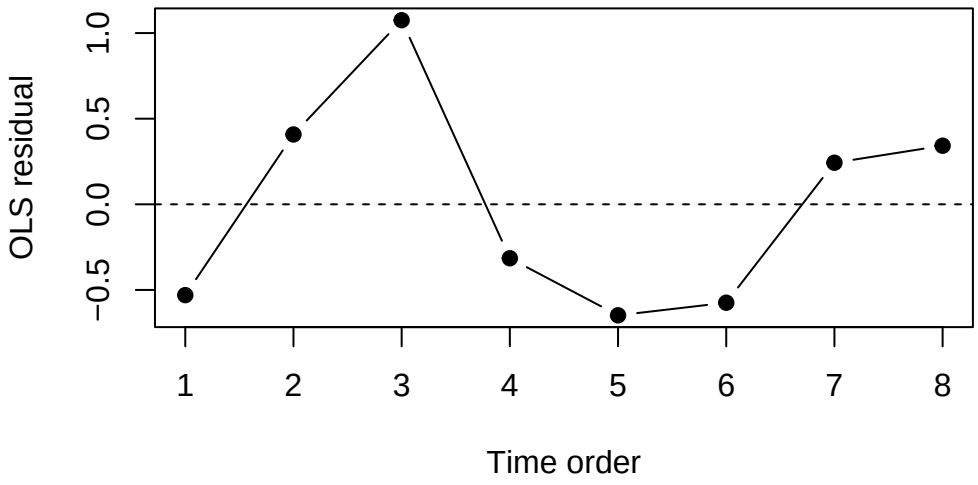
39.31 Transform the model and verify the GLS view

```r
A <- solve(chol(V))
Y_tilde <- A %*% Y
X_tilde <- A %*% X

beta_transformed <- solve(t(X_tilde) %*% X_tilde) %*% t(X_tilde) %*% Y_tilde
beta_transformed
```

```
        [,1]
   0.9738013
x2 1.6418527
```

39.32 Compare residual patterns

```r
fit_ols2 <- lm(y2 ~ x2)

plot(x2, resid(fit_ols2), type = "b", pch = 19,
     xlab = "Time order", ylab = "OLS residual")
abline(h = 0, lty = 2)
```



39.33 Simple example of grouped covariance intuition

```r
set.seed(789)
group <- rep(1:10, each = 3)
x3 <- rnorm(length(group))
u <- rnorm(10, sd = 1.2)
eps3 <- rep(u, each = 3) + rnorm(length(group), sd = 0.5)
y3 <- 4 + 2 * x3 + eps3

dat3 <- data.frame(y = y3, x = x3, group = factor(group))
fit3 <- lm(y ~ x, data = dat3)
summary(fit3)
```

```
Call:
lm(formula = y ~ x, data = dat3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3891 -0.6203 -0.2549  0.8272  1.4704

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7259     0.1740  21.417  < 2e-16 ***
x             2.1131     0.2316   9.125 6.97e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8837 on 28 degrees of freedom
Multiple R-squared:  0.7484,    Adjusted R-squared:  0.7394
F-statistic: 83.27 on 1 and 28 DF,  p-value: 6.974e-10
```

39.34 Interpreting Software Output

In base R, ordinary `lm()` directly handles OLS and weighted least squares through the `weights=` argument.

More general correlated-error models often require additional packages in practice, but the matrix formula is already enough to understand the core idea of GLS.

Students should focus on:

- what covariance structure is being assumed;
- why that structure is plausible;
- how the weighting or transformation changes the estimator;
- what practical goal is being improved.

39.35 A Practical Workflow for GLS Thinking

A sensible workflow is:

- ask whether independence and equal variance are plausible from the design of the study;
- inspect residual patterns for signs of heteroscedasticity or dependence;
- propose a covariance structure or weighting scheme when justified;
- compare OLS and GLS-style fits;
- interpret the results in light of both efficiency and model credibility.

39.36 In-Class Discussion Questions

1. Why is weighted least squares a special case of generalized least squares?
2. Why can correlated observations carry less information than independent observations?
3. What is the conceptual benefit of the transformed-model view of GLS?
4. When might robust standard errors be preferred to specifying a full covariance model?

39.37 Practice Problems

39.38 Conceptual

1. Explain why OLS is not generally efficient when the covariance matrix is not proportional to the identity.
2. Explain how GLS uses the inverse covariance matrix to weight information in the data.
3. Explain the difference between modelling the covariance structure and simply correcting standard errors.

39.39 Computational

Suppose

$$\text{Var}(\varepsilon) = \sigma^2 \mathbf{V}, \qquad \mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}.$$

1. Write down $\mathbf{V}^{-1}$.
2. Which observation receives more weight in GLS?
3. Explain why.

Now suppose

$$\hat{\beta}_{GLS} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}.$$

1. Show how this reduces to OLS when $\mathbf{V} = \mathbf{I}_n$.
2. Show how this reduces to WLS when $\mathbf{V}$ is diagonal.

39.40 Modelling Problem

You have repeated observations on each subject over time.

1. Why is the independence assumption questionable?
2. What kinds of covariance structure might be reasonable?
3. Why might GLS be more appropriate than OLS in this setting?

39.41 Suggested Homework

Complete the following tasks:

- fit an OLS model and a weighted least squares model for data with unequal variance;
- compare coefficient estimates, standard errors, and residual plots;
- compute a GLS estimator by hand for a small example with a known covariance matrix;
- explain the transformed-model view in your own words;
- write a short discussion of when a covariance model is scientifically credible enough to justify GLS.

39.42 Summary

In this week, we studied generalized least squares as an extension of ordinary least squares to models with nonidentity covariance structure.

We emphasized that:

- OLS assumes spherical errors;
- GLS allows unequal variances and correlated errors;
- WLS is a special case of GLS;
- GLS can be understood as OLS on a transformed model;
- covariance structure is part of model formulation and should be justified by both design and diagnostics.

Next week, a natural continuation is to move toward repeated-measures ideas, mixed models, or robust inference, depending on the course emphasis.

39.43 Appendix: Compact Formula Summary

General covariance model:

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

GLS criterion:

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta).$$

GLS estimator:

$$\widehat{\beta}_{GLS} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y}.$$

Transformed model idea:

$$\mathbf{A}\mathbf{Y} = \mathbf{A}\mathbf{X}\beta + \mathbf{A}\varepsilon, \qquad \mathbf{A}^\top \mathbf{A} = \mathbf{V}^{-1}.$$

Key message:

- OLS, WLS, and GLS are all part of one least squares framework;
- the difference lies in how the covariance structure is represented.

40 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Montgomery, Douglas C. 2017. *Design and Analysis of Experiments*. John wiley & Sons.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2021. *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Seber, George AF, and Alan J Lee. 2003. *Linear Regression Analysis*. John Wiley & Sons.

Part I

Appendix

41 Matrix Algebra Review

Appendix: Optional Review of Matrix Facts

Matrix dimensions

If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then $\mathbf{AB}$ is defined and is $m \times p$.

Transpose rules

For conformable matrices,

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top.$$

Inverse rule

If $\mathbf{A}$ and $\mathbf{B}$ are invertible, then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

Symmetric matrices

A matrix $\mathbf{A}$ is symmetric if

$$\mathbf{A}^\top = \mathbf{A}.$$

Covariance matrices are always symmetric.