# Course notes for MATH4/68011

# Part I: Linear Models

Dr I. Henriques-Cadby, Dr O. Johnson, Dr T. Waite

Department of Mathematics

University of Manchester

ines.henriques-cadby@manchester.ac.uk

September, 2025

## Contents

**Preliminaries** **3**

Prerequisites . . . 3

Course Format . . . 3

Intended Learning Outcomes . . . 3

Coursework and Assessment . . . 4

Recommended Reading . . . 4

Further Reading . . . 4

**1 Introduction to the Linear Model** **6**

1.1 Regression models . . . 6

1.2 Motivating examples . . . 7

1.3 Fitting models in R . . . 10

1.4 The general linear model . . . 15

1.5 Matrix form . . . 17

1.6 Categorical variables . . . 18

1.7 Regression with both qualitative and quantitative variables (ANCOVA) . . . 22

**2 Least squares estimation** **27**

2.1 Least squares and the normal equations . . . . . 27

2.2 Multivariate random variables . . . . . 31

2.3 Properties of the least squares estimator in the non-singular case . . . . . 34

2.4 The singular case . . . . . 35

2.5 Computation of a $g$-inverse . . . . . 38

**3 The fitted model and residuals** **40**

3.1 Fitted values and the hat matrix . . . . . 40

3.2 Residuals . . . . . 41

3.3 Variance estimation . . . . . 44

3.4 Residual diagnostics . . . . . 46

3.5 Coefficients of determination . . . . . 51

3.6 Leverage and influential points . . . . . 52

**4 The Gauss-Markov theorem** **56**

4.1 Estimable functions . . . . . 56

4.2 Best linear unbiased estimator . . . . . 58

**5 Statistical inference under normality** **60**

5.1 Distributional results . . . . . 60

5.2 Hypothesis tests . . . . . 62

5.3 Confidence intervals . . . . . 68

5.4 Prediction intervals . . . . . 70

**6 Testing a general linear hypothesis** **71**

6.1 Linear hypotheses and reduced models . . . . . 71

6.2 The test statistic . . . . . 74

6.3 One-way analysis of variance (ANOVA) . . . . . 77

6.4 Testing for lack of fit . . . . . 78

6.5 Proof of Proposition 6.2 . . . . . 79

6.6 Testable hypotheses . . . . . 82

# Preliminaries

Welcome to MATH4/68011 Linear Models and Nonparametric Regression! This course is aimed primarily at students on the MSc in Statistics and our MMath programmes. **These notes only cover the first part of the course which is Linear models, there is another set of notes on non-parametric regression.**

## Prerequisites

The course assumes a basic knowledge of linear algebra, probability theory and statistical methods, and ideally some experience programming in R. No problem if you have forgot some of the concepts, I will remind you, either in the notes or in the lectures.

## Course Format

Each week there will be several activities to complete.

1. Asynchronous activities (complete in your own time):
    - Watch lecture videos.
    - Complete quizzes - attempt the formative quizzes before lecture on Friday. Quiz marks do not count towards your final grade.s
    - Attempt the example sheet.
2. Synchronous activities (live activities bringing everyone together):
    - Lecture classes - These will go over the main points from the videos and formative quizzes.
    - Tutorial - Active practice of the material. It will involve discussion of the example sheet, so please attempt the example sheet beforehand (no problem if you don't complete everything!).

The week's lecture is referenced as a chapter in the lecture notes. I will like to acknowledge my colleagues Dr Tim Waite and Dr Olatunji Johnson for kindly donating this lecture notes and videos for this course.

## Intended Learning Outcomes

After successful completion of the course, students should be able to:

- formulate appropriate linear models and statistical hypotheses to investigate appropriate (scientific or real-world) questions;
- apply linear modelling and nonparametric regression techniques to make inferences and predictions about the regression relationship between covariates and a normally-distributed response variable, and to shed light on related real-world questions;
- check whether the assumptions underpinning such analyses are justified;
- explain the resulting models and analyses, with reference to the original real-world questions where appropriate, and also explain key underpinning ideas, assumptions, procedures, and theoretical results in linear modeling and nonparametric regression;
- prove both standard and unfamiliar mathematical and theoretical results underpinning the techniques studied, devising new arguments to do so where necessary;
- use R to apply the methods and to carry out appropriate simulation studies to assess the performance of the methods studied.

## Coursework and Assessment

80% Exam (3 hours) + 20% Coursework (take home)

The coursework will require you to perform data analysis on a real data-set. The analysis will be performed in R, using the methods and packages demonstrated throughout the course.

## Recommended Reading


- Weisberg, S. (2005) Applied linear regression. Wiley.
- Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012) Introduction to linear regression analysis. Wiley.
- Faraway, Julian J. Linear models with R. Chapman and Hall/CRC, 2004.
- Rawlings, J.O. (1998) Applied regression analysis: a research tool. Wadsworth and Brooks/Cole.
- Bowman, A.W. and Azzalini, A. (1998) Applied Smoothing Techniques for Data Analysis. Oxford University Press.

- Wand, M.P. and Jones, M.C. (1995) Kernel Smoothing. Chapman and Hall.
- Eubank, R.L. (1999) Nonparametric regression and spline smoothing. Dekker.
- Hardle, W. (1990) Applied Nonparametric Regression. Cambridge University Press.

# 1 Introduction to the Linear Model

## 1.1 Regression models

Regression analysis is used to study the relationship between a continuous *response* variable $Y$ and one or more *predictor* variables, denoted $x$ (single predictor) or $x_1, \ldots, x_p$ (multiple predictors). We suppose that we have data consisting of $n$ observations of these variables, denoted as follows:

Single variable case:

| Variable | $Y$ | $x$ |
|---|---|---|
| Observation 1 | $Y_1$ | $x_1$ |
| Observation 2 | $Y_2$ | $x_2$ |
| $\vdots$ | | |
| Observation $n$ | $Y_n$ | $x_n$ |

Multiple variable case:

| Variable | $Y$ | $x_1$ | $x_2$ | $\ldots$ | $x_p$ |
|---|---|---|---|---|---|
| Observation 1 | $Y_1$ | $x_{11}$ | $x_{12}$ | $\ldots$ | $x_{1p}$ |
| Observation 2 | $Y_2$ | $x_{21}$ | $x_{22}$ | $\ldots$ | $x_{2p}$ |
| $\vdots$ | | | | | |
| Observation $n$ | $Y_n$ | $x_{n1}$ | $x_{n2}$ | $\ldots$ | $x_{np}$ |

Note in this case $x_{ij}$ is the value of the $j$th predictor for the $i$th observation.

A typical regression model assumes that the expected response is a function of the predictors and some unknown parameters $\boldsymbol{\theta}$

$$E(Y) = g(x_1, \ldots, x_p; \boldsymbol{\theta})\,. \tag{1.1}$$

The form of the function $g$ must be specified by the Statistician. The parameter values $\boldsymbol{\theta}$ can be estimated by finding the values that best fit the observed data. Different models, corresponding to different functions $g$, can be fitted and compared to find the best fit to the observed data.

Once a model has been fitted it can be used to:

1. predict a future or unseen response given specified values of the predictors;
2. answer questions, or test hypotheses about, the relationship between the response and the predictors.

## 1.2 Motivating examples

### 1.2.1 Galton height data

Can we predict someone's height from the height of their parents? Below we show data gathered by Galton (1886), which shows the height of a son (in inches) versus the 'mid-parent height'. It is clear there is a positive linear association.

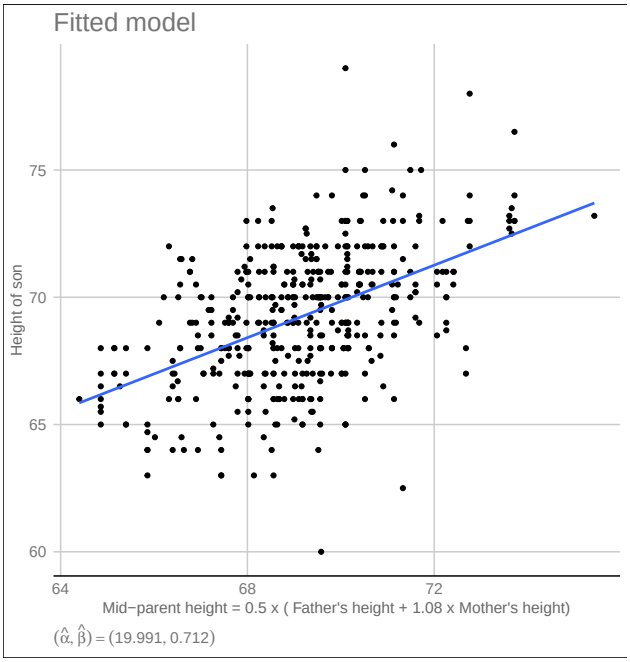



Figure 1: Plot of son's height against parental average data including the regression line $y = 19.991 + 0.712x$.

The relationship can be quantified by fitting the *simple linear regression model*

$$Y_i = \alpha + \beta x_i + \epsilon_i\,, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{1.2}$$

where the response variable $Y$ is the son's height and the predictor $x$ is the mid-parent height. Here $\alpha$ and $\beta$ denote the unknown regression parameters, which must be estimated from the data. The *random error* terms $\epsilon_i$ represent the component of the sons' height that is not due to the parents' height. Note we make the following assumptions about the random error terms:

1. the random errors are independent;
2. the random errors are normally distributed;
3. the random errors have expectation zero;

4. the random errors have constant variance.

These are common assumptions, but in practice they should be checked after fitting the model (see later chapters). Note that as a consequence of the third assumption, this model states that

$$\mathbb{E}(Y) = \alpha + \beta x \,.$$

For this data, the best-fitting values of the unknown parameters can be shown to be (more on this later)

$$\hat{\alpha} = 19.991 \quad \text{and} \quad \hat{\beta} = 0.712 \,.$$

The blue line above shows the *fitted* model $\hat{E}(Y) = \hat{\alpha} + \hat{\beta}x = 19.991 + 0.712x$. The fitted model can be used to predict the height of a son whose mid-parent height is $x = 70$ inches, via

$$\hat{\mathbb{E}}(Y)\big|_{x=70} = 19.991 + 0.712 \times 70 = 69.831 \,.$$

To test whether the son's height is associated with the mid-parent height, we would perform a significance test of the null hypothesis

$$H_0 : \beta = 0 \qquad \text{vs} \qquad H_1 : \beta \neq 0$$

There are some interesting points to note. Although there is certainly a correlation between the sons' height and the parents' height, it is not that strong. Indeed, the mid-parent height only explains around $23\%$ of the variability in the son's height (details given in page 12). Also, simple calculations (see example sheet) show that the sons are closer to the average height than their parents. This phenomenon is known as *'regression to the mean'* (hence the term *'regression analysis'*).

### 1.2.2 Soya bean data

The figure below shows data collected in an experiment to determine the optimum soil pH for growing a new strain of soya bean. How can we estimate the best pH?

We see that the data follow an approximately quadratic shape. This suggests that a reasonable model may be

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i \,, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \,,$$

where the response variable $Y$ is soya yield, the predictor variable $x$ is soil pH, and $\alpha$, $\beta$ and $\gamma$ are unknown parameters to be estimated from the data. Note that this model states that

$$\mathbb{E}(Y) = \alpha + \beta x + \gamma x^2 \,.$$

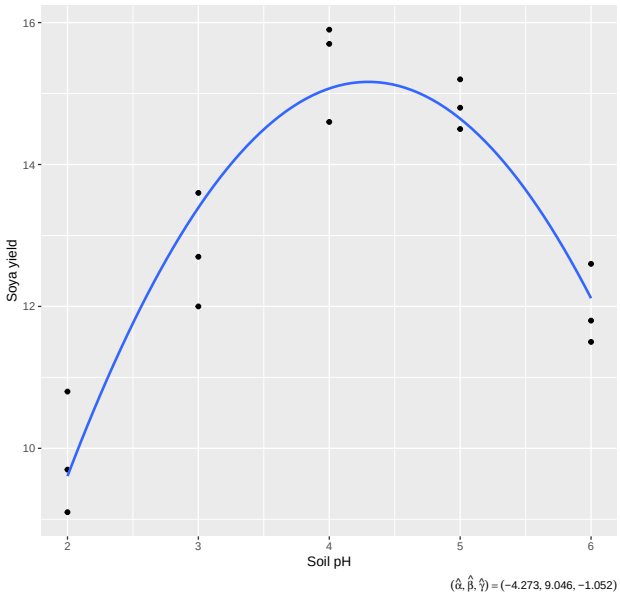



Figure 2: Plot of soya data including the quadratic regression curve

$$y = -4.273 + 9.046x - 1.052x^2$$

The best-fitting estimates of the parameters $\alpha, \beta, \gamma$ can be shown to be $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = (-4.273, 9.046, -1.052)$, giving the fitted model shown in the blue line

$$\hat{\mathbb{E}}(Y) = -4.273 + 9.046x - 1.052x^2 \,.$$

To estimate the pH $x^*$ which maximizes the expected response, we solve

$$0 = \frac{d\hat{\mathbb{E}}(Y)}{dx} = 9.046 - 2 \times 1.052x$$

giving $\hat{x}^* = 4.299$.

### 1.2.3 Alternative terminology

- The response variable is sometimes also known as the *outcome*, *output*, or *dependent* variable.
- A predictor variable is also sometime known as an *input*, *explanatory*, or *independent* variable.
- I recommend you avoid the dependent/independent terminology as it is easily confused with the idea of dependent/independent random variables, but it is important to be aware of this as you might encounter it outside this course.

## 1.3 Fitting models in R

In practice the most straightforward and flexible way to fit these kinds of models is to use R and RStudio.

### 1.3.1 The R language

- 'R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques'
- it is available to download free on your own computer from `https://cran.r-project.org`
- it is easily extensible via packages adding functions for new statistical techniques
- the most popular language among Statisticians (Python is also big in Data Science, and Julia is rapidly growing)

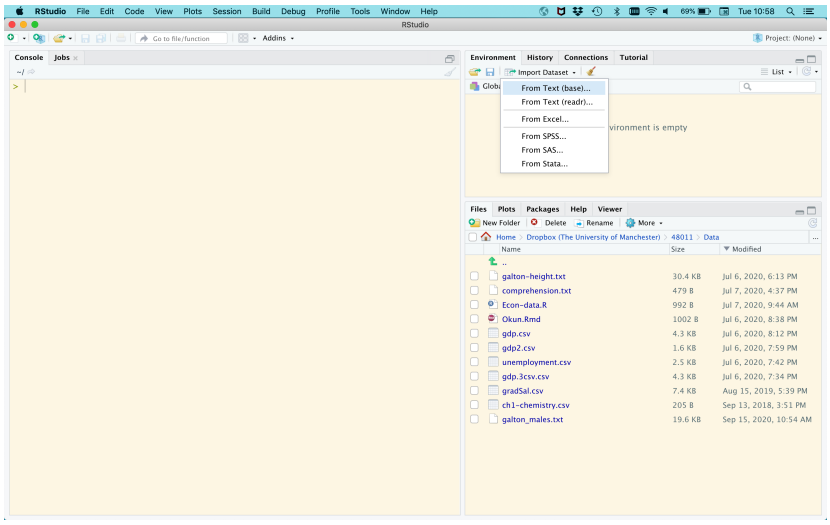
### 1.3.2 The RStudio IDE

- an *'integrated development environment' (IDE)* for the R language
- available freely at `https://rstudio.com/products/rstudio/`
- gives you many additional useful features that make it much easier to work with R, e.g. improved data import interface, automatic code completion, writing statistical reports via Markdown and Sweave, package development and more

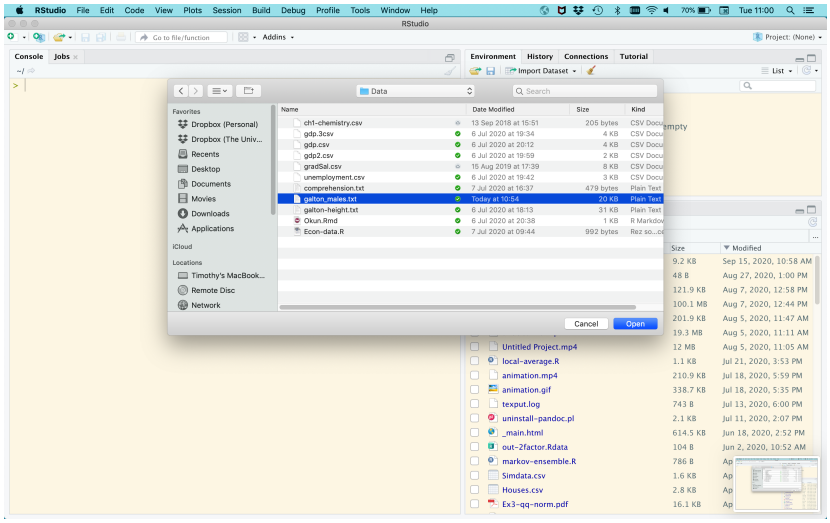
### 1.3.3 Example: Galton height data

We first load the Galton height data.

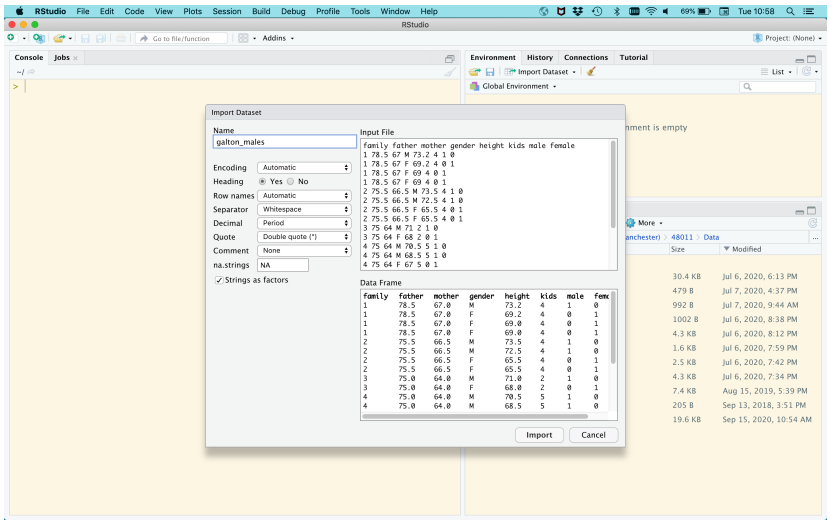
1. First download the data from Blackboard (file `galton_males.txt` in the Week 1 folder) and save it somewhere on your machine.
2. Load RStudio
3. Click on the 'import dataset' button, and select 'From Text (base)' (see figure)



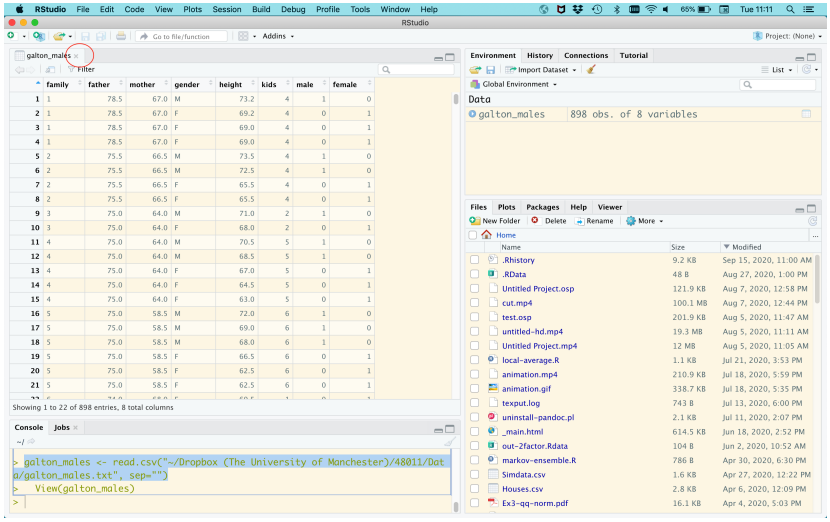

4. Navigate to the folder where you saved the data, and select the file (see figure)




5. Check the 'Data Frame' panel to ensure that R has correctly recognised the format of the data, i.e. that it correctly identified the rows and columns in the data. If not, then you can change the settings on the left until the desired result is achieved. Once you are satisified, click 'Import'.



6. The data will be imported to a data frame called `galton_males`. R will show you a panel which lets you inspect the data. When you are happy, you can close this.



Note that there other (programatic) ways of importing data, but for this is the more intuitive way.

Now that we have loaded the data, we can plot it using:

```
with(galton_males, plot(midparent,height))
```

We can use the `lm` command to fit the simple linear regression model with `height` as the response and `midparent` as the predictor. Note: the parameters are included automatically, as is the intercept.

```
galton_fit <- lm(height ~ midparent, data=galton_males)
```

the command coef can be used to obtain the parameter estimates:

```
> coef(galton_fit)
(Intercept)   midparent
 19.9909997   0.7120758
```

We see that the result agree with the estimates $\hat{\alpha} = 19.991, \hat{\beta} = 0.712$ from the previous section. More extensive information (e.g. standard errors, significance tests) can be produced using summary.

```
> summary(galton_fit)

Call:
lm(formula = height ~ midparent, data = galton_males)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5372 -1.5230  0.1891  1.5157  9.0925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.99100    4.12165    4.85 1.69e-06 ***
midparent    0.71208    0.05959   11.95  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.303 on 463 degrees of freedom
Multiple R-squared:  0.2357, Adjusted R-squared:  0.2341
F-statistic: 142.8 on 1 and 463 DF,  p-value: < 2.2e-16
```

The line of best fit can be added to your plot of the data using:

```
abline(coef(galton_fit))
```

### 1.3.4 Example: Soya bean data

To set up the data, type (or copy and paste)

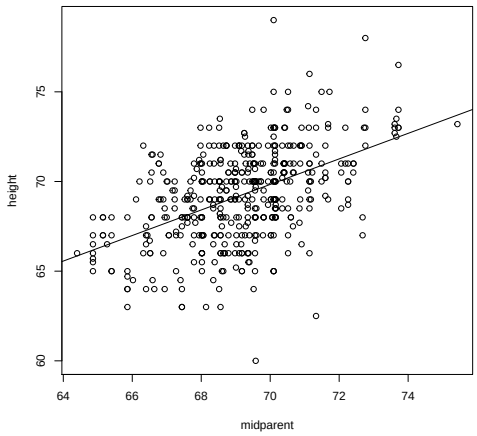



Figure 3: Using `abline` to add the regression line $y = 19.99100 + 0.71208x$ to the Galton height data.

```
soya_data <- data.frame(pH=rep(c(2,3,4,5,6),c(3,3,3,3,3)),
                        Yield=c(9.1,9.7,10.8,12,12.7,13.6,14.6,
                            15.7,15.9,15.2,14.8,14.5,12.6,11.8,11.5))
```

To fit the quadratic model with `Yield` as response and `pH` as the explanatory variable, we can again use the `lm` command:

```
soya_fit <- lm(Yield~pH+I(pH^2), data=soya_data)
```

Note that the `I()` command is important here: without it `R` will fit a simple linear model.[1] One can see more details about how to feed a formula to `R` here.

The coefficients of the regression can be obtained via

```
> coef(soya_fit)
(Intercept)          pH     I(pH^2)
  -4.273333    9.045714   -1.052381
```

Note that these estimates agree with the results in the previous section.

We can plot the data and the fitted model as follows:

```
> with(soya_data, plot(pH, Yield))
> curve(-4.273 + 9.046*x - 1.052*x^2, from=2,to=6, add=TRUE, col=4)
```

[1] In the `R` function `formula`, `I()` is used to inhibit the interpretation of operators such as "+", "-", "*" and "^" as formula operators, so they are used as arithmetical operators instead. "

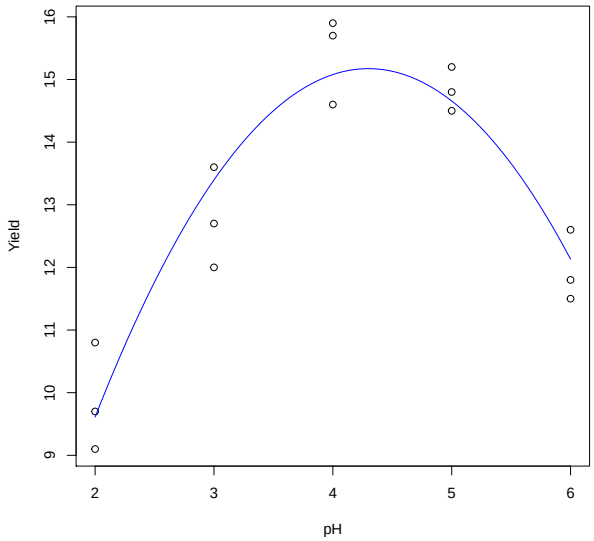



Figure 4: Using the command `curve` to add the quadratic curve $y = -4.273 + 9.046 * x - 1.052 * x^2$ to the Soya data.

## 1.4 The general linear model

A *linear model* is a regression model of the form

$$\mathbb{E}(Y) = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p \,. \tag{1.3}$$

Equivalently:

$$Y = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p + \epsilon \,, \qquad \mathbb{E}(\epsilon) = 0 \tag{1.4}$$

The key assumption is that $\mathbb{E}(Y)$ is linear in the unknown parameters $\theta_1, \ldots, \theta_p$. It does not necessarily have to be linear in the explanatory variables, because the $x_j$ are allowed to be nonlinear functions of each other (see examples below).

Note that if the above model is correct then the observed data satisfy

$$Y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \epsilon_i \qquad (i = 1, \ldots, n). \tag{1.5}$$

It is usually assumed that the random errors for different observations are uncorrelated and have zero mean and constant variance, i.e. we assume that:

- $\mathbb{E}(\epsilon_i) = 0$

- $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$
- $\mathrm{Var}(\epsilon_i) = \sigma^2$

Often we also make the stronger assumption that the $\epsilon_i$ are i.i.d. $N(0, \sigma^2)$ random variables.

Many statistical analyses can be put into the form of a linear model. By learning methodology that works for any linear model (e.g. estimation, testing, confidence intervals), we are able to solve many problems with one set of techniques.

### 1.4.1 Examples

**Simple Linear Regression model:** Suppose that there is a single explanatory variable $z$. Then the simple linear regression model can be written in general linear model form by setting $x_1 = 1$, $x_2 = z$. This gives

$$\begin{aligned}\mathbb{E}(Y) &= \theta_1 x_1 + \theta_2 x_2 \\ &= \theta_1 .1 + \theta_2 .z \\ &= \alpha + \beta z \qquad \text{(relabelling parameters)}\end{aligned}$$

**Quadratic Regression model:** Similarly the quadratic regression model can also be written in general linear model form by setting $x_1 = 1$, $x_2 = z$, $x_3 = z^2$, giving

$$\begin{aligned}\mathbb{E}(Y) &= \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \\ &= \theta_1 .1 + \theta_2 z + \theta_3 z^2 \\ &= \alpha + \beta z + \gamma z^2 \qquad \text{(relabelling parameters)}\,.\end{aligned}$$

Hence, perhaps surprisingly, a quadratic regression model (in $z$) is a linear model (in the $\theta$s)! This explains why we could fit the quadratic model using `lm` (which stands for Linear Model).

### 1.4.2 Transformations

Sometimes a nonlinear model can be made into a linear model by considering a transformation of the response. For example, the following model is not a linear model:

$$Y = Ue^{\beta x}$$

with unknown parameter $\beta$ and random error term $U$. However, taking a log transformation $Y' = \log Y$ we obtain

$$Y' = \beta x + \log U = \beta x + \epsilon$$

which is a linear model if $\mathbb{E}(\epsilon) = \mathbb{E}(\log U) = 0$.

## 1.5 Matrix form

Note that (1.5) defines a system of equations,

$$\begin{array}{rcccccccc} Y_1 & = & \theta_1 x_{11} & + & \ldots & + & \theta_p x_{1p} & + & \epsilon_1 \\ Y_2 & = & \theta_1 x_{21} & + & \ldots & + & \theta_p x_{2p} & + & \epsilon_2 \\ \vdots & & \vdots & & & & & & \vdots \\ Y_n & = & \theta_1 x_{n1} & + & \ldots & + & \theta_p x_{np} & + & \epsilon_n \end{array} \tag{1.6}$$

It will prove to be an extremely good idea to rewrite these equations in matrix form as

$$\begin{array}{ccccccc} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} & = & \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} & \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ \uparrow & & \uparrow & \uparrow & & \uparrow & \\ \mathbf{Y} & & \mathbf{X} & \boldsymbol{\theta} & & \boldsymbol{\epsilon} & \end{array} \tag{1.7}$$

giving the matrix form of the model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,. \tag{1.8}$$

The $n \times p$ matrix $\mathbf{X}$ is known as the *model matrix* or *design matrix*. The rows of $\mathbf{X}$ correspond to the different observations or cases in the data. The columns of $\mathbf{X}$ correspond to the different predictor variables in the model.

**Exercise 1.1.** *Using matrix multiplication, verify that forms* (1.6) *and* (1.7) *are equivalent.*

**Procedure 1.2** (Model matrix)**.** *To form the model matrix:*

1. *Create a column for each parameter;*
2. *For a given parameter, say $\gamma$, the entry in the $i$th row of the corresponding column is the coefficient of $\gamma$ in $\mathbb{E}(Y_i)$ (or of $Y_i$ in* (1.6) *above).*

*Note that* $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}$, *i.e.* $\mathbf{x}_1^{\mathrm{T}},\ \mathbf{x}_2^{\mathrm{T}},\ \ldots,\ \mathbf{x}_n^{\mathrm{T}}$ *are the rows of* $\mathbf{X}$.

### 1.5.1 Examples

**Simple linear regression**

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

We can write this in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}$$

**Quadratic regression**

$$Y_i = \alpha + \beta x_i^2 + \gamma x_i^2 + \epsilon_i$$

We can write this in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

## 1.6 Categorical variables

So far we have only used *quantitative* (numerical) variables. It is also possible to include *qualitative* (categorical) variables as predictors in a linear model. Note however that the response must always be quantitative and continuous.

For example, suppose we wish to investigate how salary $Y$ depends on sex, $x$. The response, salary (in £), is quantitative, but the predictor is qualitative (male/female). If your sample includes non-binary people you may wish to add a third category. To include a qualitative predictor we use what are called *dummy variables*.

### 1.6.1 Reference level

The most common approach uses a reference level, which is the level against which comparisons are made. For example, we may choose male as the reference level. Then we define a dummy variable for the other level,

$$w_i = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{otherwise} \end{cases}$$

This dummy variable can be included in the linear model together with an intercept, giving

$$Y_i = \alpha + \beta w_i + \epsilon_i\,, \qquad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

Note that

$$\mathbb{E}[Y|\text{male}] = \alpha + \beta \times 0 = \alpha$$
$$\mathbb{E}[Y|\text{female}] = \alpha + \beta \times 1 = \alpha + \beta$$

Hence $\beta = \mathbb{E}[Y|\text{female}] - \mathbb{E}[Y|\text{male}]$ corresponds to the difference in expected salary between females and males (if $\beta$ is negative, then females are paid less).

**If the categorical predictor has $k > 2$ levels, then we need to create $k - 1$ dummy variables, one corresponding to each non-reference level.** For example

$$d_i^F = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{otherwise} \end{cases}, \qquad d_i^O = \begin{cases} 1 & \text{if the } i\text{th person is non-binary} \\ 0 & \text{otherwise} \end{cases}$$

These can then be included in the model as follows:

$$Y_i = \alpha + \beta d_i^F + \gamma d_i^O + \epsilon_i\,, \qquad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

Note that $\gamma$ now denotes the difference in expected salary between a non-binary person and a male., and $\beta$ is as before.

### 1.6.2 R implementation

To show how to deal with categorical variables in R, we consider the `gradSal` dataset (available from Blackboard). This contains (fictitious) values for the sex, degree mark, and salary for 81 graduates. First load the data, then:

```
> attach(gradSal)
```

This makes the variables in the data more accessible, e.g. we can type `sal` rather than `gradSal$sal`.

**Caveat:** The command `attach` should be used with caution! If you local workspace already includes variable with the same name as a new variable in the data you are about to attach, you won't be able to refer to the new variable using its name only. This type of name conflict is common and care should be taken to avoid this. The `search()` command can be used to see the list of objects (and packages) that are currently attached in `R`.

```
> sex <- factor(sex) # declare sex as a categorical variable
> sex
  [1] M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M
 [36] M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M M
 [71] M M M M M M M M M M M M M M M M M M M M M M M M M M M F F F F F
[106] F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F
[141] F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F F
[176] F F F F F F F F F F F F F F F F F F F F F F F F F
Levels: F M
```

In the last line we see that the first level is F. This means 'female' will be treated as the reference category. To change the reference category to 'male':

```
> sex <- relevel(sex, ref="M")
```

Now we can fit the model

```
> fit <- lm(sal~sex)
```

```
> summary(fit)

Call:
lm(formula = sal ~ sex)

Residuals:
    Min      1Q  Median      3Q     Max
-15.277  -3.207   0.068   3.803  15.695

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.0222     0.5745   41.81   <2e-16 ***
sexF          0.3735     0.8125    0.46    0.646
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 5.745 on 198 degrees of freedom
Multiple R-squared:  0.001066, Adjusted R-squared:  -0.003979
F-statistic: 0.2113 on 1 and 198 DF,  p-value: 0.6463
```

We see that the estimated coefficient for females, $\hat{\beta} = 0.3735$, is positive, meaning that females appear to have slightly higher expected salary than males. The coefficient is not statistically significant. However, as we will see, there is more to the story.

### 1.6.3 Double-index notation

Another way of dealing with categorical variables is to use 'double-index notation'. Let $Y_{ij}$ denote the $j$th response in the $i$th group, where e.g. $i = 1$ corresponds to males, $i = 2$ corresponds to females, and $i = 3$ corresponds to other. Then a suitable model for the data may be

$$Y_{ij} = \mu_i + \epsilon_{ij}\,, \qquad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Note that $\mu_i$ is the expected response in the $i$th group. This formulation is commonly used in One-way ANOVA to test the hypothesis that the expected response is equal in all groups, i.e.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \,.$$

Notice that this is also a general linear model (with no intercept) because

$$E[Y] = \mu_1 d^M + \mu_2 d^F + \mu_3 d^O \,.$$

The model can be written in matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ as follows, arranging the responses by group:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ Y_{32} \\ \vdots \\ Y_{3n_3} \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \\ \epsilon_{31} \\ \epsilon_{32} \\ \vdots \\ \epsilon_{3n_3} \end{bmatrix},$$

## 1.7 Regression with both qualitative and quantitative variables (ANCOVA)

In the graduate salary example it is a good idea to take into account the graduates' degree marks $x_i$ when analysing the effect of sex. This is known as 'adjusting' for the degree mark. We can do this by fitting the model

$$Y_i = \alpha + \beta x_i + \gamma w_i + \delta x_i w_i + \epsilon_i \,, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{1.9}$$

where $w_i$ is the dummy variable for the 'female' category. Note that

$$\mathbb{E}[Y|\text{male, mark } x] = \alpha + \beta x$$
$$\mathbb{E}[Y|\text{female, mark } x] = (\alpha + \gamma) + (\beta + \delta)x \,.$$

Hence the model assumes that the relationship between degree mark and expected salary is a straight line, but the straight line may differ between males and females (both its intercept and slope).

Interpretation of the parameters:

- $\alpha$ is the regression intercept for males
- $\beta$ is difference in expected salary between a male with degree mark $x+1$ and a male with degree mark $x$ ('the effect of mark for males')
- $\gamma$ is the difference in the intercept between females and males
- $\delta$, corresponding to the product term, is known as the *interaction* effect
- the effect of degree mark for females is $\beta+\delta$
- hence the interaction $\delta$ is the difference in the effect of degree mark between females and males. If $\delta=0$, then the effect of degree mark on salary is the same for males and females, represented by the model:

$$Y_i = \alpha + \beta x_i + \gamma w_i + \epsilon_i\,, \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2) \tag{1.10}$$

To fit the model (1.9) in R use:

```
> fit2 <- lm(sal~mark+sex + sex*mark)
> summary(fit2)

Call:
lm(formula = sal ~ mark + sex + sex * mark)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4707  -3.4147   0.1865   3.2457  13.7702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.61480    2.66800   3.604 0.000398 ***
mark         0.24619    0.04483   5.492 1.22e-07 ***
sexF        -9.85186    4.39979  -2.239 0.026269 *
mark:sexF    0.10886    0.06711   1.622 0.106384
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

```
Residual standard error: 4.86 on 196 degrees of freedom
Multiple R-squared:  0.2924, Adjusted R-squared:  0.2816
F-statistic:    27 on 3 and 196 DF,  p-value: 1.162e-14
```

We note that the interaction term is not statistically significant ($p$-value $= 0.1$), so we refit the model without the interaction term as in equation (1.10) as follows:

```
> fit3 <- lm(sal~ mark+sex)
> summary(fit3)

Call:
lm(formula = sal ~ mark + sex)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3116  -3.1559   0.2033   3.2337  13.6722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7722     2.0201   3.352 0.000961 ***
mark          0.2948     0.0335   8.800 6.95e-16 ***
sexF         -2.8270     0.7801  -3.624 0.000369 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 4.88 on 197 degrees of freedom
Multiple R-squared:  0.2829, Adjusted R-squared:  0.2757
F-statistic: 38.87 on 2 and 197 DF,  p-value: 5.924e-15
```

The final model suggests that females receive around £2827 less than a male with the same degree

mark. This is surprising considering our earlier analysis indicated no difference between the groups. The explanation is that the females in our sample are more highly qualified, and so we would expect them to be paid more. This can be seen by computing the average mark in each group:

```
> aggregate(mark, by=list(sex), mean)
  Group.1        x
1       M 58.52052
2       F 69.37802
```

This shows how it is important to account for potential confounding factors, here the degree mark factor.

Note the model including interaction can be written in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & w_1 & x_1 w_1 \\ 1 & x_2 & w_2 & x_2 w_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & w_n & x_n w_n \end{bmatrix}$$

### 1.7.1 Double index notation

An alternative approach is to use double index notation. Again let $i = 1$ correspond to males and $i = 2$ correspond to females, with $j$ indexing individuals within a group, (and let the graduate's degree marks be denoted by $x_{ij}$). Then we could fit the model

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}\,, \qquad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)\,,$$

noting that now the parameters depend on $i$, i.e. they differ between groups. This is also a linear model, since it can be written as

$$\mathbb{E}[Y] = \alpha_1 d^M + \alpha_2 d^F + \beta_1 d^M x + \beta_2 d^F x$$

Now

$$\mathbb{E}[Y|\text{male}] = \alpha_1 + \beta_1 x$$

and

$$\mathbb{E}[Y|\text{female}] = \alpha_2 + \beta_2 x$$

The above can be written in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{12} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ 0 & 0 & 1 & x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{bmatrix}$$

# 2 Least squares estimation

## 2.1 Least squares and the normal equations

The true parameter values $\boldsymbol{\theta}$ are unknown, but given a particular vector $\mathbf{b}$ of possible parameter values we can compute the deviations

$$e_i = Y_i - x_{i1}b_1 - x_{i2}b_2 - \ldots - x_{ip}b_p$$

These can be gathered together into a vector, $\mathbf{e} = (e_1, e_2, \ldots, e_n)^{\mathrm{T}}$ which satisfies

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

In order to estimate the true $\boldsymbol{\theta}$, it seems natural to choose the vector $\mathbf{b}$ to make the deviations 'small' in some sense.

To quantify the meaning of 'small' versus 'large' deviations, we introduce the sum of squared deviations,

$$S(\mathbf{b}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\mathbf{b})\,.$$

The principle of least squares says that we should choose our estimate, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ to minimize $S(\hat{\boldsymbol{\theta}})$.

**Definition 2.1.** *Given $\mathbf{Y}$, a vector $\hat{\boldsymbol{\theta}}$ is called a* least squares estimate *if it satisfies*

$$S(\hat{\boldsymbol{\theta}}) \leq S(\mathbf{b}) \quad \text{for all } \mathbf{b} \in \mathbb{R}^p\,.$$

**Theorem 2.2.** *$\hat{\boldsymbol{\theta}}$ is a least squares estimate if and only if it is a solution of the* normal equations

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\theta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}\,. \tag{2.1}$$

The matrix equation (2.1) can be written as a system of $p$ linear scalar equations in the $p$ scalars $\hat{\theta}_1, \ldots, \hat{\theta}_p$. The $p \times p$ matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is known as the *information matrix*.

It can be shown that there always exists a solution to (2.1). If $\det \mathbf{X}^{\mathrm{T}}\mathbf{X} \neq 0$, then the information matrix is non-singular, i.e. invertible, and the solution is unique. In particular

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,.$$

If $\det \mathbf{X}^{\mathrm{T}}\mathbf{X} = 0$ then the information matrix is singular and there are infinitely many solutions. The singular case is studied more in Sections 2.4–2.5.

Before proving Theorem 2.2 we give another result and an example. The following is helpful for calculation:

**Proposition 2.3.** *(i) The information matrix is symmetric.*

*(ii) The $(j,k)$th entry of the information matrix is*

$$[\mathbf{X}^{\mathrm{T}}\mathbf{X}]_{jk} = \langle \mathrm{col}_j(\mathbf{X}), \mathrm{col}_k(\mathbf{X}) \rangle \,,$$

*where $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{n} a_i b_i$ denotes the inner product/dot product of $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$ and $\mathbf{b} = (b_1, \ldots, b_n)^{\mathrm{T}}$, and $\mathrm{col}_j(\mathbf{A})$ denotes the $j$th column of matrix $\mathbf{A}$.*

*(iii) $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is a $p \times 1$ column vector, with $j$th entry*

$$[\mathbf{X}^{\mathrm{T}}\mathbf{Y}]_j = \langle \mathrm{col}_j(\mathbf{X}), \mathbf{Y} \rangle \,.$$

*Proof of Proposition 2.3.* For part (i) recall that a matrix $\mathbf{S}$ is *symmetric* if $\mathbf{S}^{\mathrm{T}} = \mathbf{S}$. Also, for matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$. Finally note that $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ as required.

For part (ii), recall that for matrices $\mathbf{A}$ and $\mathbf{B}$ the $(j,k)$th entry of $\mathbf{AB}$ is

$$[\mathbf{AB}]_{jk} = \langle \mathrm{row}_j(\mathbf{A}), \mathrm{col}_k(\mathbf{B}) \rangle \,,$$

where $\mathrm{row}_j(\mathbf{A})$ denotes the $j$th row of $\mathbf{A}$. Hence $[\mathbf{X}^{\mathrm{T}}\mathbf{X}]_{jk} = \langle \mathrm{row}_j(\mathbf{X}^{\mathrm{T}}), \mathrm{col}_k(\mathbf{X}) \rangle = \langle \mathrm{col}_j(\mathbf{X}), \mathrm{col}_k(\mathbf{X}) \rangle$. Part (iii) follows similarly. □

### 2.1.1 Example: simple linear regression

The simple linear regression model is

$$Y_i = \alpha + \beta x_i + \epsilon_i \,, \quad i = 1, \ldots, n \,, \tag{2.2}$$

with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, for $i = 1, \ldots, n$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

The above is sometimes also known as the *'first-order model'*. Define

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$s_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \,.$$

We now show that the least squares estimates of $\alpha$ and $\beta$ are

$$\hat{\alpha} = \bar{Y} - \frac{\bar{x} s_{xY}}{s_{xx}} \,, \quad \hat{\beta} = \frac{s_{xY}}{s_{xx}} \,. \tag{2.3}$$

Recall that (2.2) can be rewritten in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Using Proposition 2.3 we have that

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & (s_{xx} + n\bar{x}^2) \end{bmatrix}, \quad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix} = \begin{bmatrix} n\bar{Y} \\ s_{xY} + n\bar{x}\bar{Y} \end{bmatrix}.$$

Recall that for a $2 \times 2$ matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ we have $\det(\mathbf{A}) = ad - bc$, and provided $\det \mathbf{A} \neq 0$, $\mathbf{A}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. Note that

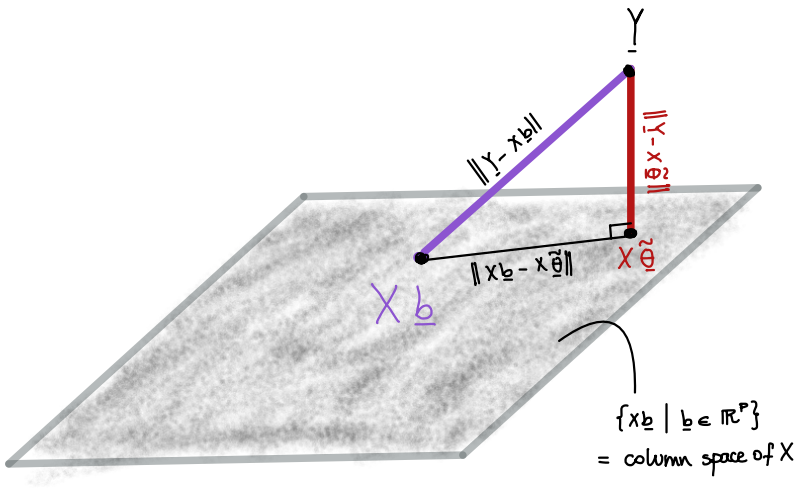
$$\det(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = n s_x x + n^2\bar{x}^2 - n^2\bar{x}^2 = n s_{xx}\,,$$

which is non-zero provided the design points are not all equal. In this case

$$\begin{aligned} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} &= \frac{1}{n s_{xx}} \begin{bmatrix} (s_{xx} + n\bar{x}^2) & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ s_{xY} + n\bar{x}\bar{Y} \end{bmatrix} \\ &= \frac{1}{n s_{xx}} \begin{bmatrix} (s_{xx} + n\bar{x}^2)n\bar{Y} - n\bar{x}(s_{xY} + n\bar{x}\bar{Y}) \\ -(n\bar{x})(n\bar{Y}) + n(s_{xY} + n\bar{x}\bar{Y}) \end{bmatrix} = \frac{1}{n s_{xx}} \begin{bmatrix} n s_{xx}\bar{Y} + n^2\bar{x}^2\bar{Y} - n\bar{x}s_{xY} - n^2\bar{x}^2\bar{Y} \\ -n^2\bar{x}\bar{Y} + n s_{xY} + n^2\bar{x}\bar{Y} \end{bmatrix} \\ &= \frac{1}{n s_{xx}} \begin{bmatrix} n s_{xx}\bar{Y} - n\bar{x}s_{xY} \\ n s_{xY} \end{bmatrix} = \begin{bmatrix} \bar{Y} - \frac{\bar{x}s_{xY}}{s_{xx}} \\ \frac{s_{xY}}{s_{xx}} \end{bmatrix} \quad \text{as required.} \end{aligned}$$

**Geometry of least squares**

To prove Theorem 2.2 we adopt a geometrical viewpoint. Think about $\mathbf{Y}$ as a point in $\mathbb{R}^n$. Then the linear model defines a subspace, $\mathcal{M} = \{\mathbf{X}\boldsymbol{\theta} | \boldsymbol{\theta} \in \mathbb{R}^p\}$, of $\mathbb{R}^n$, visualised as a plane in the figure below.




Consider the 'orthogonal projection of $\mathbf{Y}$ into $\mathcal{M}$', which is the point $\mathbf{X}\tilde{\boldsymbol{\theta}}$ satisfying the following property:

$$\langle(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}), \mathbf{X}\mathbf{b}\rangle = 0 \qquad \text{for all } \mathbf{b} \in \mathbb{R}^p \tag{2.4}$$

in other words the vector from $\mathbf{Y}$ to the projection is orthogonal to every vector in the subspace. Such a point always exists. Moreover, condition (2.4) holds if and only if $\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}) = \mathbf{0}$, i.e. if $\tilde{\boldsymbol{\theta}}$ is a solution of the normal equations.

*Proof of Theorem 2.2.* We have two statements to prove.

(i) *If $\hat{\boldsymbol{\theta}}$ is a solution of the normal equations, then it is a least squares estimate.* Suppose that $\hat{\boldsymbol{\theta}}$ is a solution of the normal equations. Then it satisfies (2.4) above. Let $\mathbf{b}$ be any other vector in $\mathbb{R}^p$. Then by Pythagoras' theorem

$$S(\mathbf{b}) = ||\mathbf{Y} - \mathbf{X}\mathbf{b}||^2 = ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}||^2 + ||\mathbf{X}(\hat{\boldsymbol{\theta}} - \mathbf{b})||^2 \geq ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}||^2 = S(\hat{\boldsymbol{\theta}})\,.$$

As $\mathbf{b}$ was arbitrary, $\hat{\boldsymbol{\theta}}$ is a least squares estimate. This proves (i).

(ii) *If $\hat{\boldsymbol{\theta}}$ is a least squares estimate, then it is a solution of the normal equations.* Suppose that $\hat{\boldsymbol{\theta}}$ is a least squares estimate and $\mathbf{X}\tilde{\boldsymbol{\theta}}$ be the orthogonal projection. By part (i), $\tilde{\boldsymbol{\theta}}$ is also a least squares estimate. Hence we have that $S(\hat{\boldsymbol{\theta}}) = S(\tilde{\boldsymbol{\theta}})$. On the other hand, by Pythagoras' theorem

$$S(\tilde{\boldsymbol{\theta}}) = ||\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}||^2 = ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}||^2 + ||\mathbf{X}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})||^2 = S(\hat{\boldsymbol{\theta}}) + ||\mathbf{X}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})||^2$$

hence we must have that $||\mathbf{X}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})||^2 = 0$, i.e. $\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}\tilde{\boldsymbol{\theta}}$. In this case, for any $\mathbf{b}$ we have that

$$\langle \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}, \mathbf{X}\mathbf{b}\rangle = \langle \mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}, \mathbf{X}\mathbf{b}\rangle = 0\,,$$

i.e. condition (2.4) also holds for $\hat{\boldsymbol{\theta}}$. Hence $\hat{\boldsymbol{\theta}}$ is a solution of the normal equations. This proves (ii).

□

## 2.2 Multivariate random variables

Before continuing with properties of the least squares estimator, we now discuss some elementary properties of multivariate random variables, i.e. random vectors and matrices. This will be needed since $\hat{\boldsymbol{\theta}}$ is a random vector.

If $U_1, \ldots, U_p$ are real-valued (scalar) random variables, then we say that

$$\mathbf{U} = (U_1, \ldots, U_p)^{\mathrm{T}}$$

is a *random (column) vector* of length $p$. Let also $\mathbf{W} = (W_1, \ldots, W_p)^{\mathrm{T}}$ be a random vector of length $q$. Then we define

$$\mathbb{E}(\mathbf{U}) = [\mathbb{E}(U_1), \ldots, \mathbb{E}(U_p)]^{\mathrm{T}}$$

$$\mathrm{Var}(\mathbf{U}) = \begin{bmatrix} \mathrm{Var}(U_1) & \mathrm{Cov}(U_1, U_2) & \ldots & \mathrm{Cov}(U_1, U_p) \\ \mathrm{Cov}(U_2, U_1) & \mathrm{Var}(U_2) & \ldots & \mathrm{Cov}(U_2, U_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(U_p, U_1) & \mathrm{Cov}(U_p, U_2) & \ldots & \mathrm{Var}(U_p) \end{bmatrix}$$

$$\mathrm{Cov}(\mathbf{U}, \mathbf{W}) = \begin{bmatrix} \mathrm{Cov}(U_1, W_1) & \mathrm{Cov}(U_1, W_2) & \ldots & \mathrm{Cov}(U_1, W_q) \\ \mathrm{Cov}(U_2, W_1) & \mathrm{Cov}(U_2, W_2) & \ldots & \mathrm{Cov}(U_2, W_q) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(U_p, W_1) & \ldots & \ldots & \mathrm{Cov}(U_p, W_q) \end{bmatrix}.$$

Note that:

- the variance matrix $\mathrm{Var}(\mathbf{U})$ is a $p \times p$ matrix with $i, j$th entry

$$[\mathrm{Var}(\mathbf{U})]_{ij} = \mathrm{Cov}(U_i, U_j)\,.$$

The $i$th entry on the diagonal is $\mathrm{Cov}(U_i, U_i) = \mathrm{Var}(U_i) \geq 0$. As $\mathrm{Cov}(U_i, U_j) = \mathrm{Cov}(U_j, U_i)$, the variance matrix is symmetric – that is, $\mathrm{Var}(\mathbf{U})^{\mathrm{T}} = \mathrm{Var}(\mathbf{U})$;

- the covariance matrix $\mathrm{Cov}(\mathbf{U}, \mathbf{W})$ is a $p \times q$ matrix with $i, j$th entry

$$[\mathrm{Cov}(\mathbf{U}, \mathbf{W})]_{ij} = \mathrm{Cov}(U_i, W_j)\,.$$

In addition, for a $p \times q$ *random matrix*

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1q} \\ M_{21} & M_{22} & \dots & M_{2q} \\ \vdots & & & \\ M_{p1} & M_{p2} & \dots & M_{pq} \end{bmatrix},$$

whose entries $M_{ij}$ are scalar random variables, the expectation is defined as

$$\mathbb{E}(\mathbf{M}) = \begin{bmatrix} \mathbb{E}(M_{11}) & \mathbb{E}(M_{12}) & \dots & \mathbb{E}(M_{1q}) \\ \mathbb{E}(M_{21}) & \mathbb{E}(M_{22}) & \dots & \mathbb{E}(M_{2q}) \\ \vdots & & & \\ \mathbb{E}(M_{p1}) & \mathbb{E}(M_{p2}) & \dots & \mathbb{E}(M_{pq}) \end{bmatrix}.$$

Note that a random column vector of length $p$ is also a random matrix of dimension $p \times 1$.

The expectation, variance and covariance of multivariate random variables have several useful properties.

**Lemma 2.4.** *Assume that* $\mathbf{A}$ *is an* $m \times p$ *non-random matrix,* $\mathbf{B}$ *is an* $n \times q$ *non-random matrix,* $\mathbf{M}$ *is a* $p \times q$ *random matrix,* $\mathbf{U}$ *is a random vector of length* $p$ *and* $\mathbf{W}$ *is a random vector of length* $q$*. Then:*

$$\mathbb{E}(\mathbf{AU}) = \mathbf{A}\mathbb{E}\mathbf{U}$$

$$\mathbb{E}(\mathbf{AMB}^{\mathrm{T}}) = \mathbf{A}\mathbb{E}(\mathbf{M})\mathbf{B}^{\mathrm{T}}$$

$$\mathrm{Cov}(\mathbf{U}, \mathbf{W}) = \mathbb{E}(\mathbf{U}\mathbf{W}^{\mathrm{T}}) - \mathbb{E}(\mathbf{U})\mathbb{E}(\mathbf{W})^{\mathrm{T}}$$

$$\mathrm{Cov}(\mathbf{AU}, \mathbf{BW}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{U}, \mathbf{W})\mathbf{B}^{\mathrm{T}}$$

$$\mathrm{Var}(\mathbf{AU}) = \mathbf{A}\,\mathrm{Var}(\mathbf{U})\mathbf{A}^{\mathrm{T}}$$

**Definition 2.5.** *A random vector* $\mathbf{W}$ *of length* $p$ *is said to have a* multivariate normal distribution *with mean* $\boldsymbol{\mu} \in \mathbb{R}^p$ *and* $p \times p$ *variance matrix* $\boldsymbol{\Sigma}$*, written*

$$\mathbf{W} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,,$$

*if there exists a $p \times r$ matrix $\mathbf{B}$, with $\mathbf{B}\mathbf{B}^{\mathrm{T}} = \boldsymbol{\Sigma}$, such that*

$$\mathbf{W} = \boldsymbol{\mu} + \mathbf{B}\mathbf{Z}\,,$$

*where $\mathbf{Z} = (Z_1, \ldots, Z_r)^{\mathrm{T}}$ is a random vector of length $r$ whose components are distributed as $Z_i \sim N(0,1)$ independently. (Note that $\mathrm{Var}(\mathbf{Z}) = \mathbf{I}$.)*

Recall that the real symmetric $p \times p$ matrix $\boldsymbol{\Sigma}$ is *non-negative definite* if $\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^p$. The multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be shown to exist for any real symmetric non-negative definite matrix $\boldsymbol{\Sigma}$. The distribution has several important properties.

**Lemma 2.6.** *(i) If $\mathbf{W} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\mathbb{E}(\mathbf{W}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{W}) = \boldsymbol{\Sigma}$.*

*(ii) If also $\mathbf{A}$ is a non-random $m \times p$ matrix, and $\mathbf{c}$ is a non-random vector of length $m$, then*

$$\mathbf{A}\mathbf{W} + \mathbf{c} \sim N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}})$$

Thus if $\mathbf{W}$ is multivariate normal, then any linear transformation of $\mathbf{W}$ is multivariate normal, with mean and variance that are easily calculated.

Let $\mathbf{U} = (U_1, \ldots, U_p)^{\mathrm{T}}$ be an $\mathbb{R}^p$-valued random vector and $\mathbf{V} = (V_1, \ldots, V_q)^{\mathrm{T}}$ be an $\mathbb{R}^q$-valued random vector.

**Definition 2.7.** $\mathbf{U}$ *and* $\mathbf{V}$ *are* independent *if, for all 'suitably nice' subsets $A \subseteq \mathbb{R}^p$, $B \subseteq \mathbb{R}^q$,*

$$\mathbb{P}(\mathbf{U} \in A \text{ and } \mathbf{V} \in B) = \mathbb{P}(\mathbf{U} \in A)\,\mathbb{P}(\mathbf{V} \in B)\,.$$

Defining what is meant precisely by 'suitably nice' requires measure theory, and is outside the scope of this course. However, practically all sets you can actually think of are 'suitably nice'.

A key result is that if two random vectors $\mathbf{U}_1$ and $\mathbf{U}_2$ are *jointly* normally distributed and uncorrelated, then they are independent.

**Proposition 2.8.** *Suppose that*

$$\begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathrm{T}} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

*Then $\mathbf{U}_1$ and $\mathbf{U}_2$ are independent if and only if $\mathrm{Cov}(\mathbf{U}_1, \mathbf{U}_2) = \boldsymbol{\Sigma}_{12} = \mathbf{0}$.*

Care must be taken with this result. Even if $\mathbf{U}_1$ and $\mathbf{U}_2$ are both normally distributed vectors, it does not necessarily follow that $\mathbf{U}_1$ and $\mathbf{U}_2$ are jointly normally distributed. Thus, to show that $\mathbf{U}_1$ and $\mathbf{U}_2$ are independent, it is not enough to show that $\mathbf{U}_1$ is normal and $\mathbf{U}_2$ is normal and $\mathrm{Cov}(\mathbf{U}_1, \mathbf{U}_2) = \mathbf{0}$.

## 2.3 Properties of the least squares estimator in the non-singular case

**Proposition 2.9.** *If $\mathbf{X}$ is non-random and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is non-singular, then the unique least squares estimator, $\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, satisfies:*

*(i) the estimator $\hat{\boldsymbol{\theta}}$ is unbiased, i.e. $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.*

*(ii) the sampling variance of $\hat{\boldsymbol{\theta}}$ is given by*

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,.$$

*(iii) if the errors are normally distributed, i.e. $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then*

$$\hat{\boldsymbol{\theta}} \sim N[\boldsymbol{\theta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]\,.$$

**Example 2.10** (Simple linear regression)**.** *From Section 2.1.1, we know that provided that not all $x_i$ are equal, the information matrix is invertible and*

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}}{s_{xx}} \\ -\frac{\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{bmatrix}.$$

*Hence, by Proposition 2.9 we have that*

$$\mathrm{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$$
$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{s_{xx}}\,.$$

Proposition 2.9 shows that the information matrix plays a crucial role in the performance of least squares estimation. It depends on both the model that we wish to fit and the design that has been used, i.e. the choice of the $x$-values in the experiment. Choosing a good $\mathbf{X}$ is, essentially, one of the main goals of the subject of Optimal Design of Experiments. For further details, see MATH68082, taught by Dr. Alex Donev.

To prove Proposition 2.9, we will need the following result from Linear Algebra:

**Lemma 2.11.** *If $\mathbf{B}$ is a symmetric invertible matrix, then $(\mathbf{B})^{-1}$ is also symmetric.*

*Proof of Lemma 2.11.* This can be seen since since

$$\mathbf{B}(\mathbf{B}^{-1})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}(\mathbf{B}^{-1})^{\mathrm{T}} = (\mathbf{B}^{-1}\mathbf{B})^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}\,.$$

Premultiplying the left hand side by $\mathbf{B}^{-1}$ we obtain

$$(\mathbf{B}^{-1})^{\mathrm{T}} = \mathbf{B}^{-1}\,.$$

□

*Proof of Proposition 2.9.* For part (i), note that

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbb{E}(\mathbf{Y}) \\
&\qquad \text{using } \mathbb{E}(\mathbf{AU}) = \mathbf{A}\mathbb{E}(\mathbf{U}), \text{ as } \mathbf{X} \text{ is non-random} \\
&= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\theta}) \qquad \text{since } \mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta} \\
&= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\theta} \\
&= \mathbf{I}\boldsymbol{\theta} = \boldsymbol{\theta}\,.
\end{aligned}$$

For part (ii), let $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ and note that

$$\begin{aligned}
\mathbf{A}^{\mathrm{T}} &= [(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}]^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]^{\mathrm{T}} \\
&= \mathbf{X}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]^{\mathrm{T}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \qquad (2.5) \\
&\qquad \text{since } \mathbf{X}^{\mathrm{T}}\mathbf{X} \text{ is symmetric.}
\end{aligned}$$

Hence we have that

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\theta}}) &= \mathrm{Var}[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}] = \mathrm{Var}(\mathbf{AY}) = \mathbf{A}\,\mathrm{Var}(\mathbf{Y})\mathbf{A}^{\mathrm{T}} \\
&= \mathbf{A}(\sigma^2\mathbf{I})\mathbf{A}^{\mathrm{T}} = \sigma^2\mathbf{A}\mathbf{A}^{\mathrm{T}} \\
&= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \qquad \text{using (2.5)} \\
&= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,.
\end{aligned}$$

For part (iii), note that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \sim N_n(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I})$ by Lemma 2.6 and similarly

$$\hat{\boldsymbol{\theta}} = \mathbf{AY} \sim N_p(\mathbf{AX}\boldsymbol{\theta}, \sigma^2\mathbf{A}\mathbf{A}^{\mathrm{T}})$$

Hence $\hat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$ as we have shown above that $\mathbf{AX}\boldsymbol{\theta} = \boldsymbol{\theta}$ and $\mathbf{A}\mathbf{A}^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. □

## 2.4 The singular case

If $\det(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = 0$, then the information matrix is singular and has no inverse. In this case the normal equations have infinitely many solutions. We illustrate this with an example.

### 2.4.1 Illustrative example: One-way classification model

Suppose we have a categorical variable with two levels, and two observations per level. In Chapter 1, we used the model

$$Y_{ij} = \mu_i + \epsilon_{ij}\,, \qquad i = 1, 2; j = 1, 2.$$

for this set up, where $Y_{ij}$ is the response for the $j$th observation in the $i$th group. This model is 'identifiable', i.e. there will be a unique solution to the normal equations.

However, another common model is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}\,, \quad i = 1, 2;\; j = 1, 2\,,$$

where

- $\mu$ is an intercept term
- $\tau_i$ denotes the effect of the $i$th treatment
- $\epsilon_{ij}$ is the random error term

The new model is 'over-parameterized': it has 3 parameters ($\mu$, $\tau_1$, $\tau_2$) even though there are only two treatment groups. As a result many different values of the parameters give the same predictions. For example if $\tau_1$ and $\tau_2$ are replaced by $\tau_1 + 1$, $\tau_2 + 1$ and $\mu$ is replaced by $\mu - 1$ then $\mathbb{E}(Y)$ remains the same, since

$$\mathbb{E}(Y) \to (\mu + 1) + (\tau_i - 1) = \mu + \tau_i\,.$$

As a consequence of this, the normal equations have no unique solution; indeed they have infinitely many.

To see this in detail consider the matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \end{bmatrix}.$$

The normal equations $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$ are

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{bmatrix} = \begin{bmatrix} Y_{1\cdot} + Y_{2\cdot} \\ Y_{1\cdot} \\ Y_{2\cdot} \end{bmatrix}$$

where $Y_{1\cdot} = \sum_j Y_{1j}$, $Y_{2\cdot} = \sum_j Y_{2j}$. Note that $\det(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = 4(4) - 2(4) + 2(-4) = 16 - 8 - 8 = 0$, hence the information matrix is singular and there is no unique solution. Moreover,

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

so if $\hat{\boldsymbol{\theta}}$ is a solution then so too is $\hat{\boldsymbol{\theta}} + \lambda(1, -1, -1)^{\mathrm{T}}$ because

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\theta}} + \lambda \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}) = \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\theta}} + \lambda\mathbf{X}^{\mathrm{T}}\mathbf{X} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{0} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}\,.$$

By varying $\lambda$ we can obtain infinitely many solutions.

### 2.4.2 Use of constraints to find a solution

Often it is possible to obtain a unique solution to the normal equations in the singular case by imposing additional constraints on the solution. For example, in the classification model above, in practice there are three common choices of constraint:

1. *Zero constraint on the intercept*, i.e. $\hat{\mu} = 0$.

2. *Zero constraint on the effect of the first treatment*, i.e. $\hat{\tau}_1 = 0$.

3. *Sum-to-zero constraint* on the treatment effects, i.e. $\hat{\tau}_1 + \hat{\tau}_2 = 0$.

Although the constraint method works in many cases, in general it can be difficult to identify a suitable set of constraints that yields a unique solution of the normal equations. Moreover, it is difficult to assess which statistical conclusions are independent of the choice of constraint.

A more systematic approach is to use a 'generalized inverse' (see next section). This also enables theoretical results to be developed to clarify which statistical conclusions are independent of the choice of a particular solution, and to describe the properties of estimators, hypothesis tests and so on.

### 2.4.3 Least squares estimation via $g$-inverses

In the non-singular case, we saw that the unique least squares estimator was

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,.$$

In the singular case, the normal equations have infinitely many solutions. Moreover, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible, so we can not define a solution as above.

However, it is natural to consider an estimator of the similar form

$$\hat{\boldsymbol{\theta}}_G = \mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,,$$

with $\mathbf{G}$ a non-random constant matrix (i.e. not depending on $\mathbf{Y}$). An interesting question is whether it is possible to find a $\mathbf{G}$ such that $\hat{\boldsymbol{\theta}}_G$ is a least squares estimate of $\boldsymbol{\theta}$ *whatever the value of* $\mathbf{Y}$. In other words, can we find a $\mathbf{G}$ such that $\hat{\boldsymbol{\theta}}_G$ is a solution to the normal equations for all possible values of $\mathbf{Y}$?

To answer this question, we need the idea of a 'generalized inverse', or $g$-inverse.

**Definition 2.12.** *Given an $m \times n$ matrix $\mathbf{A}$, the $n \times m$ matrix $\mathbf{G}$ is a $g$-inverse of $\mathbf{A}$ if*

$$\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A}\,.$$

*The notation $\mathbf{A}^-$ is often used for a $g$-inverse of $\mathbf{A}$.*

A $g$-inverse of $\mathbf{A}$ always exists but is typically not unique, unless $\mathbf{A}$ is invertible, in which case the unique $g$-inverse is the regular inverse $\mathbf{A}^{-1}$. The importance of $g$-inverses in the theory of Linear Models is the following:

**Proposition 2.13.** *The estimator $\hat{\boldsymbol{\theta}}_G = \mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is a solution of the normal equations for all values of $\mathbf{Y} \in \mathbb{R}^n$ if and only if $\mathbf{G}$ is a $g$-inverse of the information matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.*

Thus, provided we can find a $g$-inverse of the information matrix, although the normal equations do not have a unique solution, we are able to obtain a particular solution. Moreover, the estimator $\hat{\boldsymbol{\theta}}_G$ is a linear transformation of $\mathbf{Y}$, so we will be able to compute its statistical properties (see Example Sheet 3, Q2).

We need to take care to ensure that the statistical conclusions we make are independent of the choice of $g$-inverse. This will be the case provided we restrict ourselves to so-called 'estimable' functions of the parameters (see Chapter 4).

## 2.5 Computation of a $g$-inverse

**Procedure 2.14.** *To compute a $g$-inverse, $\mathbf{G}$, of a matrix $\mathbf{A}$ with $\mathrm{rank}(\mathbf{A}) = r$:*

1. *Delete rows and columns of $\mathbf{A}$ until an invertible $r \times r$ matrix $\mathbf{B}$ remains;*

2. *Calculate $\mathbf{B}^{-1}$;*

3. *For each deleted* ***row*** *of* $\mathbf{A}$*, fill the corresponding* ***column*** *of* $\mathbf{G}$ *with zeroes;*

4. *For each deleted* ***column*** *of* $\mathbf{A}$*, fill the corresponding* ***row*** *of* $\mathbf{G}$ *with zeroes;*

5. *Fill the remaining entries with the entries of* $\mathbf{B}^{-1}$*.*

Note that for a given matrix, there will be several possible choices of submatrix $\mathbf{B}$, leading to several different $g$-inverses. However, this method does not give all possible $g$-inverses.

**Example 2.15.** *Find a* $g$*-inverse of the* $3 \times 4$ *matrix*

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 & 0 \\ 1 & 1 & 5 & 15 \\ 3 & 1 & 3 & 5 \end{bmatrix}.$$

Note that Gaussian elimination shows there are two independent rows, so that $\mathrm{rank}(\mathbf{A}) = 2$. Thus we need to identify an invertible $2 \times 2$ submatrix. Deleting Columns 1 and 3 and Row 1, we obtain the submatrix

$$\mathbf{B} = \begin{bmatrix} 1 & 15 \\ 1 & 5 \end{bmatrix}$$

which has $\det(\mathbf{B}) = -10$ and

$$\mathbf{B}^{-1} = \begin{bmatrix} -\frac{1}{2} & \frac{3}{2} \\ \frac{1}{10} & -\frac{1}{10} \end{bmatrix}.$$

Using the above procedure, the corresponding $g$-inverse is

$$\mathbf{A}^{-} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{3}{2} \\ 0 & 0 & 0 \\ 0 & \frac{1}{10} & -\frac{1}{10} \end{bmatrix}.$$

Note that as *Columns* 1 and 3 were deleted from $\mathbf{A}$, $\mathbf{A}^{-}$ has zeroes in *Rows* 1 and 3. As *Row* 1 was deleted from $\mathbf{A}$, *Column* 1 of $\mathbf{A}^{-}$ is filled with zeroes.

# 3 The fitted model and residuals

## 3.1 Fitted values and the hat matrix

Recall the model equation,

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}\,.$$

Suppose that we have a $g$-inverse, $\mathbf{G}$, of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and the corresponding least squares estimate, $\hat{\boldsymbol{\theta}}_G = \mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, of the parameters, $\boldsymbol{\theta}$. Then the fitted *fitted values* are defined as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}_G\,,$$

i.e. we substitute the estimated parameter values into the model equation to estimate $\mathbb{E}(\mathbf{Y})$.

Note using the definition of $\hat{\boldsymbol{\theta}}_G$ that

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}\mathbf{Y}\,,$$

where above the *hat matrix*, $\mathbf{H}$, is defined as $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}$. It is called the 'hat matrix' because it puts the hat on $\mathbf{Y}$. It is of dimension $n \times n$. In the full rank case, $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and so $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is defined uniquely. However, in the rank-deficient case naïvely it appears that the above definition of the hat matrix depends on the choice of $g$-inverse $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$. Proposition 3.1 below shows that, surprisingly, it does not, and so the hat matrix is well-defined. It also gives many other important properties of the hat matrix, which will be useful later when proving properties about residuals, sums of squares, and the Gauss-Markov theorem in Chapter 4.

**Proposition 3.1** (Properties of the hat matrix). *If $\mathbf{G}$ is a $g$-inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}$ is the hat matrix, then:*

*(i)* $\mathbf{H}\mathbf{X} = \mathbf{X}$*;*

*(ii)* $\mathbf{H}$ *does not depend on the choice of* $g$*-inverse;*

*(iii)* $\mathbf{H} = \mathbf{H}^{\mathrm{T}}$ *is symmetric;*

*(iv)* $\mathbf{H}^2 = \mathbf{H}$*, i.e.* $\mathbf{H}$ *is idempotent;*

*(v)* $\mathbf{I} - \mathbf{H}$ *is symmetric and idempotent;*

*(vi)* $\mathrm{tr}(\mathbf{H}) = \mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$.

Note that property (i) also implies that the fitted values $\hat{\mathbf{Y}} = \mathbf{HY}$ are well-defined independently of the choice of $g$-inverse. Property (ii) implies that $\mathbf{XGX}^{\mathrm{T}} = \mathbf{XG}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}$ even if $\mathbf{G}$ is not symmetric, which usually it will not be.

In the case where $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, the proof of Proposition 3.1 is easy (see Example Sheet 3, Q4). In the singular case, the proof is rather long and technical, and does not provide you with many more statistical skills. As a result, the proof in the non-singular case is non-examinable, but it is included in the additional notes for this chapter for the interested reader (it is quite an interesting piece of linear algebra!).

The following result shows the usefulness of the hat matrix in calculating the statistical properties of the fitted values:

**Proposition 3.2** (Properties of fitted values). *The fitted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}_G = \mathbf{HY}$ have the following properties:*

*(i)* $\mathbb{E}(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\theta}$

*(ii)* $\mathrm{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$.

*(iii) If $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$, then $\hat{\mathbf{Y}} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{H})$.*

Hence, for example,

$$\mathrm{Var}(\hat{Y}_i) = \sigma^2 h_{ii} = \sigma^2(\mathbf{xGx}^{\mathrm{T}})_{ii} = \sigma^2\mathbf{x}_i\mathbf{G}\mathbf{x}_i^{\mathrm{T}}$$

$$\mathrm{Cov}(\hat{Y}_i, \hat{Y}_j) = \sigma^2 h_{ij} = \sigma^2(\mathbf{xGx}^{\mathrm{T}})_{ij} = \sigma^2\mathbf{x}_i\mathbf{G}\mathbf{x}_j^{\mathrm{T}}$$

$$\mathrm{Corr}(\hat{Y}_i, \hat{Y}_j) = \frac{\mathrm{Cov}(\hat{Y}_i, \hat{Y}_j)}{\sqrt{\mathrm{Var}(\hat{Y}_i)\,\mathrm{Var}(\hat{Y}_j)}} = \frac{h_{ij}}{\sqrt{h_{ii}h_{jj}}}\,.$$

where $\mathbf{x}_k$ denotes the $k^{\text{th}}$ row of $\mathbf{X}$, for $k = 1, \ldots, n$.

The proof of Proposition 3.2 is left as an exercise for the reader (see Example Sheet 3, Q5).

## 3.2 Residuals

The *residuals* are defined as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}\,.$$

As the notation suggests, the residuals can be considered to be an estimate of the errors $\boldsymbol{\epsilon}$. However, it must be emphasized that **the residuals $\hat{\epsilon}$ are different from the errors $\epsilon$, and have different**

**properties**. In particular, although the *errors* are uncorrelated with constant variance, i.e. $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$, the *residuals* are correlated and have different variance, as shown below.

**Proposition 3.3.** *The residuals satisfy the following:*

*(i)* $\mathbb{E}(\hat{\boldsymbol{\epsilon}}) = \mathbf{0}$;

*(ii)* $\mathrm{Var}(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$, *in other words*

$$\mathrm{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$
$$\mathrm{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma^2 h_{ij}\,, \quad \text{for } i \neq j\,,$$

*where* $h_{ij}$ *is the* $(i,j)$*th entry of* $\mathbf{H}$;

*(iii) If* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, *then* $\hat{\boldsymbol{\epsilon}} \sim N[\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})]$;

*(iv)* $\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \mathbf{0}$;

*(v)* $\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = 0$.

Note that if the model contains an intercept term, then one of the columns of $\mathbf{X}$ will be $(1, 1, \ldots, 1)^{\mathrm{T}}$ and so property (iv) above will imply that $\sum_{i=1}^n \hat{\epsilon}_i = 0$.

*Proof.* For parts (i) and (ii),

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\epsilon}}) &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\mathbb{E}(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\theta} \\
&= \mathbf{X}\boldsymbol{\theta} - \mathbf{H}\mathbf{X}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\theta} \quad (\text{as } \mathbf{H}\mathbf{X} = \mathbf{X}) \\
&= \mathbf{0}\,, \\
\mathrm{Var}(\hat{\boldsymbol{\epsilon}}) &= \mathrm{Var}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\,\mathrm{Var}[\mathbf{Y}](\mathbf{I} - \mathbf{H})^{\mathrm{T}} \\
&= (\mathbf{I} - \mathbf{H})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{H})^{\mathrm{T}} = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \\
&\qquad \text{as } \mathbf{I} - \mathbf{H} \text{ is symmetric} \\
&= \sigma^2(\mathbf{I} - \mathbf{H}) \quad \text{as } \mathbf{I} - \mathbf{H} \text{ is idempotent.}
\end{aligned}$$

For part (iii), note that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is multivariate normal, then $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\theta} + (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$ is a linear transformation of $\boldsymbol{\epsilon}$ and so also multivariate normal by Lemma 2.6(ii). Specifically $\hat{\boldsymbol{\epsilon}} \sim N_n\left(\mathbb{E}(\hat{\boldsymbol{\epsilon}}), \mathrm{Var}(\hat{\boldsymbol{\epsilon}})\right) = N_n\left(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})\right)$ by part (i). For part (iv), note that $\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_G) = \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\theta}}_G = \mathbf{0}$, as $\hat{\boldsymbol{\theta}}_G$ is a solution of the normal equations. Part (v) follows immediately, since $\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = (\mathbf{X}\hat{\boldsymbol{\theta}}_G)^{\mathrm{T}}\boldsymbol{\epsilon} = \hat{\boldsymbol{\theta}}_G^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\theta}}_G^{\mathrm{T}}\mathbf{0} = 0$ by (iv). □

**Example: simple linear regression**

As we have seen in Section 2.1.1, for the simple linear regression model,

$$Y_i = \alpha + \beta x_i + \epsilon_i\,, \quad i = 1, \ldots, n,$$

with $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, $i = 1, \ldots, n$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$, we have

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad \mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

and provided there are at least two distinct values among the $x_i$, the information matrix is non-singular with

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}}{s_{xx}} \\ -\frac{\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{pmatrix}.$$

Hence we have that the only $g$-inverse is $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and the covariances of the fitted values are given by

$$\begin{aligned}
\mathrm{Cov}(\hat{Y}_i, \hat{Y}_j) = \sigma^2 h_{ij} &= \sigma^2 \mathbf{x}_i^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{x}_j \\
&= \sigma^2 \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}}{s_{xx}} \\ -\frac{\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{pmatrix} \begin{pmatrix} 1 \\ x_j \end{pmatrix} \\
&= \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} - \frac{\bar{x}x_j}{s_{xx}} - \frac{\bar{x}x_i}{s_{xx}} + \frac{x_i x_j}{s_{xx}} \right\} \\
&= \sigma^2 \left\{ \frac{1}{n} + \frac{1}{s_{xx}}(x_i x_j - x_i\bar{x} - x_j\bar{x} + \bar{x}^2) \right\} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{s_{xx}}(x_i - \bar{x})(x_j - \bar{x})\,, \\
\mathrm{Var}(\hat{Y}_i) = \sigma^2 h_{ii} &= \frac{\sigma^2}{n} + \frac{\sigma^2}{s_{xx}}(x_i - \bar{x})^2\,.
\end{aligned}$$

Similarly, the covariances of the residuals are

$$\begin{aligned}
\mathrm{Cov}(\hat{\epsilon}_i, \hat{\epsilon}_j) &= -\sigma^2 h_{ij} = -\frac{\sigma^2}{n} - \frac{\sigma^2}{s_{xx}}(x_i - \bar{x})(x_j - \bar{x}) \\
\mathrm{Var}(\hat{\epsilon}_i) &= \sigma^2(1 - h_{ii}) = \frac{(n-1)\sigma^2}{n} - \frac{\sigma^2}{s_{xx}}(x_i - \bar{x})^2\,.
\end{aligned}$$

## 3.3 Variance estimation

We often wish to estimate the variance $\sigma^2$, either because this is of direct interest, or because it is needed when making inferences about the parameters $\boldsymbol{\theta}$, e.g. in the calculation of confidence intervals and hypothesis tests in Chapters 5 and 6.

We first define the *residual sum of squares*,

$$\text{SSE} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^{\mathrm{T}} \hat{\boldsymbol{\epsilon}} \,.$$

Note that, as $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, and $\mathbf{I} - \mathbf{H}$ is idempotent by Proposition 3.1(v), an alternative expression is

$$\begin{aligned} \text{SSE} &= \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\theta}}_G \,. \end{aligned}$$

The final formula can used to compute the RSS without explicitly calculating the residuals, which is useful: this is the formula most used in practice.

The primary aim of this section will be to prove the following.

**Proposition 3.4.** *Let $r = \operatorname{rank}(\mathbf{X})$. The estimator*

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - r}$$

*is an unbiased estimator of $\sigma^2$.*

Note that if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, then $\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = p$ and so in the non-singular case the above becomes

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - p} \,.$$

Before proving Proposition 3.4 we recall some general properties of the trace of a matrix.

**Lemma 3.5.** *(i) If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{B}$ is an $m \times n$ matrix, then $\mathbf{AB}$ and $\mathbf{BA}$ exist and*

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}) \,.$$

*(ii) If $\mathbf{A}$ is a random $n \times m$ matrix, then*

$$\mathbb{E}(\operatorname{tr} \mathbf{A}) = \operatorname{tr}(\mathbb{E}\mathbf{A}) \,.$$

*(iii) If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $n \times m$ *matrices, and* $c$ *is a scalar (a* $1 \times 1$ *matrix), then*

$$\begin{aligned}\operatorname{tr}(\mathbf{A}+\mathbf{B}) &= \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) \\ \operatorname{tr}(c\mathbf{A}) &= c \operatorname{tr}(\mathbf{A}) \\ \operatorname{tr}(c) &= c\,.\end{aligned}$$

*Proof of Proposition 3.4.*

$$\begin{aligned}\mathbb{E}[\mathsf{SSE}] &= \mathbb{E}[\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}}] = \mathbb{E}(\operatorname{tr}[\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}}]) \quad \text{as } c = \operatorname{tr}(c) \text{ and } \hat{\boldsymbol{\epsilon}}^{\mathrm{T}}\hat{\boldsymbol{\epsilon}} \text{ is a scalar} \\ &= \mathbb{E}(\operatorname{tr}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}]) \quad \text{since } \operatorname{tr}\mathbf{AB} = \operatorname{tr}\mathbf{BA} \\ &= \operatorname{tr}(\mathbb{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}]) \quad \text{by part (ii) of the above Lemma.} \qquad (3.1)\end{aligned}$$

However, note that by Proposition 3.3 and Lemma 2.4 we have

$$\begin{aligned}\sigma^2(\mathbf{I}_{n\times n} - \mathbf{H}) &= \operatorname{Var}(\hat{\boldsymbol{\epsilon}}) = \operatorname{Cov}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}) = \mathbb{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}] - \mathbb{E}[\hat{\boldsymbol{\epsilon}}]\mathbb{E}[\hat{\boldsymbol{\epsilon}}]^{\mathrm{T}} \\ &= \mathbb{E}[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathrm{T}}] - \mathbf{0}\mathbf{0}^{\mathrm{T}}\,,\end{aligned}$$

so that $\mathbb{E}\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^{\mathrm{T}} = \sigma^2(\mathbf{I}_{n\times n} - \mathbf{H})$. Combining this with 3.1 we have

$$\begin{aligned}\mathbb{E}[\mathsf{SSE}] &= \operatorname{tr}[\sigma^2(\mathbf{I}_{n\times n} - \mathbf{H})] = \sigma^2 \operatorname{tr}(\mathbf{I}_{n\times n} - \mathbf{H}) \\ &= \sigma^2[\operatorname{tr}(\mathbf{I}_{n\times n}) - \operatorname{tr}\mathbf{H}] \\ &= \sigma^2[n - \operatorname{rank}(\mathbf{X})] \\ &= \sigma^2[n-r]\,.\end{aligned}$$

Hence

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left[\frac{\mathsf{SSE}}{n-r}\right] = \frac{1}{n-r}\mathbb{E}[\mathsf{SSE}] = \frac{(n-r)\sigma^2}{(n-r)} = \sigma^2\,,$$

and so $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2$ as claimed. □

**Standard errors**

In the non-singular case, $\hat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$ with $\operatorname{Var}\hat{\boldsymbol{\theta}} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Moreover $\hat{\theta}_j$ is unbiased for $\theta_j$ and

$$\text{s. d.}(\hat{\theta}_j) = \sqrt{\operatorname{Var}(\hat{\theta}_j)} = \sqrt{[\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]_{jj}} = \sigma\sqrt{m^{(jj)}}\,,$$

where $m^{(jj)}$ denotes the $j,j$th entry of the inverse information matrix $\mathbf{M}^{-1} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. (Be careful: this is different from the inverse of the $j,j$th entry of the information matrix, $m_{jj}^{-1}$!). However, the above is unknown due to dependence on $\sigma^2$.

The *standard error* of $\hat{\theta}_j$ is obtained by substituting the unbiased estimate of $\sigma^2$ above, giving

$$\mathrm{s.e.}(\hat{\theta}_j) = \hat{\sigma}\sqrt{m^{(jj)}}\,.$$

This will be useful later, e.g. we will see that $\hat{\theta}_j \pm t_{\frac{\alpha}{2};n-p}\,\mathrm{s.e.}(\hat{\theta}_j)$ gives a $100(1-\alpha)\%$ confidence interval for $\theta_j$.

## 3.4 Residual diagnostics

Before carrying out any inferences (e.g. confidence intervals or hypothesis tests) it is a good idea to check the validity of the model assumptions, namely:

- *Structural form of the model* – does the model need any additional terms? Are transformations of the response or predictors needed?
- *Random error assumptions* – do the assumptions of constant variance, normality, and independence hold?

This can be done using residual diagnostic plots. The most useful of these are:

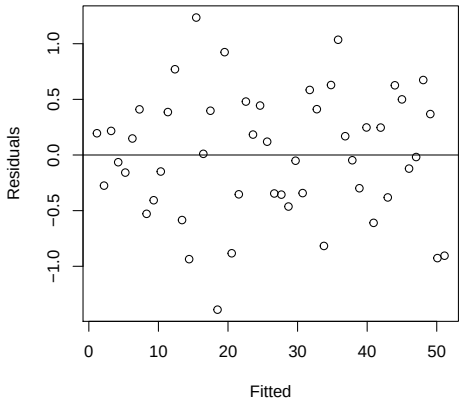
- plot of $\hat{y}$ vs $\hat{\epsilon}$
- plot of each predictor $x_j$ vs $\hat{\epsilon}$
- qq-plot of the residuals

### Plot of $\hat{y}$ vs $\hat{\epsilon}$

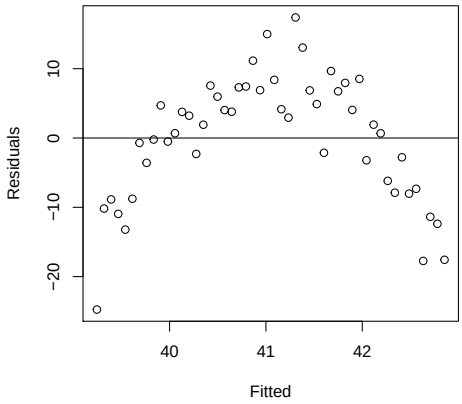
This is a scatter plot of the points $(\hat{y}_i, \hat{\epsilon}_i)$, $i = 1, \ldots, n$. It can be made in R as follows:

```
simple_fit <- lm(y~x)
plot(fitted(simple_fit), residuals(simple_fit), xlab="Fitted", ylab="Residuals")
abline(h=0)
```

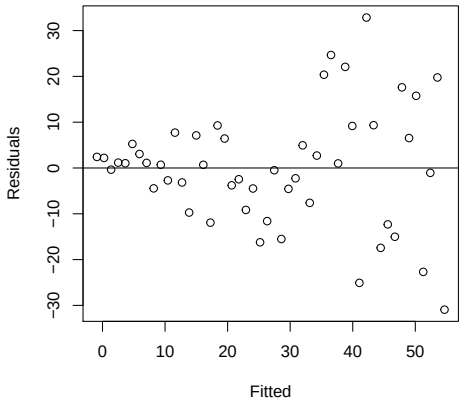
If all the assumptions are correct, the points should be randomly scattered around the line $\hat{\epsilon} = 0$, similar to the following:




A systematic trend may indicate a problem with the structural form of the model, e.g. in the following a simple linear model has been fitted, but a quadratic model is a better description of the data:




Such problems can often be improved by including extra terms in the model. Patterns in the spread of the residuals indicate that the constant variance assumption is incorrect. In the figure below, the variance is an increasing function of the mean:

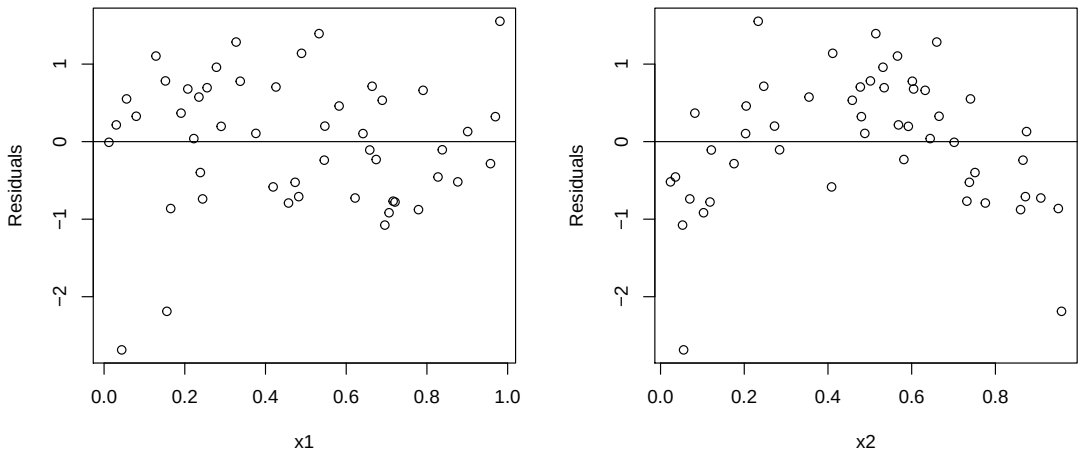



Non-constant variance can sometimes be eliminated by transforming the response (e.g. using a log transformation or a Box-Cox transformation), if not then it should be accounted for in the model and the inference, e.g. by using weighted least squares.

### Plot of each predictor versus $\hat{\epsilon}$

Here we plot $(x_{ij}, y_i)$, $i = 1, \ldots, n$. There is a different plot for each predictor variable. If the assumptions are correct, then the plot for each variable should consist of points randomly scattered around the line $\hat{\epsilon} = 0$.

Systematic trends may indicate that extra terms need to be added to the model. For example, here are plots for the model $y = \alpha + \beta x_1 + \gamma x_2 + \epsilon$. It is clear that an $x_2^2$ term may improve the model.



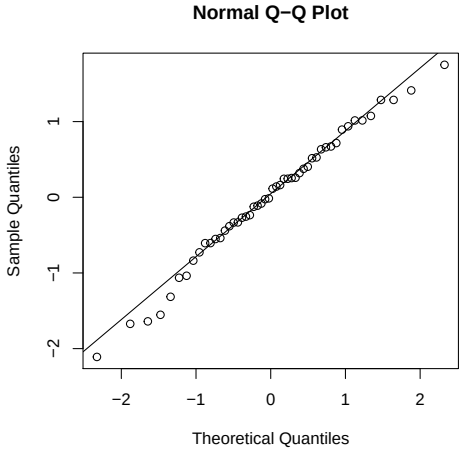

Patterns in the spread of the residuals in the plot for $x_j$ indicate that the variance may depend on $x_j$.

### qq-plot of the residuals

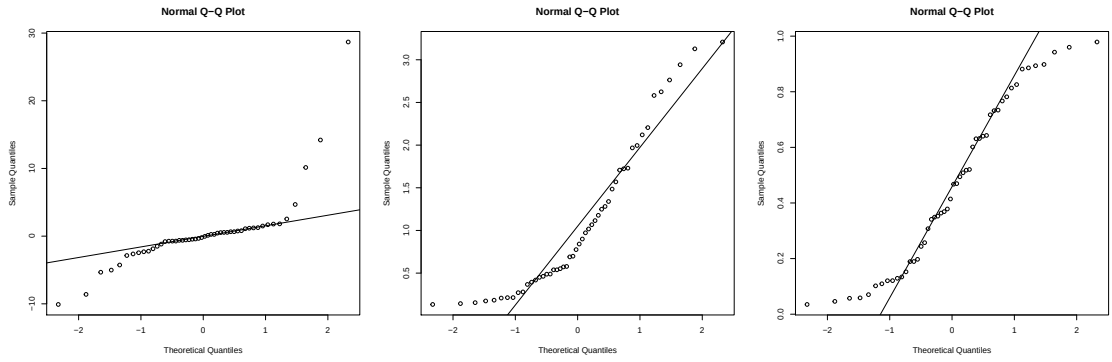
A qq-plot is used to check the normality assumption. We plot the sorted residuals against the theoretical quantiles of the normal distribution, $\Phi^{-1}(\frac{i}{n+1})$. If the normality assumption is correct, then the points should approximately follow a straight line. Such a plot can be produced in R as follows:

```
simple_fit <- lm(y~x)
qqnorm(residuals(simple_fit)); qqline(residuals(simple_fit))
```

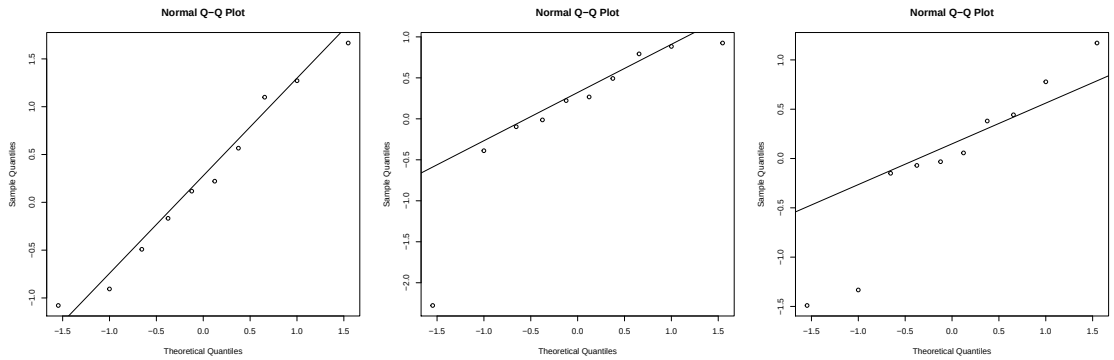
Here is an example where the residuals look normal:




To give you an idea of what can go wrong in a qq-plot, below we show qq-plots for random samples from three different non-normal distributions (from left to right, a heavy-tailed, skewed, and light-tailed distribution).




Note that it is difficult to make any conclusions when the sample size is small: to see this, consider the plots below. Each of these corresponds to a random sample from a $N(0,1)$ with $n = 10$.




### 3.4.1 Studentized residuals

There is an issue with using the raw residuals $\hat{\epsilon}_i$ in diagnostic plots, namely that the variance $\mathrm{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ is known to be non-constant. Therefore patterns in the spread of the residuals could, in principle, be due to trends in $(1 - h_{ii})$ rather than non-constant variance of the random errors. In practice, this is very rare as the terms $(1 - h_{ii})$ are usually not that different.

As a result it may be better in diagnostic plots to use *standardized residuals*,

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\sqrt{\tilde{\sigma}^2(1 - h_{ii})}},$$

where $\tilde{\sigma}^2$ is an estimate of the variance. These have approximately constant variance. There are two main methods.

- *Internal studentization*. The $i$th internally studentized residual is

  $$r_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}},$$

  where $\hat{\sigma}^2 = \mathrm{RSS}/(n - r)$ is the variance estimate from Section 3.3. These can be computed using `rstandard()` in R.

  These should be approximately distributed as $r_i \sim N(0, 1)$ if the model is correct and $n - r$ is large. Thus in practice the internally studentized residuals are usually compared with a standard normal distribution. However, if $n - r$ is not large, then such comparisons may not be justified.

- *External studentization*. The $i$th externally studentized residual is

  $$t_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}},$$

  where $\hat{\sigma}^2_{(i)}$ denotes the unbiased estimate of $\sigma^2$ computed from the regression after omitting the $i$th case from the data. The $t_i$ can be computed using `rstudent()` in R. If the model is correct, then $t_i \sim t(n - 1 - p)$.

## 3.5 Coefficients of determination

In this section, we introduce the coefficients of determination $R^2$ and $R_a^2$. These are commonly used to measure the predictive power of a fitted model. First we must introduce the idea of corrected sums of squares.

### 3.5.1 Corrected sums of squares

For models with an intercept, we define the *corrected total sum of squares* as $\mathsf{SST}_c = \sum_{i=1}^n (Y_i - \bar{Y})^2 \geq 0$ and the corrected regression sum of squares as $\mathsf{SSR}_c = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \geq 0$.

**Proposition 3.6.** *If the model contains an intercept term, then $SST_c = SSR_c + SSE$.*

*Proof.* To see this, note that:

$$\begin{aligned}
\mathsf{SST}_c &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 \\
&= \sum_{i=1}^n \left[ (Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2 \right] \\
&= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^n \hat{\epsilon}_i (\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\
&= \mathsf{SSE} + 2\sum_{i=1}^n \hat{\epsilon}_i \hat{Y}_i - 2\sum_{i=1}^n \hat{\epsilon}_i \bar{Y} + \mathsf{SSR}_c \,.
\end{aligned}$$

Note that $\sum_{i=1}^n \hat{\epsilon}_i \hat{Y}_i = \hat{\boldsymbol{\epsilon}}^{\mathrm{T}} \hat{\mathbf{Y}} = 0$ by Proposition 3.3(v). In addition $\sum_{i=1}^n \hat{\epsilon}_i \bar{Y} = \bar{Y} \sum_{i=1}^n \hat{\epsilon}_i = 0$ using Proposition 3.3(iv) together with the assumption that model contains an intercept so $\mathbf{X}$ contains a column of ones. □

### 3.5.2 R-squared and adjusted R-squared

The coefficients of determination are a measure of the predictive power of a fitted model. The first of these is

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

**Proposition 3.7.** *The coefficient of determination satisfies $0 \leq R^2 \leq 1$.*

*Proof.* To see this, first note that $\mathsf{SSR}_c, \mathsf{SST}_c \geq 0$, so $R^2 = \mathsf{SSR}_c / \mathsf{SST}_c \geq 0$. Secondly, by Proposition 3.6 we have that

$$R^2 = \frac{\mathsf{SST}_c - \mathsf{SSE}}{\mathsf{SST}_c} = 1 - \frac{\mathsf{SSE}}{\mathsf{SST}_c} \,.$$

Since $\mathsf{SSE}, \mathsf{SST}_c \geq 0$, we have that $R^2 \leq 1$. □

$R^2$ is interpreted as the proportion of variability in the data that is explained by the regression model. If $R^2 \approx 1$ then the model has high predictive power, while $R^2 \approx 0$ indicates low predictive power. High $R^2$ does not necessarily indicate a 'good fit': as we will see, the model can still fail a lack-of-fit test even with $R^2 \approx 1$. Similarly, a value $R^2 \approx 0$ does not necessarily indicate a lack of fit; low $R^2$ can occur when the model is correct and $\sigma$ is large compared to the $\theta_j$.

$R^2$ will always increase if an additional predictor is included in the model, even if the predictor is not significantly associated with the response. Hence $R^2$ cannot be used to compare competing models. A better way to compare models is using *adjusted* $R^2$:

$$R_a^2 = 1 - \frac{\mathsf{SSE}/(n-r)}{\mathsf{SST}_c/(n-1)} = 1 - \frac{\hat{\sigma}^2}{\mathsf{SST}_c/(n-1)}\,.$$

The latter will increase only if the additional predictor reduces $\hat{\sigma}^2$. However it is even better to use hypothesis tests or information criteria such as AIC or BIC.

It is straightforward to get R to calculate $R^2$ and $R_a^2$ using the `summary()` command.

## 3.6 Leverage and influential points

It is important to consider the sensitivity of the model fit to the values of the different observations. Sometimes there are a small number of influential observations which drive the model fit. These influential observations should be checked carefully: if they are erroneous, then the fitted model may be inaccurate.

### 3.6.1 Leverage

The *leverage* of the $i$th observation is defined as $h_{ii}$, the $i$th diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}$. Note that $\hat{Y}_i = \sum_{j=1}^n h_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j$ , and so

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}\,.$$

Thus the leverage measures the sensitivity of the fitted value at $\mathbf{x}_i$ to small changes in $Y_i$. An observation with high leverage is known as a *leverage point*.

To quantify 'high leverage', note that $0 \leq h_{ii} \leq 1$ because $\mathrm{Var}(\hat{Y}_i) = \sigma^2 h_{ii} \geq 0$ and $\mathrm{Var}(\hat{\epsilon}_i) = \sigma^2(1-h_{ii}) \geq 0$. Moreover, $\sum_i^n h_{ii} = \mathrm{tr}(\mathbf{H}) = \mathrm{rank}(\mathbf{X}) = r$. Thus the mean leverage is $\frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{r}{n}$. This suggests that observations with $h_{ii}$ close to 1, or $h_{ii}$ 'significantly' greater than $\frac{r}{n}$ should be considered as high leverage points. A common rule of thumb is to treat the observation as a leverage point if $h_{ii} > \frac{2r}{n}$. A leverage point may potentially be influential in the model fit, but is not necessarily so.

### 3.6.2 Influence

If an observation has both high leverage and a large residual, then it will substantially affect the parameter estimates of the fitted model. This idea of influence is quantified by *Cook's distance*, which measures how much the fitted values change when the $i$th observation is removed:

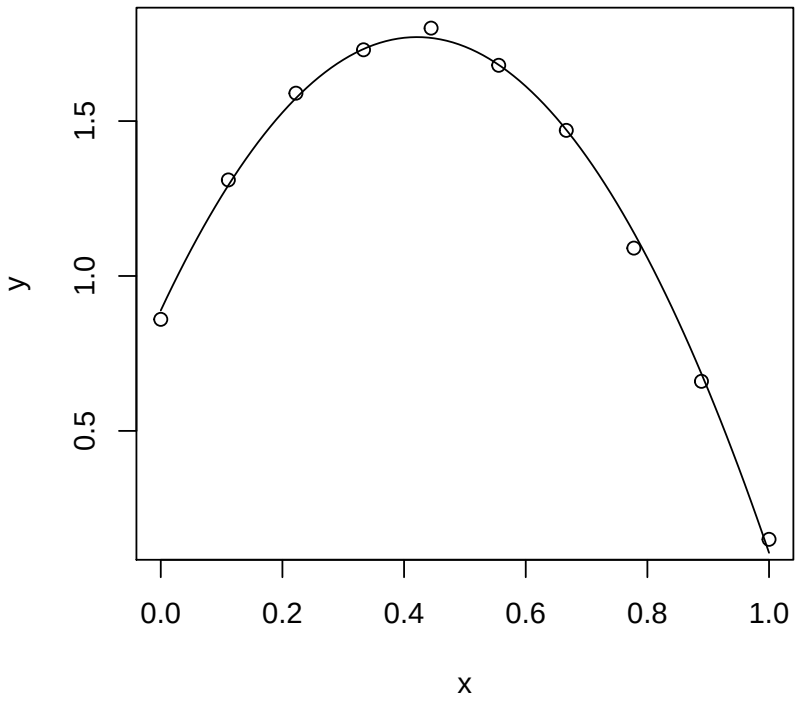
$$D_i = \frac{||\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)}||^2}{p\hat{\sigma}^2} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}\,, \quad i = 1, \ldots, n\,,$$

where $\hat{\mathbf{Y}}_{(i)}$ denotes the vector of fitted values obtained after leaving out the $i$th observation, and $r_i$ denotes the $i$th (internally) studentized residual.

An observation is deemed *influential* if $D_i > 1$. It is usually worth refitting the model with influential observations excluded to see what impact this has on the fitted model. There exist several other related measures of influence, such as DFFITS and DFBETAS, which we do not have time to discuss.

### 3.6.3 Example: chemistry data

We are given data (file `chemistry.csv` on Blackboard) from a chemistry experiment consisting of measurements of reagent concentration (mol dm$^{-3}$, $x$) and product yield (mol dm$^{-3}$, $Y$). The data follow a roughly quadratic pattern:



The plot above is produced by the following code:

```
library(readr)
chemistry <- read_csv("chemistry.csv")
plot(chemistry)
quadratic <- lm(y~x+I(x^2),data=chemistry)
beta <- coef(quadratic)
curve(beta[1]+beta[2]*x+beta[3]*x^2, from=0,to=1, add=T)
```

We now identify leverage and influential points. We have $r = p = 3$ and $n = 10$, so that an observation is of high leverage if $h_{ii} > 0.6$. The leverage can easily be computed in R as follows:

```
> influence(quadratic)$hat
        1         2         3         4         5
0.6181818 0.2787879 0.1833333 0.1954545 0.2242424
        6         7         8         9        10
0.2242424 0.1954545 0.1833333 0.2787879 0.6181818
```

We see that there are two high leverage points: observations 1 and 10. These are the observations with the smallest and largest values of $x$.

The calculation below shows that observations 1 and 10 are also influential:

```
>  cooks.distance(quadratic)
           1            2            3            4            5
1.1154690280 0.0509724360 0.0224314550 0.0006236501 0.1188788378
           6            7            8            9           10
0.0002512963 0.0003181888 0.2143038182 0.1014139679 2.5523436288
```

We can refit the model without the most influential observation (obs 10):

```
quadratic.subset <- lm(y~x+I(x^2),data=chemistry[-c(10),])
```

Recalculation of Cook's distance shows that observation 9 is now influential ($D_9 > 1$). Refitting again without observations 9 and 10 gives a fit with no influential observations as measured by Cook's distance. The parameter estimates change from $\hat{\boldsymbol{\theta}} = (0.8887, 4.1883, -4.9705)^{\mathrm{T}}$ (based on all the data) to $\hat{\boldsymbol{\theta}} = (0.8671, 4.4309, -5.3180)^{\mathrm{T}}$ (omitting observations 9 and 10).

# 4 The Gauss-Markov theorem

## 4.1 Estimable functions

**Definition 4.1.** *A function* $\psi : \mathbb{R}^p \to \mathbb{R}$ *is said to be a* linear parametric function *if*

$$\psi(\theta_1, \ldots, \theta_p) = \lambda_1\theta_1 + \ldots + \lambda_p\theta_p\,,$$

*also written as* $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ *, is a linear combination of the parameters for a fixed vector* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}} \in \mathbb{R}^p$*.*

**Definition 4.2.** *A linear parametric function* $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ *is* (linearly) estimable *if there exists a vector* $\mathbf{a} \in \mathbb{R}^n$ *(not depending on* $\boldsymbol{\theta}$*) such that:*

$$\text{for all } \boldsymbol{\theta} \in \mathbb{R}^p, \quad \mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{Y}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} \quad \text{for some vector of observations } \mathbf{Y}. \tag{4.1}$$

In other words, the parametric function is estimable if there exists an unbiased *linear* estimator (linear in $\mathbf{Y}$).

In the context of $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$ a linear model., we have the following result:

**Proposition 4.3.** $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ *is estimable if and only if* $\boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{a}$ *for some* $\mathbf{a} \in \mathbb{R}^n$*.*

That is, for $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ to be estimable, the coefficient vector $\boldsymbol{\lambda}$ must be a linear combination of the rows of $\mathbf{X}$.

*Proof of Proposition 4.3.* If the function $\psi(\theta) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable, then there exists $a \in \mathbb{R}^n$ such that for all $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$, we have $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{Y})$. That means

$$\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{Y}) \overset{\text{Lemma(2.4)}}{=} \mathbf{a}^{\mathrm{T}}\mathbb{E}(\mathbf{Y}) = \mathbf{a}^{\mathrm{T}}\mathbb{E}(\mathbf{X}\boldsymbol{\theta} + \epsilon) = \mathbf{a}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\theta} + \mathbb{E}(\epsilon)) \overset{\mathbb{E}(\epsilon)=0}{=} \mathbf{a}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}\,. \tag{4.2}$$

So $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{a}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}$ holds for all $\boldsymbol{\theta} \in \mathbb{R}^p$ , implying that $\boldsymbol{\lambda}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}\mathbf{X}$ or $\boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{a}$.

Conversely suppose that $\boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^n$. Then for all $\boldsymbol{\theta} \in \mathbb{R}^p$, and $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$, we have $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = (\mathbf{X}^{\mathrm{T}}\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{a}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{a}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\theta}) \overset{\mathbb{E}(\epsilon)=0}{=} \mathbf{a}^{\mathrm{T}}\mathbb{E}(\mathbf{X}\boldsymbol{\theta} + \epsilon) = \mathbf{a}^{\mathrm{T}}\mathbb{E}(\mathbf{Y}) \overset{\text{Lemma(2.4)}}{=} \mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{Y})$, showing that $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable. □

In the non-singular case, i.e. when $\det(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \neq 0$, then we do not have to worry about estimability, due to the following result.

**Corollary 4.4.** *If* $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ *is non-singular, then all functions of the form* $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ *are estimable.*

Note that this implies that in the non-singular case each individual component, $\theta_j$, of the vector $\boldsymbol{\theta}$ is estimable, since $\theta_j = \mathbf{e}_j^{\mathrm{T}}\boldsymbol{\theta}$, with

$$\mathbf{e}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j\text{th place.}$$

*Proof of Corollary 4.4.* If $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is non-singular, then $\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is the unique least squares estimator. Let $\boldsymbol{\lambda} \in \mathbb{R}^p$ be arbitrary. Then $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable, as $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}} = \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is an unbiased linear estimator of $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$. Unbiasedness is verified by noting that $\mathbb{E}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$. □

Let $\mathbf{G}$ be a $g$-inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and $\hat{\boldsymbol{\theta}}_{\mathbf{G}} = \mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ denote the corresponding solution of the normal equations.

**Proposition 4.5.** *If $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable, then:*

*(i)* $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G$ *is independent of the choice of $g$-inverse* $\mathbf{G}$*;*

*(ii)* $\mathbb{E}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$*;*

*(iii)* $\mathrm{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) = \sigma^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}$*;*

*(iv) if* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$*, then*

$$\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G \sim N(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda})\,.$$

Thus, in the rank-deficient / singular case, even though the least squares estimator of $\boldsymbol{\theta}$ is not unique, the least squares estimator of an estimable function $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is unique. Moreover, the least squares estimator of an estimable function is an unbiased estimator. The Gauss-Markov theorem below shows that in addition, the LS estimator is optimal in a sense to be defined.

Note that if $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ were not estimable, we would have that $\mathrm{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) = \sigma^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{G}^{\mathrm{T}}\boldsymbol{\lambda}$. The simplification in part (iii) only occurs for estimable functions.

*Proof of Proposition 4.5.* For part (i), note that

$$\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \quad (\text{as } \boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{a} \text{ by estimability})$$

$$= \mathbf{a}^{\mathrm{T}}\mathbf{H}\mathbf{Y}\,,$$

and recall from Lemma 3.1 that the hat matrix $\mathbf{H}$ does not depend on the choice of $g$-inverse, therefore $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G$ does not depend on the choice of $g$-inverse.

For part (ii) note that

$$\begin{aligned}\mathbb{E}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) &= \mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbb{E}(\mathbf{Y}) = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}\\ &= \mathbf{a}^{\mathrm{T}}\mathbf{H}\mathbf{X}\boldsymbol{\theta} = \mathbf{a}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} \quad \text{since } \mathbf{H}\mathbf{X} = \mathbf{X}\\ &= \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}\end{aligned}$$

For part (iii), note

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) &= \boldsymbol{\lambda}^{\mathrm{T}}\operatorname{Var}(\hat{\boldsymbol{\theta}}_G)\boldsymbol{\lambda} = \mathbf{a}^{\mathrm{T}}\mathbf{X}\operatorname{Var}(\hat{\boldsymbol{\theta}}_G)\mathbf{X}^{\mathrm{T}}\mathbf{a}\\ &= \mathbf{a}^{\mathrm{T}}\mathbf{X}\operatorname{Var}(\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y})\mathbf{X}^{\mathrm{T}}\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\operatorname{Var}(\mathbf{Y})\mathbf{X}\mathbf{G}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{a}\\ &= \mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}(\sigma^2\mathbf{I})\mathbf{X}\mathbf{G}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{a}\\ &= \sigma^2\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{G}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{a}\\ &= \sigma^2\mathbf{a}^{\mathrm{T}}\mathbf{H}^2\mathbf{a} = \sigma^2\mathbf{a}^{\mathrm{T}}\mathbf{H}\mathbf{a} \qquad \text{by symmetry and idempotence of } \mathbf{H}\\ &= \sigma^2\mathbf{a}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{a} = \sigma^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda} \quad \text{as } \boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{a}\text{, by estimability}\end{aligned}$$

□

## 4.2 Best linear unbiased estimator

**Definition 4.6.** $\mathbf{a}^{\mathrm{T}}\mathbf{Y}$ *is a* best linear unbiased estimator *(b.l.u.e.) of* $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ *if it is an unbiased estimator and it has smallest variance among all linear unbiased estimators.*

**Theorem 4.7** (Gauss-Markov theorem)**.** $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G$ *is a best linear unbiased estimator of* $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$*. Moreover it the unique linear estimator having this property.*

*Proof of Theorem 4.7.* Suppose that $\mathbf{b}^{\mathrm{T}}\mathbf{Y}$ is another unbiased linear estimator of $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$, so that $\boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{b}$ [2]. Recall from Proposition 3.1 that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent. We will show that $\operatorname{Var}(\mathbf{b}^{\mathrm{T}}\mathbf{Y}) \geq \operatorname{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G)$, so $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}_G$ is the minimum variance linear estimator.

[2] Indeed, this amounts to $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ being (linearly) estimable, thus $\boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{b}$ from the proof of proposition 4.3.

$$\begin{aligned}
&\mathrm{Var}(\mathbf{b}^{\mathrm{T}}\mathbf{Y}) - \mathrm{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) = \sigma^2\mathbf{b}^{\mathrm{T}}\mathbf{b} - \sigma^2\boldsymbol{\lambda}\mathbf{G}\boldsymbol{\lambda}^{\mathrm{T}} \quad \text{by the previous proposition} \\
&= \sigma^2\mathbf{b}^{\mathrm{T}}\mathbf{b} - \sigma^2\mathbf{b}^{\mathrm{T}}\mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{b} \quad \text{since } \boldsymbol{\lambda} = \mathbf{X}^{\mathrm{T}}\mathbf{b} \\
&= \sigma^2\mathbf{b}^{\mathrm{T}}(\mathbf{I} - \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}})\mathbf{b} \\
&= \sigma^2\mathbf{b}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{b} = \sigma^2\mathbf{b}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})\mathbf{b} \quad \text{as } \mathbf{I} - \mathbf{H} \text{ is idempotent} \\
&= \sigma^2\mathbf{b}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^{\mathrm{T}}\mathbf{b} \quad \text{as } \mathbf{I} - \mathbf{H} \text{ is symmetric} \\
&= \sigma^2[(\mathbf{I} - \mathbf{H})\mathbf{b}]^{\mathrm{T}}[(\mathbf{I} - \mathbf{H})\mathbf{b}] \\
&= \sigma^2||(\mathbf{I} - \mathbf{H})\mathbf{b}||^2 \geq 0
\end{aligned}$$

Hence $\mathrm{Var}(\mathbf{b}^{\mathrm{T}}\mathbf{Y}) - \mathrm{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) \geq 0$ so $\mathrm{Var}(\mathbf{b}^{\mathrm{T}}\mathbf{Y}) \geq \mathrm{Var}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G)$. □

# 5 Statistical inference under normality

## 5.1 Distributional results

In this section we state and prove a number of distributional results that are useful for constructing confidence intervals and hypothesis tests. Throughout, we will assume that the errors are normally distributed, i.e.

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \,.$$

We start with a simple lemma.

**Lemma 5.1.** *Suppose that* $\mathbf{U}$ *and* $\mathbf{V}$ *are independent random vectors, and* $g$ *and* $h$ *are functions. Then* $g(\mathbf{U})$ *and* $h(\mathbf{V})$ *are independent.*

*Proof.* Let $A$ and $B$ be 'suitably nice' subsets of Euclidean space. Then

$$\begin{aligned}
\mathbb{P}(g(\mathbf{U}) \in A \text{ and } h(\mathbf{V}) \in B) &= \mathbb{P}(\mathbf{U} \in g^{-1}(A) \text{ and } \mathbf{V} \in h^{-1}(B)) \\
&= \mathbb{P}(\mathbf{U} \in g^{-1}(A))\mathbb{P}(\mathbf{V} \in h^{-1}(B)) \qquad \text{by independence of } \mathbf{U} \text{ and } \mathbf{V} \\
&= \mathbb{P}[g(\mathbf{U}) \in A]\mathbb{P}[h(\mathbf{V}) \in B]
\end{aligned}$$

where $g^{-1}(A) = \{\mathbf{u} | g(\mathbf{u}) \in A\}$ and $h^{-1}(B) = \{\mathbf{v} | h(\mathbf{v}) \in B\}$. □

The following result shows that a suitably scaled version of $\hat{\sigma}^2$ has a chi squared distribution.

**Proposition 5.2.** *If* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$*, and* $r = \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$*, then*

*(i)*

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-r) \,,$$

*(ii)* $\hat{\sigma}^2$ *is independent of* $\hat{\boldsymbol{\theta}}_G$*.*

[Recall that $\chi^2(k)$ is the distribution of $Y = Z_1^2 + \ldots + Z_k^2$ where $Z_1, \ldots, Z_k \overset{\text{i.i.d.}}{\sim} N(0,1)$.]

*Proof.* (i) Note that

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \frac{1}{\sigma^2}\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$$

As $\mathbf{I} - \mathbf{H}$ is a real symmetric matrix, it can be diagonalised, giving $\mathbf{I} - \mathbf{H} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$ where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\mathbf{P}$ is a matrix of orthogonal eigenvectors, so that $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I}$. Hence the above becomes

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \lambda_i [\mathbf{P}^{\mathrm{T}}\boldsymbol{\epsilon}]_i^2 = \sum_{i=1}^{n} \lambda_i Z_i^2 \qquad (\dagger)$$

where $Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$, since $\mathbf{P}^{\mathrm{T}}\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{P}^{\mathrm{T}}\mathbf{P}) = N(\mathbf{0}, \sigma^2\mathbf{I})$.

Now we consider the eigenvalues. As $\mathbf{I}-\mathbf{H}$ is idempotent, its eigenvalues must all be equal to 0 or 1. Moreover, since $n - r = \mathrm{tr}(\mathbf{I}-\mathbf{H}) = \mathrm{tr}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{P}) = \mathrm{tr}(\boldsymbol{\Lambda}\mathbf{P}\mathbf{P}^{\mathrm{T}}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_i \lambda_i$, we must have that $n-r$ of the $\lambda_i$ are equal to 1 and $r$ of the $\lambda_i$ are equal to 0. Substituting this into equation (†) above, we see that $\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-r)$ as claimed.

(ii) Note that $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}) = \mathrm{Cov}[\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{Y}, (\mathbf{I}-\mathbf{H})\mathbf{Y}] = \mathbf{G}\mathbf{X}^{\mathrm{T}}\,\mathrm{Cov}(\mathbf{Y},\mathbf{Y})(\mathbf{I}-\mathbf{H})^{\mathrm{T}} = \sigma^2\mathbf{G}[\mathbf{X} - \mathbf{H}\mathbf{X}]^{\mathrm{T}} = \mathbf{0}$, so $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\epsilon}}$ are uncorrelated. Moreover,

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\epsilon}}\end{bmatrix} = \begin{bmatrix}\mathbf{G}\mathbf{X}^{\mathrm{T}} \\ \mathbf{I}-\mathbf{H}\end{bmatrix}\mathbf{Y}$$

is a linear transformation of a multivariate normal vector, so is itself normal, i.e. $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\epsilon}}$ are jointly normal. Hence they are independent. As $\hat{\sigma}^2 = \frac{1}{n-r}||\hat{\boldsymbol{\epsilon}}||^2$ is a function of $\hat{\boldsymbol{\epsilon}}$, we have that $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ are independent.

□

The next result is the main one used when constructing confidence intervals and hypothesis tests. Recall that for an estimable function $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G$ the least squares estimator of the estimable function $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ satisfies

$$\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G \sim N(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda})\,, \tag{5.1}$$

and its standard error is $\mathrm{s.e.}(\lambda^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G) = \sqrt{\hat{\sigma}^2\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}$.

**Proposition 5.3.** *If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is an estimable function, then*

$$\frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\textit{s.e.}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G)} \sim t(n-r)\,.$$

[Recall that $t(k)$ is the distribution of $T = \frac{Z}{\sqrt{V/k}} \sim t(k)$, where $Z \sim N(0,1)$, and $V \sim \chi^2(k)$ independently.]

*Proof of Proposition 5.3.* Recall that if $Y \sim N(\mu, \sigma^2)$, then $\frac{Y-\mu}{\sigma} \sim N(0,1)$; this is known as standardization. Considering (5.1) and standardizing $\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G$, we see that

$$Z = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\sigma\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} \sim N(0,1)\,,$$

Recall that $V = (n-r)\hat{\sigma}^2/\sigma \sim \chi^2(n-r)$ and $\hat{\boldsymbol{\theta}}_G$ are independent from Proposition 5.2(ii). Since $Z$ is a function of $\hat{\boldsymbol{\theta}}_G$, we also have by Lemma 5.1 that $V$ and $Z$ are independent. Hence

$$T = \frac{Z}{\sqrt{V/(n-r)}} \sim t(n-r)\,.$$

However, the left hand side satisfies

$$\frac{Z}{\sqrt{V/(n-r)}} = \frac{\frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\sigma\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}}}{\sqrt{\frac{(n-r)\hat{\sigma}^2}{(n-r)\sigma^2}}} = \frac{\sqrt{\sigma^2}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta})}{\sigma\sqrt{\hat{\sigma}^2}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\mathrm{s.e.}(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G)}$$

Hence we have proved Proposition 5.3. □

## 5.2 Hypothesis tests

### 5.2.1 Testing estimable functions

Suppose that $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is an estimable function. We wish to test the null hypothesis,

$$H_0 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \psi\,,$$

where $\psi$ is a known real number, in favour of the two-sided alternative hypothesis,

$$H_1 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} \neq \psi\,.$$

Let $t_{\alpha;k}$ denote the *upper $\alpha$ point* of the $t$ distribution with $k$ degrees of freedom. That is, if $W \sim t(k)$, then $\mathbb{P}(W > t_{\alpha;k}) = \alpha$.

Recall that the *significance level* of a hypothesis test is

$$\mathbb{P}(\text{Type I error} \,|\, H_0) = \mathbb{P}(\text{Reject } H_0 \,|\, H_0)\,.$$

**Proposition 5.4.** *If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, and $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is an estimable function, then the test statistic*

$$T = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \psi}{\hat{S}_{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G}} = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \psi}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}}\,, \tag{5.2}$$

*with rejection region*

$$\textit{Reject } H_0 \textit{ if } |T| > t_{\frac{\alpha}{2};n-r}\,, \tag{5.3}$$

*gives a test of $H_0$ vs $H_1$ with significance level $\alpha$.*

*Proof.* First note that if $H_0$ is true then $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \psi$ and so $T \sim t(n-r)$ by Proposition 5.3. Hence, we see that

$$\begin{aligned}
\mathbb{P}(\text{Reject } H_0 \,|\, H_0) &= \mathbb{P}\left(|T| > t_{\frac{\alpha}{2};n-r} \,\middle|\, H_0\right) = \mathbb{P}\left(T > t_{\frac{\alpha}{2};n-r} \text{ or } T < -t_{\frac{\alpha}{2};n-r}\right) \\
&= \mathbb{P}\left(T > t_{\frac{\alpha}{2};n-r} \,\middle|\, H_0\right) + \mathbb{P}\left(T < -t_{\frac{\alpha}{2};n-r} \,\middle|\, H_0\right) \\
&\quad \text{since if } A \text{ and } B \text{ are disjoint events, then } \mathbb{P}(A \text{ or } B) = P(A) + P(B) \\
&= \mathbb{P}\left(T > t_{\frac{\alpha}{2};n-r} \,\middle|\, H_0\right) + \mathbb{P}\left(T > t_{\frac{\alpha}{2};n-r} \,\middle|\, H_0\right) \\
&\quad \text{by symmetry of the } t \text{ distribution around zero} \\
&= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha\,.
\end{aligned}$$

□

Note that **one-sided alternatives** are also easily tested, by amending the form of the rejection region:

- For $H_1 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} > \psi$, reject $H_0$ if $T > t_{\alpha;n-r}$;
- For $H_1 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} < \psi$, reject $H_0$ if $T < -t_{\alpha;n-r}$.

**Statistical tables in the exam**

Note that the statistical tables in the exam use a slightly different notation to the lecture notes for the points of a $t$ distribution.

In particular, in the examination tables, $t_{\nu,q}$ denotes the $q$-quantile of the $t(\nu)$ distribution, i.e. if $T \sim t(\nu)$ then

$$\mathbb{P}(T \leq t^{\text{tables}}_{\nu,q}) = q\,.$$

In other words, the tables give the *lower* $q$ point of the distribution whereas in the notes we use the *upper* $\alpha$ point.

The two are related as follows:

$$t^{\text{notes}}_{\alpha;k} = t^{\text{tables}}_{k,1-\alpha}\,,$$

since

$$\mathbb{P}(T \leq t^{\text{notes}}_{\alpha;k}) = 1 - \mathbb{P}(T > t^{\text{notes}}_{\alpha;k}) = 1 - \alpha = \mathbb{P}(T \leq t^{\text{tables}}_{k,1-\alpha})\,.$$

The upshot of this is that to calculate $t^{\text{notes}}_{\alpha;k}$, you first need to calculate $q = 1 - \alpha$, and then look up $t^{\text{tables}}_{k,q}$ in the tables.

**Example 5.5.** *Calculate the upper 0.05 point of a $t(7)$ distribution, i.e. $t_{0.05;7}$. Solution: first we calculate $q = 1 - \alpha = 1 - 0.05 = 0.95$. Then we look up $t^{\textit{tables}}_{7,0.95} = 1.8946$ to give $t_{0.05;7} = 1.8946$.*

### 5.2.2 Testing individual parameters

Recall that if the information matrix is non-singular, i.e. if $\det(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \neq 0$, then by Proposition 4.4, each component $\theta_j = \mathbf{e}_j^{\mathrm{T}}\boldsymbol{\theta}$ is estimable. Moreover,

$$\mathrm{Var}(\mathbf{e}_j^{\mathrm{T}}\hat{\boldsymbol{\theta}}) = \sigma^2 \mathbf{e}_j^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{e}_j = \sigma^2 m^{(jj)}\,,$$

where $m^{(jj)}$ denotes the $j$th diagonal element of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

Recall also that in this case $r = \mathrm{rank}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = p$. Thus in the non-singular case we can test the null hypothesis

$$H_0 : \theta_j = \psi \quad \text{vs.} \quad H_1 : \theta_j \neq \psi$$

at significance level $\alpha$ using the test statistic

$$T = \frac{\hat{\theta}_j - \psi}{\hat{\sigma}\sqrt{m^{(jj)}}}\,, \tag{5.4}$$

and rejection region

$$\text{Reject } H_0 \text{ if } |T| > t_{\frac{\alpha}{2};n-p}\,.$$

A case of particular importance occurs when the null value $\psi$ is equal to zero, i.e.

$$H_0 : \theta_j = 0\,.$$

This is often simply known as a 'significance test for the parameter $\theta_j$'. When $H_0$ is rejected it is often said that '$\theta_j$ is significant'.

Strictly speaking, and especially when talking to applied scientists, it is better to say that rejection of $H_0 : \theta_j = 0$ implies that the data contain '*statistically* significant evidence that the parameter is non-zero'. (Often knowing that $\theta_j \neq 0$ is not enough to show that it is *scientifically* significant. We may need evidence of a *large* effect, rather than just a non-zero one.)

### 5.2.3 Example: soya bean data

A new strain of soya has been developed for use in soil with a low pH value (acidic soil). The following data were collected in an experiment in order to determine the pH at which the expected yield of the plant is maximized.

| $i$ | pH, $x_i$ | Yield, $Y_i$ |
|---|---|---|
| 1 | 2 | 9.1 |
| 2 | 2 | 9.7 |
| 3 | 2 | 10.8 |
| 4 | 3 | 12.0 |
| 5 | 3 | 12.7 |
| 6 | 3 | 13.6 |
| 7 | 4 | 14.6 |
| 8 | 4 | 15.7 |
| 9 | 4 | 15.9 |
| 10 | 5 | 15.2 |
| 11 | 5 | 14.8 |
| 12 | 5 | 14.5 |
| 13 | 6 | 12.6 |
| 14 | 6 | 11.8 |
| 15 | 6 | 11.5 |

Table: Soya bean data

A quadratic model is proposed, i.e.

$$Y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \epsilon_i\,, \quad i = 1, \ldots, 15,$$

with the $\epsilon_i \sim N(0, \sigma^2)$ independently.

We are given that

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = 2586.83\,, \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} \frac{79}{15} & -\frac{14}{5} & \frac{1}{3} \\ -\frac{14}{5} & \frac{109}{70} & -\frac{4}{21} \\ \frac{1}{3} & -\frac{4}{21} & \frac{1}{42} \end{bmatrix}, \quad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} 194.5 \\ 796.8 \\ 3607.2 \end{pmatrix}.$$

(i) Using a significance level of $\alpha = 0.05$, test $H_0 : \theta_3 = 0$ vs $H_1 : \theta_3 \neq 0$.

(ii) Using a significance level of $\alpha = 0.05$, test the null hypothesis that the maximum expected yield occurs at pH level $x = 4$.

For part (i), by (5.4) the appropriate test statistic is

$$t = \frac{\hat{\theta}_3}{\hat{\sigma}\sqrt{m^{(33)}}}\,.$$

To calculate this, we first need to compute $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}$. The least squares estimates are

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} \frac{79}{15} & -\frac{14}{5} & \frac{1}{3} \\ -\frac{14}{5} & \frac{109}{70} & -\frac{4}{21} \\ \frac{1}{3} & -\frac{4}{21} & \frac{1}{42} \end{pmatrix} \begin{pmatrix} 194.5 \\ 796.8 \\ 3607.2 \end{pmatrix}$$
$$= \begin{pmatrix} -4.273333 \\ 9.045714 \\ -1.052381 \end{pmatrix}.$$

The residual sum of squares is

$$\begin{aligned} \mathsf{SSE} &= \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \hat{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= 2586.83 - (-4.273333 \times 194.5 + 9.045714 \times 796.8 - 1.052381 \times 3607.2) \\ &= 6.517097\,. \end{aligned}$$

Hence $\hat{\sigma}^2 = \mathsf{SSE}/(n-p) = 6.517097/(15-3) = 0.5430914 = 0.7369^2$. Thus the test statistic is

$$t = \frac{-1.052381}{0.7369\sqrt{\frac{1}{42}}} = \frac{-1.052381\sqrt{42}}{0.7369} = -9.25527\,.$$

The critical value is

$$t_{\frac{0.05}{2};15-3} = t_{0.025;12} = t^{\mathsf{tables}}_{12,0.975} = 2.1788\,.$$

Since $|t| = 9.25527 > 2.1788$, we reject $H_0 : \beta_3 = 0$, and conclude there is statistically significant evidence that the regression line should include a quadratic term.

For part (ii), note that the regression model is

$$\mathbb{E}(Y) = \theta_1 + \theta_2 x + \theta_3 x^2\,,$$

and that the maximum expected response occurs where

$$0 = \frac{d\,\mathbb{E}(Y)}{dx} = \theta_2 + 2\theta_3 x\,.$$

Hence the hypothesis that the maximum expected response occurs at $x = 4$ is equivalent to

$$H_0 : \theta_2 + 8\theta_3 = 0\,,$$

or equivalently

$$H_0 : \begin{pmatrix} 0 & 1 & 8 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = 0\,,$$

which is of the form $H_0 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = 0$ with $\boldsymbol{\lambda} = (0, 1, 8)^{\mathrm{T}}$.

Since $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, by Corollary 4.4 any linear combination of the parameters is estimable, including $\theta_2 + 8\theta_3$. The unique $g$-inverse of the information matrix is $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Using Proposition 5.4, the test statistic is

$$t = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\lambda}}} = \frac{\hat{\theta}_2 + 8\hat{\theta}_3}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\lambda}}}$$

Note that

$$\begin{aligned}
\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\boldsymbol{\lambda} &= \begin{pmatrix} 0 & 1 & 8 \end{pmatrix} \begin{pmatrix} \frac{79}{15} & -\frac{14}{5} & \frac{1}{3} \\ -\frac{14}{5} & \frac{109}{70} & -\frac{4}{21} \\ \frac{1}{3} & -\frac{4}{21} & \frac{1}{42} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 8 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 1 & 8 \end{pmatrix} \begin{pmatrix} -0.1333333 \\ 0.03333333 \\ 0 \end{pmatrix} \\
&= 0.03333333\,.
\end{aligned}$$

Hence the test statistic becomes

$$t = \frac{9.045714 + 8 \times (-1.052381)}{0.7369\sqrt{0.03333333}} = 4.657879.$$

Again this is compared to the critical value $t_{0.025,12} = 2.1788$. Since $|t| = 4.657879 > 2.1788$ we reject the null hypothesis that the maximum expected response is located at $x = 4$. Note that in section 1.2.2 we concluded that the maximum expected response was located at $\hat{x}^* = 4.299$ instead.

## 5.3 Confidence intervals

Recall that for an unknown fixed quantity $\eta$, the random interval

$$I(\mathbf{Y}) = [a(\mathbf{Y}), b(\mathbf{Y})] \,,$$

whose end-points are functions of the random data, is a $100(1-\alpha)\%$ *confidence interval* for $\eta$ if the *coverage probability*,

$$\mathbb{P}[a(\mathbf{Y}) \leq \eta \leq b(\mathbf{Y})] = 1 - \alpha \,.$$

(E.g. if $\alpha = 0.05$, we obtain a $100(1-0.05) = 100(0.95) = 95\%$ confidence interval.)

A confidence interval can be thought of as a set of plausible values for $\eta$.

**Proposition 5.6.** *If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is an estimable function, then the following is a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$,*

$$I(\mathbf{Y}) = \left[\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}},\ \boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G + t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}\right] .$$

Note that the end-points of the above interval are of the form

$$\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G \pm t_{\frac{\alpha}{2};n-r}\,\hat{S}_{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G} \,.$$

Recall that in the full rank (non-singular) case, we have that each individual parameter $\theta_j$ is estimable, $r = p$, and $\mathbf{G} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

Thus the formula above can be applied with $\boldsymbol{\lambda} = \mathbf{e}_j$ to give the following confidence interval for $\theta_j$:

$$\left[\hat{\theta}_j - t_{\frac{\alpha}{2};n-p}\,\hat{\sigma}\sqrt{m^{(jj)}},\ \hat{\theta}_j + t_{\frac{\alpha}{2};n-p}\,\hat{\sigma}\sqrt{m^{(jj)}}\right] , \tag{5.5}$$

where again $m^{(jj)}$ denotes the $j$th diagonal element of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

**Example 5.7.** *For the Soya bean example (Section 5.2.3), compute a 95% confidence interval for the coefficient, $\theta_3$, of the $x^2$ term.*

Recall that

$$\hat{\sigma} = 0.7369\,, \quad \hat{\theta}_3 = -1.052381\,, \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{pmatrix} \frac{79}{15} & -\frac{14}{5} & \frac{1}{3} \\ -\frac{14}{5} & \frac{109}{70} & -\frac{4}{21} \\ \frac{1}{3} & -\frac{4}{21} & \frac{1}{42} \end{pmatrix}.$$

To obtain a 95% interval, set $1-\alpha = 0.95$, i.e. $\alpha = 0.05$ and note that $t_{\frac{\alpha}{2};n-p} = t_{0.025;12} = 2.1788$. Hence, by (5.5), we have that a 95% confidence interval is given by the end points

$$-1.052381 \pm 2.1788 \times 0.7369 \times \sqrt{\frac{1}{42}}\,,$$

i.e. the 95% confidence interval is $[-1.3001, -0.8046]$. Since this interval does not contain zero, it is not plausible that $\theta_3 = 0$, which is why the null hypothesis $H_0 : \theta_3 = 0$ was rejected.

*Proof of Proposition 5.6.* From Proposition 5.3, we know that if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable, then

$$T = \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} \sim t_{n-r}\,,$$

and so

$$\begin{aligned}
&\mathbb{P}(-t_{\frac{\alpha}{2};n-r} < T < t_{\frac{\alpha}{2};n-r}) = \mathbb{P}(T < t_{\frac{\alpha}{2};n-r}) - \mathbb{P}(T < -t_{\frac{\alpha}{2};n-r}) \\
&= [1 - \mathbb{P}(T > t_{\frac{\alpha}{2};n-r})] - \mathbb{P}(T > t_{\frac{\alpha}{2};n-r}) \\
&\qquad \text{by symmetry of the } t \text{ distribution around zero} \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha\,.
\end{aligned}$$

Hence

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(-t_{\frac{\alpha}{2};n-r} < \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} < t_{\frac{\alpha}{2};n-r}\right) \\
&= \mathbb{P}\left(-t_{\frac{\alpha}{2};n-r} < \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} \quad \text{and} \quad \frac{\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}}{\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}} < t_{\frac{\alpha}{2};n-r}\right) \\
&= \mathbb{P}\Big(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} < \boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G + t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}} \\
&\qquad\qquad \text{and} \quad \boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}} < \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}\Big) \\
1-\alpha &= \mathbb{P}\left(\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}} < \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} < \boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G + t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}\right).
\end{aligned}$$

Note that the final line above says precisely that the random interval

$$I(\mathbf{Y}) = \left[\boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}},\ \boldsymbol{\lambda}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G + t_{\frac{\alpha}{2};n-r}\,\hat{\sigma}\sqrt{\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{G}\boldsymbol{\lambda}}\right]$$

contains $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta}$ with probability $1-\alpha$, and so $I(\mathbf{Y})$ is a $100(1-\alpha)\%$ confidence interval as claimed. □

## 5.4 Prediction intervals

The interval $[a(\mathbf{Y}), b(\mathbf{Y})]$ is a $100(1-\alpha)\%$ *prediction interval* for a random variable $W$ if $P(a(\mathbf{Y}) \leq W \leq b(\mathbf{Y})) = 1-\alpha$. This is similar to a confidence interval, but for a random quantity rather than a fixed quantity.

Suppose that we have fitted a model $Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta} + \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ to response data $Y_1, \ldots, Y_n$, and that we now wish to predict a future response,

$$\tilde{Y} = \tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\theta} + \tilde{\epsilon}\,,$$

from the same model, where $\tilde{\epsilon} \sim N(0, \sigma^2)$ independently of $\epsilon_1, \ldots, \epsilon_n$.

**Proposition 5.8.** *Provided that $\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\theta}$ is estimable, the following is a $100(1-\alpha)\%$ prediction interval for $\tilde{Y}$:*

$$\left[\tilde{\mathbf{x}}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G - t_{\frac{\alpha}{2};n-r}\hat{\sigma}\sqrt{\tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{G}\tilde{\mathbf{x}}+1},\quad \tilde{\mathbf{x}}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_G + t_{\frac{\alpha}{2};n-r}\hat{\sigma}\sqrt{\tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{G}\tilde{\mathbf{x}}+1}\right]\,.$$

Compared to the confidence interval for $\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\theta}$ in Proposition 5.6, there is an extra $+1$ inside the square root. This reflects the fact that when we predict $\tilde{Y}$, there is additional uncertainty due to not knowing $\tilde{\epsilon}$, as well as the uncertainty due to not knowing $\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\theta}$. For guidance on the proof of this result, see Example Sheet 5, Q3.

# 6 Testing a general linear hypothesis

## 6.1 Linear hypotheses and reduced models

Suppose that the following general linear model holds

$$F: \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})\,.$$

In Chapter 5, we showed how to test a hypothesis about a single estimable linear function of the parameters, i.e. $H_0 : \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\theta} = \psi$. In this chapter, we derive tests for hypotheses consisting of $k$ independent linear constraints on the parameters.

We suppose that the null and alternative hypotheses are of the form

$$H_0 : \begin{cases} \mathbf{c}_1^{\mathrm{T}}\boldsymbol{\theta} = d_1\,, \\ \mathbf{c}_2^{\mathrm{T}}\boldsymbol{\theta} = d_2\,, \\ \quad\vdots \\ \mathbf{c}_k^{\mathrm{T}}\boldsymbol{\theta} = d_k\,, \end{cases} \qquad H_1 : \mathbf{c}_l^{\mathrm{T}}\boldsymbol{\theta} \neq d_l \quad \text{for some } l = 1, 2, \ldots, k\,.$$

where $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathbb{R}^p$ are known linearly independent vectors, and $d_1, \ldots, d_k$ are known constants. The hypothesis $H_0$ above is known as the *general linear hypothesis*. The above can be rewritten as

$$H_0 : \mathbf{C}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}^{\mathrm{T}}\boldsymbol{\theta} \neq \mathbf{d}\,,$$

where $\mathbf{C}$ is a $p \times k$ matrix with $\mathrm{col}_j(\mathbf{C}) = \mathbf{c}_j$ and $\mathbf{d} = (d_1, \ldots, d_k)^{\mathrm{T}}$. It is assumed that there exists at least one $\mathbf{b} \in \mathbb{R}^p$ such that $\mathbf{C}^{\mathrm{T}}\mathbf{b} = \mathbf{d}$, i.e. that the constraints can actually be satisfied for certain possible parameter values.

It can be shown that for any general linear hypothesis, there exists a corresponding *reduced model*,

$$R: \quad \tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})\,,$$

such that $H_0$ holds if and only if model $R$ is true. The response, $\tilde{Y}_i$, in the reduced model may differ from $Y_i$ by a constant, as we will see.

### 6.1.1 Example: one-way classification model

Consider the one-way classification model with $t$ treatment groups, each containing $m$ experimental units. The response for the $j$th unit in the $i$th group is

$$Y_{ij} = \mu_i + \epsilon_{ij}\,,$$

$$\epsilon_{ij} \sim N(0, \sigma^2)\,, \quad i = 1, \ldots, t;\ j = 1, \ldots, m\,.$$

Above, $\mu_i$ denotes the expected response under the $i$th treatment. A common null hypothesis is that all of the treatments give the same expected response, i.e. that

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_t \,.$$

This null hypothesis corresponds to the statement that there is no difference in the effects of the different treatments.

Note that $H_0$ can be rewritten equivalently as

$$H_0 : \begin{cases} \mu_1 - \mu_2 = 0 \,, \\ \mu_1 - \mu_3 = 0 \,, \\ \vdots \\ \mu_1 - \mu_t = 0 \,. \end{cases}$$

$$H_0 : \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & -1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_t \end{pmatrix} = \mathbf{0} \,,$$

thus $H_0$ is indeed a linear hypothesis, with $t - 1$ constraints.

If $H_0$ is true then $\mu_1 = \mu_2 = \ldots = \mu_t = \mu$, and the full model becomes the reduced model

$$Y_{ij} = \mu + \epsilon_{ij} \,,$$

with a single free parameter $\mu$.

### 6.1.2 Example: quadratic model

In the quadratic model

$$F : \quad Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i \,, \quad i = 1, \ldots, n \,,$$

assume that $\gamma < 0$, so that the turning point is a maximum. Consider the null hypothesis that BOTH (i) the maximum expected response occurs at $x = 4$, and (ii) the maximum expected response is 16. The maximum expected response occurs where

$$0 = \frac{d\mathbb{E}(Y)}{dx} = \beta + 2\gamma x \,,$$

thus the maximum occurs at $x = 4$ if and only if

$$\beta + 8\gamma = 0\,.$$

The value of the expected response at $x = 4$ is 16 if and only if

$$16 = \alpha + 4\beta + 16\gamma\,.$$

Thus our null hypothesis is equivalent to

$$H_0 : \begin{cases} \beta + 8\gamma = 0 \\ \alpha + 4\beta + 16\gamma = 16\,, \end{cases}$$

which is clearly a linear hypothesis with $k = 2$ constraints. Note that in addition $H_0$ is true if any only if

$$\begin{aligned} \beta &= -8\gamma \\ \alpha &= 16 + 16\gamma\,. \end{aligned}$$

Hence, under $H_0$ the parameters $\alpha$ and $\beta$ can be eliminated from the full model, as follows:

$$\begin{aligned} &\alpha + \beta x + \gamma x^2 = 16 + 16\gamma - 8\gamma x + \gamma x^2 \\ &= 16 + \gamma(16 - 8x + x^2) = 16 + \gamma(x-4)^2 \\ &Y_i = 16 + \gamma(x_i - 4)^2 + \epsilon_i\,. \end{aligned}$$

Thus, setting $\tilde{Y}_i = Y_i - 16$, the reduced model is

$$R: \quad \tilde{Y}_i = \gamma(x_i - 4)^2 + \epsilon_i\,.$$

The reduced model has a single free parameter, $\gamma$.

### 6.1.3 The general case

We now show that a general linear hypothesis always corresponds to a reduced model. To see this we consider the general form of the solutions of the equation from the null hypothesis, $\mathbf{C}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{d}$. Specifically, $\mathbf{t}$ is a solution of this equation if and only if

$$\mathbf{t} = \mathbf{A}\boldsymbol{\beta} + \mathbf{b}\,, \qquad \text{for some } \boldsymbol{\beta} \in \mathbb{R}^{r_R}$$

where $\mathbf{b}$ is a particular solution, and $\mathbf{A}$ is a matrix whose columns form a basis of the null space (or *kernel*) of $\mathbf{C}^{\mathrm{T}}$. Note that $\mathbf{C}^{\mathrm{T}}\mathbf{A} = \mathbf{0}$.

Hence

$$\begin{aligned} H_0 : \mathbf{C}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{d} \text{ holds} &\iff \boldsymbol{\theta} = \mathbf{A}\boldsymbol{\beta} + \mathbf{b} \\ &\iff \mathbf{Y} = \mathbf{X}(\mathbf{A}\boldsymbol{\beta} + \mathbf{b}) + \boldsymbol{\epsilon} \\ &\iff \mathbf{Y} = \mathbf{X}\mathbf{A}\boldsymbol{\beta} + \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon} \\ &\qquad \text{(NB this is not quite a linear model due to the constant term)} \\ &\iff \mathbf{Y} - \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &\iff \tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

with $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$, $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{A}$. This notation will be used in later proofs.

## 6.2 The test statistic

The statement that the null hypothesis $H_0$ is true is equivalent to the statement that the reduced model $R$ is correct. Hence testing $H_0$ is equivalent to testing the null hypothesis that the reduced model $R$ is correct, versus the alternative hypothesis that the full model $F$ is required to explain the data.

Let

$$\begin{aligned} \mathsf{SSE}_F &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_G)^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_G) \\ \mathsf{SSE}_R &= (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{G_R})^{\mathrm{T}}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{G_R}) \end{aligned}$$

denote the residual sums of squares from fitting the full model and the reduced model respectively. Above, $\hat{\boldsymbol{\beta}}_{G_R}$ denotes a least squares estimate of $\boldsymbol{\beta}$ obtained using a $g$-inverse, $\mathbf{G}_R$, of $\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$. It can be shown that $\mathsf{SSE}_R$ does not depend on the choice of $\mathbf{A}$ and $\mathbf{b}$.

Note that if either $H_0$ or $H_1$ is true, then

$$\mathbb{E}\left(\frac{\mathsf{SSE}_F}{n - r_F}\right) = \sigma^2\,,$$

by Proposition 3.4 (since both $H_0$ and $H_1$ imply that $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$; $H_0$ just restricts the values of the parameters).

In addition, if $H_0$ is true then, also by Proposition 3.4,

$$\mathbb{E}\left(\frac{\mathsf{SSE}_R}{n - r_R}\right) = \sigma^2\,.$$

However, if $H_0$ is false, then $\text{SSE}_R/(n - r_R)$ will usually be inflated due to the lack of fit of the reduced model. This suggests that a reasonable procedure for testing $H_0$ versus $H_1$ is to reject $H_0$ if $\text{SSE}_R/(n - r_R)$ is significantly larger than $\text{SSE}_F/(n - r_F)$.

A test statistic based on this idea is given below, together with its distribution when the null hypothesis is true.

**Definition 6.1.** *If $V_1 \sim \chi^2(d_1)$ and $V_2 \sim \chi^2(d_2)$ independently, we say that the random variable*

$$W = \frac{V_1/d_1}{V_2/d_2} \sim F(d_1, d_2)\,,$$

*i.e. $W$ has an $F$* distribution with $d_1$ numerator degrees of freedom and $d_2$ denominator degrees of freedom.

**Proposition 6.2.** *If $H_0$ is true, and $d = r_F - r_R$ is the difference in the ranks of the full and reduced models, then the test statistic*

$$F = \frac{(\mathit{SSE}_R - \mathit{SSE}_F)/d}{\mathit{SSE}_F/(n - r_F)} \sim F(d, n - r_F)\,.$$

(N.B. the quantity $\text{SS}_{H_0} = \text{SSE}_R - \text{SSE}_F$ in the numerator is known as the *extra sum of squares*).

This leads to the following test of $H_0$ versus $H_1$ at significance level $\alpha$:

$$\text{Reject } H_0 \text{ if } \quad \frac{(\text{SSE}_R - \text{SSE}_F)/d}{\text{SSE}_F/(n - r_F)} > F_{\alpha;d,n-r_F}\,, \tag{6.1}$$

where $F_{\alpha;d,n-r_F}$ denotes the upper $\alpha$ point of an $F(d, n - r_F)$ distribution. This critical value can be looked up in all good statistical tables.

We will prove the result above later, but first we give some examples.

### 6.2.1 Example: quadratic model, soya bean data

For the soya bean data in Section 5.2.3, and the quadratic model

$$F: \quad Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i\,, \quad \epsilon_i \sim N(0, \sigma^2)$$

$i = 1, \ldots, 15$, we wish to test the null hypothesis that both

(i) the maximum expected response occurs at $x = 4$, **and**

(ii) the maximum expected response equals 16.

We showed in Section 6.1.2 that this null hypothesis is equivalent to

$$H_0 : \begin{cases} \beta + 8\gamma = 0 \\ \alpha + 4\beta + 16\gamma = 16\,, \end{cases}$$

and that the corresponding reduced model is

$$R: \quad \tilde{Y}_i = \gamma(x_i - 4)^2 + \epsilon_i\,, \quad \epsilon_i \sim N(0, \sigma^2)\,,$$

with $\tilde{Y}_i = Y_i - 16$. Note that this model does not have an intercept.

The above can be written in matrix form as $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with

$$\tilde{\mathbf{Y}} = \begin{pmatrix} Y_1 - 16 \\ Y_2 - 16 \\ \vdots \\ Y_{15} - 16 \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} (x_1 - 4)^2 \\ (x_2 - 4)^2 \\ \vdots \\ (x_{15} - 4)^2 \end{pmatrix}, \quad \boldsymbol{\beta} = (\gamma)\,,$$

Moreover we are given that

$$\begin{aligned} \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} &= \left(\sum_{i=1}^{15}(x_i - 4)^4\right) = 102 \\ \tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{Y}} &= \left(\sum_{i=1}^{15}(x_i - 4)^2(Y_i - 16)\right) = -135.2 \\ \tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{Y}} &= \sum_{i=1}^{15}(Y_i - 16)^2 = 202.83 \end{aligned}$$

Thus the least squares estimates of the reduced model parameters are

$$\begin{aligned} (\hat{\gamma}) = \hat{\boldsymbol{\beta}} &= (\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{Y}} \\ &= (102)^{-1}(-135.2) = -1.32549 \end{aligned}$$

and the residual sum of squares for the reduced model is

$$\mathsf{SSE}_R = \tilde{\mathbf{Y}}^{\mathrm{T}}\tilde{\mathbf{Y}} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{Y}} = 202.83 - (-1.32549)(-135.2) = 23.62375\,.$$

Recall from Section 5.2.3 that $\mathsf{SSE}_F = 6.517097$. Hence the test statistic (6.1) is

$$F = \frac{(23.62375 - 6.517097)/(2)}{(6.517097)/(15 - 3)} = 15.74933\,.$$

Testing $H_0$ at the 5% significance level, we have that $15.749 = F > F_{0.05;2,12} = 3.89$, and so $H_0$ is rejected. Thus, there is statistically significant evidence at the 5% level that either the maximum expected response is not at $x = 4$, or the maximum value of the expected response is not 16.

**Calculations using R**

The calculations can be done in R easily as follows.

```
> soyabean <- data.frame(x=rep(c(2,3,4,5,6),c(3,3,3,3,3)),
+                        y=c(9.1,9.7,10.8,12,12.7,13.6,14.6,
+                            15.7,15.9,15.2,14.8,14.5,12.6,11.8,11.5))
> fitF <- lm(y ~ 1 + x + I(x^2), data=soyabean)
> fitR <- lm((y-16) ~ I((x-4)^2) -1, data=soyabean)
```

Note in the final line, the term '-1' in the regression formula tells R not to fit an intercept in the reduced model.

```
> sseF <- sum(residuals(fitF)^2)
> sseR <- sum(residuals(fitR)^2)
> ( (sseR-sseF)/2 ) / ( sseF/(15-3) )
[1] 15.75043
> qf(0.95, 2,12)
[1] 3.885294
```

## 6.3 One-way analysis of variance (ANOVA)

Consider the one-way classification model

$$F: \quad Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, p;\, j = 1, \ldots, n_i\,, \tag{6.2}$$

with $p$ groups and $n_i$ observations in the $i$th group. We wish to test the hypotheses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_k \text{ for some } i, k\,.$$

The null hypothesis corresponds to the statement that all of the groups have the same expected response. The alternative states that not all of the groups have identical expected responses.

We showed in Section 6.1.1 that testing $H_0$ corresponds to a comparison of model $F$ with the reduced model

$$R: \quad Y_{ij} = \mu + \epsilon_{ij}\,,$$

with one free parameter, $\mu$. It can be shown that the test statistic simplifies as follows:

$$F = \frac{(\mathsf{SSE}_R - \mathsf{SSE}_F)/(r_F - r_R)}{\mathsf{SSE}_F/(n - r_F)} = \frac{\mathsf{SS}_{\mathsf{between}}/(p-1)}{\mathsf{SS}_{\mathsf{within}}/(n-p)}\,,$$

where $\mathsf{SS}_{\mathsf{between}} = \sum_{i=1}^{n} n_i(\bar{Y}_i - \bar{Y})^2$ and $\mathsf{SS}_{\mathsf{within}} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ denote respectively the *between-group sum of squares* and the *within-group sum of squares*. The null hypothesis is rejected at significance level $\alpha$ if $F > F_{\alpha;p-1,n-p}$.

## 6.4 Testing for lack of fit

Suppose that we wish to check the fit of a particular linear model, e.g. quadratic regression,

$$R: \quad \mathbb{E}(Y) = \alpha + \beta x + \gamma x^2\,.$$

We shall do so by conducting a test of the hypotheses:

$$H_0: \mathbb{E}(Y) = \alpha + \beta x + \gamma x^2 \quad \text{vs} \quad H_1: \mathbb{E}(Y) = \mu(x)\,,$$

where $\mu(\cdot)$ is an arbitrary, unspecified function.

Such a hypothesis can be tested using an $F$-test provided that there is at least one setting of $x$ with *replication*, i.e. multiple observations. In the case with multiple predictors we require there to be a *combination* of settings of all the predictor variables with more than one observation.

Let us write the distinct settings of $x$ used as $x_1, x_2, \ldots, x_g$. We can then group together the responses according to the level of $x$. Let $Y_{ij}$ denote the $j$th observation with setting $x_i$. Then $H_0$ corresponds to the following:

$$R: \quad Y_{ij} = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_{ij}\,, \quad i = 1, \ldots, g;\ j = 1, \ldots, n_i\,,$$

$\epsilon_{ij} \sim N(0, \sigma^2)$ independently. In addition $H_1$ corresponds to the following:

$$\begin{aligned} F: \quad Y_{ij} &= \mu(x_i) + \epsilon_{ij} \\ &= \mu_i + \epsilon_{ij}\,, \end{aligned} \tag{6.3}$$

again with the $\epsilon_{ij} \sim N(0, \sigma^2)$ independently. Note that (6.3) corresponds to the one-way classification model (6.2) used in the one-way ANOVA in Section 6.3, with the level of $x$ used as the grouping factor.

The hypothesis test comparing models $F$ and $R$ can be done in R as follows. For the soya bean data, we find insufficient evidence ($p$-value 0.1843) to reject the hypothesis that the expected response function is quadratic.

```
> fitR <- lm(y~x+I(x^2), data=soyabean)
> fitF <- lm(y~factor(x), data=soyabean)
> anova(fitR,fitF)
## Analysis of Variance Table
##
## Model 1: y ~ x + I(x^2)
## Model 2: y ~ factor(x)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     12 6.5168
## 2     10 4.6467  2    1.8701 2.0123 0.1843
```

It can be shown that the general form of the test statistic is

$$F = \frac{(\mathsf{SSE}_R - \mathsf{SSE}_F)/(r_F - r_R)}{\mathsf{SSE}_F/(n - r_F)} = \frac{\mathsf{SS}_{\mathsf{LoF}}/(g - r_R)}{\mathsf{SSE}_{\mathsf{PE}}/(n - g)}\,,$$

on $g - r_R$ numerator and $n - g$ denominator degrees of freedom, where $\mathsf{SS}_{\mathsf{LoF}} = \sum_i n_i(\hat{Y}_i^R - \bar{Y}_i)^2$ and $\mathsf{SS}_{\mathsf{PE}} = \mathsf{SSE}_F = \sum_{i,j}(Y_{ij} - \bar{Y}_i)^2$ are the *lack-of-fit sum of squares* and *pure error sum of squares* respectively. There is statistically significant evidence that the mean function in model $R$ is incorrect if $F > F_{\alpha;g-r_R,n-g}$.

## 6.5 Proof of Proposition 6.2

We aim to show that if $H_0$ is true then

$$F = \frac{(\mathsf{SSE}_R - \mathsf{SSE}_F)/d}{\mathsf{SSE}_F/(n - r_F)} \sim F(d, n - r_F)\,,$$

where $d = r_F - r_R$.

We will use the following lemma, which can be proved using moment generating functions.

**Lemma 6.3.** *If $V_2 \sim \chi^2(d_2)$ and $V_3 \sim \chi^2(d_3)$,*

$$V_1 = V_2 - V_3\,,$$

*and $V_1$ and $V_3$ are independent, then $V_1 \sim \chi^2(d_2 - d_3)$.*

Let $\hat{\boldsymbol{\theta}}_R = \mathbf{A}\hat{\boldsymbol{\beta}}_{G_R} + \mathbf{b}$ denote a restricted least squares estimate of the full model parameters calculated under the constraint $H_0$. Let $\hat{\boldsymbol{\epsilon}}_F$ denote the residuals from the full model. Then we have the following.

**Lemma 6.4.** *$\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)$ and $\hat{\boldsymbol{\epsilon}}_F$ are jointly normal with $\mathrm{Cov}(\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G), \hat{\boldsymbol{\epsilon}}_F) = \mathbf{0}$, and so are independent by Lemma 2.8.*

(We omit the proof of the above lemma, which is quite lengthy, but is essentially just matrix algebra.)

Now note that

$$\begin{aligned}
\mathsf{SSE}_R &= (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{G_R})^{\mathrm{T}}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{G_R}) \\
&= (\mathbf{Y} - \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{A}\hat{\boldsymbol{\beta}}_{G_R})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{A}\hat{\boldsymbol{\beta}}_{G_R}) \\
&= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_R)^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_R) \\
&= S(\hat{\boldsymbol{\theta}}_G) + (\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G) \quad \text{by results from Chapter 2} \\
&= \mathsf{SSE}_F + (\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)\,.
\end{aligned}$$

Hence the extra sum of squares is

$$\mathsf{SS}_{H_0} = \mathsf{SSE}_R - \mathsf{SSE}_F = (\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)\,. \tag{6.4}$$

From Proposition 5.2, we know that if $H_0$ is true, then

$$\begin{aligned}
V_2 &= \frac{\mathsf{SSE}_R}{\sigma^2} \sim \chi^2(n - r_R) \\
V_3 &= \frac{\mathsf{SSE}_F}{\sigma^2} \sim \chi^2(n - r_F)\,.
\end{aligned}$$

Letting

$$V_1 = V_2 - V_3 = \frac{\mathsf{SSE}_R - \mathsf{SSE}_F}{\sigma^2} = \frac{\mathsf{SS}_{H_0}}{\sigma^2}\,,$$

we see that $V_1$ and $V_3$ are independent, since $V_1$ is a function of $\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)$ by (6.4), $V_3$ is a function of $\hat{\boldsymbol{\epsilon}}_F$, and $\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G)$ and $\hat{\boldsymbol{\epsilon}}_F$ are independent by Lemma 6.4.

Thus, $V_1$, $V_2$, and $V_3$ satisfy the conditions of Lemma 6.3 and we have that

$$V_1 \sim \chi^2(d)\,,$$

with $d = (n - r_R) - (n - r_F) = r_F - r_R$ . Hence

$$\frac{V_1/d}{V_3/(n - r_F)} \sim F(d, n - r_F)\,.$$

However,

$$\frac{V_1/d}{V_3/(n-r_F)} = \frac{(\mathsf{SSE}_R - \mathsf{SSE}_F)/(d\sigma^2)}{\mathsf{SSE}_F/((n-r_F)\sigma^2)}$$
$$= \frac{(\mathsf{SSE}_R - \mathsf{SSE}_F)/d}{\mathsf{SSE}_F/(n-r_F)} = F\,.$$

Hence $F \sim F(d, n-r_F)$ if $H_0$ is true as claimed.

*Proof of Lemma 6.4.*

$$\mathrm{Cov}(\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G), \hat{\boldsymbol{\epsilon}}) = \mathbf{0}\,.$$

By Lemma 2.8, this is enough to show that they are independent.

For joint normality, note that

$$\begin{pmatrix} \mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G) \\ \hat{\boldsymbol{\epsilon}} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_R\mathbf{Y} - \mathbf{H}_F\mathbf{Y} \\ (\mathbf{I}-\mathbf{H}_F)\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_R - \mathbf{H}_F \\ \mathbf{I}-\mathbf{H}_F \end{pmatrix} \mathbf{Y} = \mathbf{W}\mathbf{Y}\,,$$

where $\mathbf{H}_F = \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{H}_R = \tilde{\mathbf{X}}\mathbf{G}_R\tilde{\mathbf{X}}^{\mathrm{T}}$ denote the hat matrices for the full model and the reduced model, respectively. Hence, since $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I})$,

$$\begin{pmatrix} \mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G) \\ \hat{\boldsymbol{\epsilon}} \end{pmatrix} \sim N(\mathbf{W}\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{W}\mathbf{W}^{\mathrm{T}})\,.$$

$$\begin{aligned}
\mathrm{Cov}[\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G), \hat{\boldsymbol{\epsilon}}] &= \mathrm{Cov}[\mathbf{H}_R\mathbf{Y} - \mathbf{H}_F\mathbf{Y}, (\mathbf{I}-\mathbf{H}_F)\mathbf{Y}] \\
&= \mathrm{Cov}[\mathbf{H}_R\mathbf{Y}, (\mathbf{I}-\mathbf{H}_F)\mathbf{Y}] - \mathrm{Cov}[\mathbf{H}_F\mathbf{Y}, (\mathbf{I}-\mathbf{H}_F)\mathbf{Y}] \\
&= \sigma^2\left\{\mathbf{H}_R(\mathbf{I}-\mathbf{H}_F)^{\mathrm{T}} - \mathbf{H}_F(\mathbf{I}-\mathbf{H}_F)^{\mathrm{T}}\right\} \\
&= \sigma^2\left\{\mathbf{H}_R(\mathbf{I}-\mathbf{H}_F) - \mathbf{H}_F(\mathbf{I}-\mathbf{H}_F)\right\} \\
&= \sigma^2\left\{\mathbf{H}_R - \mathbf{H}_R\mathbf{H}_F - \mathbf{H}_F + \mathbf{H}_F^2\right\} \\
&= \sigma^2\left\{\mathbf{H}_R - \mathbf{H}_R\mathbf{H}_F\right\}\,. \qquad (6.5)
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_R\mathbf{H}_F &= \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{X}}\mathbf{G}_R\tilde{\mathbf{X}}^{\mathrm{T}} \\
&= \mathbf{X}\mathbf{G}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A}\mathbf{G}_R\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} \\
&= \mathbf{X}\mathbf{A}\mathbf{G}_R\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} \\
&= \tilde{\mathbf{X}}\mathbf{G}_R\tilde{\mathbf{X}}^{T} \\
&= \mathbf{H}_R\,. \qquad (6.6)
\end{aligned}$$

$$\mathrm{Cov}[\mathbf{X}(\hat{\boldsymbol{\theta}}_R - \hat{\boldsymbol{\theta}}_G), \hat{\boldsymbol{\epsilon}}] = \mathbf{0}$$

□

## 6.6 Testable hypotheses

**Definition 6.5.** *The hypothesis* $\mathbf{C}^{\mathrm{T}}\boldsymbol{\theta} = \mathbf{d}$ *is said to be* testable *if each of the functions,* $\mathbf{c}_1^{\mathrm{T}}\boldsymbol{\theta}, \mathbf{c}_2^{\mathrm{T}}\boldsymbol{\theta}, \ldots, \mathbf{c}_k^{\mathrm{T}}\boldsymbol{\theta}$*, in the constraints is an estimable function.*

If the null hypothesis is testable, then there is another way of testing $H_0$ based on using the (unbiased) least squares estimator, $\mathbf{C}^{\mathrm{T}}\hat{\boldsymbol{\theta}} \sim N(\mathbf{C}^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\mathbf{C}^{\mathrm{T}}\mathbf{G}\mathbf{C})$. The test statistic is

$$F = \frac{(\mathbf{C}^{\mathrm{T}}\hat{\boldsymbol{\theta}} - \mathbf{d})^{\mathrm{T}}(\mathbf{C}^{\mathrm{T}}\mathbf{G}\mathbf{C})^{-1}(\mathbf{C}^{\mathrm{T}}\hat{\boldsymbol{\theta}} - \mathbf{d})/k}{\mathsf{SSE}_F/(n - r_F)} \overset{H_0}{\sim} F(k, n - r_F)$$

and we reject $H_0$ if $F > F_{\alpha;k,(n-r_F)}$. It can be shown that for testable hypotheses this procedure is equivalent to the test based on the difference of the residual sums of squares. Note that the above procedure does not work if the hypothesis is not testable, as in this case the least squares estimator is biased.