DR RICHARD WILKINSON

LINEAR MODELS

UNIVERSITY OF NOTTINGHAM

THESE NOTES ARE SPLIT into two parts (theory and case-studies) in the hope that this makes it easier for you to revise the theory for the exam, and to learn how to put this theory into practice for the two pieces of assessed coursework. The disadvantage of this structure is that **you will need to bring both parts to each lecture**.

This section of notes contains only the theory and some simple theoretical examples. I have significantly rewritten and reorganised the notes this year based on feedback from last year's students. Greater focus has now been placed on application rather than theory and proof, and I have removed some technical material that is not necessary in the age of ready accessibility to computers and free statistical software.

The case-study handout contains a detailed analysis of some real datasets. The case-studies contain a lot of material - this is to help you do as good a job as possible with the coursework and for use as a reference. It is not necessary for you to memorise everything in the case studies for the exam. However, it is important that you master both theory and practice.

Please email any corrections or suggestions to

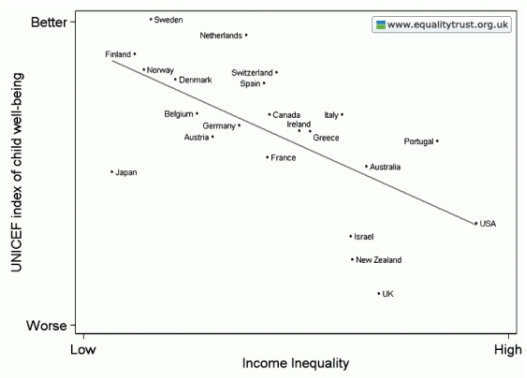
r.d.wilkinson@nottingham.ac.uk

1

Introduction to linear models

1.1 What is regression?

Regression is the name given to a huge collection of statistical techniques used to analyse the relationship between two or more variables. Regession aims to exploit the pattern of correlations in the data to provide simple models that have explanatory and/or predictive power.



The first regression analysis was performed by Legendre in 1805 and by Gauss in 1809 to determine the orbits of bodies about the Sun from astronomical observations. Since then, regression has been used in pretty much every field imaginable.

Linear regression is the topic of this module. It is the simplest and most important case of regression analysis, and despite the myriad of techniques developed by statisticians over the past two centuries, it is still one of the most common types of statistical analysis performed, and the basis of much of the analysis done in the empirical sciences.

1.2 Data and variables

- We will consider data of the form:

  $$\left\{(\boldsymbol{x}_i^T, y_i) \ : \ i = 1, \ldots, n\right\} \qquad (\boldsymbol{x}_i^T \text{ – row vector of dimension } k)$$

  i.e. there are n *experimental units* or *cases*, each yielding a set of measurements

  1. y_i is the response or outcome for the ith case,
  2. $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is the i^{th} covariate consisting of k input variables.

- The covariates may be
  1. continuous
  2. discrete: e.g. binary, nominal scale (no ordering) or ordinal scale (some ordering).

  Discrete variables are called *factors* and the values taken by a factor are called *levels*.

Warning: In previous modules all random variables are given a capital letter, e.g. X, and so are matrices e.g. $\boldsymbol{A}$. This module contains both matrices and random vectors, and so you need to keep track of what is random and what isn't yourself.

1.3 A good strategy for statistical modelling

A. Propose a *model* for the data (i.e. a parametric formula linking the response variable with the input variables, recognising the stochastic nature of the response).

B. Fit the model (e.g. find the best set of parameters).

C. Ask 'Is the model adequate?' (i.e. consistent with the data). Does it allow the main questions of the analysis to be answered?

D. Fit other plausible models, compare them, and choose the best.

E. Use the best model to answer the questions of interest.

This isn't the order we cover the material. We do something like B, D,E, C with A covered throughout.

1.4 The class of linear models

Consider models of the form

$$y_i = f\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) + \epsilon_i \qquad i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $p-$vector of parameters. Without loss of generality we can assume $\mathbb{E}[\epsilon_i] = 0$.

This is the most general class of *regression models*. The aim in regression analysis is to find a good choice of the value of $\boldsymbol{\beta}$, and perhaps also of the regression functions f.

In this module we consider the restricted class of **linear models**

$$f\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) = \boldsymbol{g}\left(\boldsymbol{x}_i\right)^T \boldsymbol{\beta} \tag{1.2}$$

where $\boldsymbol{g}$ is some *known* (i.e. chosen by us) function that transforms the $k-$vector $\boldsymbol{x}_i$ into a $p-$vector $\boldsymbol{g}\left(\boldsymbol{x}_i\right)$ (generally, p need not equal k). We also assume that for $i \neq j$, $\mathbb{C}\mathrm{ov}(\epsilon_i, \epsilon_j) = 0$ and $\mathbb{V}\mathrm{ar}(\epsilon_i) = \mathbb{V}\mathrm{ar}(\epsilon_j)$.

For example,

- If $k = 1$, then $g(x) = (1, x)^\top$ is a straight line with an intercept and $g(x) = (1, x, x^2)^\top$ is a quadratic.

1.4.1 The design matrix

If we collect the $\boldsymbol{g}(\boldsymbol{x}_i)$'s into rows $\boldsymbol{g}(\boldsymbol{x}_i)^T$, $i = 1, 2, \ldots, n$ of a $(n \times p)$ matrix $\boldsymbol{Z}$, we then obtain a matrix representation for equations (1.1) and (1.2):

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{Z}\,\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}}.$$

This is the general form of a linear model.

- $\boldsymbol{y}$ is called the **response**
- $\boldsymbol{x}$ is called the **explanatory variable, the regressors, or the covariates**.
- Z is called the **design matrix**.

Notes:

1. $E[\boldsymbol{y}] = \boldsymbol{Z}\boldsymbol{\beta}$ is called the *linear structure of the model*.
2. A linear model must be linear in the parameters (β_j's) but not necessarily linear in the input variables ($\boldsymbol{x}_1$,$\boldsymbol{x}_2$,...,$\boldsymbol{x}_k$).
3. Once we have found $\boldsymbol{Z}$, we can do without $\boldsymbol{g}(\boldsymbol{x})$ and so usually we will skip straight to $\boldsymbol{Z}$.
4. One dataset can lead to many different models – i.e. many different design matrices can be formed from the same cases $\times$ variables array.

E.g. $y = a + bx^2 + c\log x + \epsilon$ is still a linear model, but $y = a + \cos(bx) + \epsilon$ is not.

Why use linear models?

1. It is a surprisingly large class of useful models.
2. The theory is well developed (i.e. analytically tractable).
3. Numerical aspects are relatively easy.
4. Lots of nice properties (see Chapter 2).
5. Long and proven history of success.

1.5 Simple and multiplelinear regression

Simple linear regression refers to the case where there is a single covariate and we wish to fit the linear model

$$y_i = a + bx_i + \epsilon_i, \text{ where } i = 1, \ldots, n. \tag{1.3}$$

In this model we have just two parameters and one input variable and we fit a straight line through the data.

What are $\boldsymbol{\beta}$, $\boldsymbol{g}(x)^T$ and $\boldsymbol{Z}$ here?

Multiple linear regression refers to the case where there are multiple covariates, $x_1, x_2, \ldots, x_k$ say, and we wish to fit the model

$$y_i = a + b_1 x_{1i} + \ldots + b_k x_{ki} + \epsilon_i, \text{ where } i = 1, \ldots, n. \qquad (1.5)$$

What are $\boldsymbol{\beta}$, $\boldsymbol{g}(x)^T$ and $\boldsymbol{Z}$ here?

Once you have done this a few times, you will be able to go directly from (1.5) to Z skipping g.

1.6 A salutary example

$y_i =$ final mark received in the Linear Models module for student i

$$x_i = \begin{cases} 0 \text{ if student i attended at most 1 problem class} \\ 1 \text{ if student i attended at least 2 problem classes} \end{cases}$$

$$\boldsymbol{y}^T = (73, 41, 85, 49, 69, 63, 54, 70, 59, 49, 44, 74, 70, 0, \ldots, 46, 41, 70)$$
$$\boldsymbol{x}^T = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, \ldots, 0, 0, 1).$$

A possible model of these data is

$$y_i = \begin{cases} a + \epsilon_i & \text{if } x_i = 0 \\ a + b + \epsilon_i & \text{if } x_i = 1 \end{cases}$$

What are $\boldsymbol{Z}$ and $\boldsymbol{\beta}$?

What quantity does b correspond to?

After fitting the model, we find $\widehat{a} = 53.0$, $\widehat{b} = 13.1$ – What conclusion do you draw?

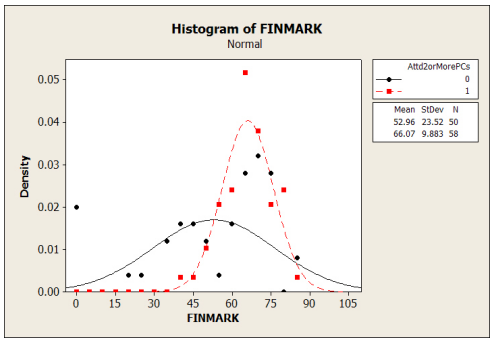


Figure 1.1: Histograms of the student for each category.

This can also be expressed as

$$y_i = a + bx_i + \epsilon_i$$

,

This is called a one-way analysis of variance (ANOVA) model, as there is one covariate, which is a discrete factor.

2

Model fitting: Least squares estimation

2.1 The least squares estimator

See Section 2.10 for a reminder of some useful matrix algebra results.

Consider the linear model $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.

Definition 1. *The (ordinary) least squares (OLS) estimator of $\boldsymbol{\beta}$ is the vector $\widehat{\boldsymbol{\beta}}$ which minimizes the sum of squared differences*

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i])^2. \end{aligned}$$

Proposition 1. *Assume that $\boldsymbol{Z}$ is of rank p, so that $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ exists. Then the ordinary least squares estimate of β is*

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{y}.$$

2.1.1 Examples

TWO OBSERVATIONS with $\mathbb{E}[y_1] = \theta, \mathbb{E}[y_2] = 2\theta$:

CONSIDER THE STRAIGHT LINE[1]

$$y_i = a + b\,(x_i - \overline{x}) + \epsilon_i\ ,\ i = 1, \ldots, n.$$

[1] This is the simple linear regression model, as there is only one covariate. This gives the simple formulae you saw in G11STA. You should always use the matrix form of the estimator for this module!

2.1.2 **Proof of Proposition 1 (R0)**

We want to prove that

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}.$$

First find the stationary points of $S(\boldsymbol{\beta})$.

At the stationary point $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, these derivatives must be zero.

$$-2\boldsymbol{Z}^T\boldsymbol{y} + 2\boldsymbol{Z}^T\boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{0} \qquad (2.1)$$

(2.1) are the **normal equations**.

Next, we show that any solution to (2.1) is a minimum of $S(\boldsymbol{\beta})$:

Recall that $\widehat{\boldsymbol{\beta}}$ minimizes $S(\boldsymbol{\beta})$ if the Hessian matrix,

$$\frac{d^2S}{d\boldsymbol{\beta}^2}$$

is a positive definite matrix.

Finally we solve the normal equations to find $\widehat{\boldsymbol{\beta}}$ explicitly.
In general $\boldsymbol{Z}$ is chosen to be of full rank, i.e.,

$$\operatorname{rank}(\boldsymbol{Z}) = p$$

thus

$$\operatorname{rank}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right) = p$$

(i.e. $\boldsymbol{Z}^T\boldsymbol{Z}$ is non-singular and so its inverse exists). Then from (2.1)

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}.$$

Note that if Z was a square invertible matrix, then $\widehat{\boldsymbol{\beta}} = \boldsymbol{Z}^{-1}\boldsymbol{y}$. In general Z won't be a square matrix (and so cannot be inverted). The term $\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ is acting as a pseudo-inverse, and is sometimes called the Moore-Penrose pseudo-inverse of Z.

Minimising the sum of squared differences $S(\boldsymbol{\beta})$ is called the **method of least squares**. Since our estimator uses $\boldsymbol{y}$ (a random vector) to produce an estimate for $\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\beta}}$ is also a random vector. We call $\widehat{\boldsymbol{\beta}}$ the **least-squares estimate** of $\boldsymbol{\beta}$.

2.2 Some definitions

Definition 2. *The **fitted values** are given by $\widehat{\boldsymbol{y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$, i.e. $\widehat{y}_i = \boldsymbol{z}_i\widehat{\boldsymbol{\beta}}$ where $\boldsymbol{z}_i$ is the i^{th} row of $\boldsymbol{Z}$. $\widehat{y}_i$ is the expected value of the fitted model for observation i.*

Definition 3. *The i^{th} **residual** is given by $\widehat{\epsilon}_i = y_i - \widehat{y}_i$. The residuals are the difference between the observed values and the fitted values. We write $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ for the vector of residuals.*

2.2.1 Generalised Expectation and Variance

Before we can prove some properties of $\widehat{\boldsymbol{\beta}}$, we require a generalised form of expectation and variance.

Let $\boldsymbol{X} = \left[X_{ij}\right]$ be a matrix of random variables (r.v.s) with ijth entry X_{ij}. Then $\mathbb{E}\left[\boldsymbol{X}\right]_{ij} := \mathbb{E}\left[X_{ij}\right]$, i.e., the expectation of a matrix is the matrix of expected values.

Recall that $\mathbb{E}$ is a linear operator: if $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ are constant matrices, and if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are vectors of r.v.'s., then

$$\mathbb{E}\left[\boldsymbol{AXB} + \boldsymbol{CY} + \boldsymbol{D}\right] = \boldsymbol{A}\mathbb{E}(\boldsymbol{X})\boldsymbol{B} + \boldsymbol{C}\mathbb{E}(\boldsymbol{Y}) + \boldsymbol{D}.$$

Definition 4. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be vectors of random variables of length p and q respectively. The* **covariance matrix** *$\mathbb{C}ov\left(\boldsymbol{X}, \boldsymbol{Y}\right)$ is defined to be a $p \times q$ matrix with*

$$\mathbb{C}ov\left(\boldsymbol{X}, \boldsymbol{Y}\right)_{ij} \quad := \quad \mathbb{C}ov\left(X_i, Y_j\right)$$

where

$$\mathbb{C}ov\left(X_i, Y_j\right) = \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(Y_j - \mathbb{E}[Y_j]\right)\right]$$

is the usual univariate covariance.

Properties:

(i) $\mathbb{C}\mathrm{ov}\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right]\right)\left(\boldsymbol{Y} - \mathbb{E}\left[\boldsymbol{Y}\right]\right)^T\right]$

$$= \mathbb{E}\left[\boldsymbol{X}\boldsymbol{Y}^T\right] - \mathbb{E}\left[\boldsymbol{X}\right]\mathbb{E}\left[\boldsymbol{Y}\right]^T$$

(ii) If $\boldsymbol{a}$, $\boldsymbol{b}$ are vectors of constants with lengths p and q respectively, then

$$\mathbb{C}\text{ov}\left(\boldsymbol{X}-\boldsymbol{a},\ \boldsymbol{Y}-\boldsymbol{b}\right) = \mathbb{C}\text{ov}\left(\boldsymbol{X},\boldsymbol{Y}\right)$$

(iii) If $\boldsymbol{A}$, $\boldsymbol{B}$ are constant matrices (of the correct size) then

$$\mathbb{C}\text{ov}\left(\boldsymbol{A}\boldsymbol{X},\ \boldsymbol{B}\boldsymbol{Y}\right) = \boldsymbol{A}\mathbb{C}\text{ov}\left(\boldsymbol{X},\ \boldsymbol{Y}\right)\boldsymbol{B}^T.$$

Definition 5. *For a $(p \times 1)$ vector $\boldsymbol{X}$, we define the* **variance-covariance matrix**, $\mathbb{V}ar\left(\boldsymbol{X}\right)$ *to be*

$$\mathbb{V}ar\left(\boldsymbol{X}\right) := \mathbb{C}ov\left(\boldsymbol{X},\ \boldsymbol{X}\right)$$

and so

$$\mathbb{V}ar\left(\boldsymbol{X}\right) = \begin{bmatrix} \mathbb{V}ar(X_1) & \mathbb{C}ov(X_1,X_2) & \dots & \mathbb{C}ov(X_1,X_p) \\ \mathbb{C}ov(X_2,X_1) & \mathbb{V}ar(X_2) & & \\ \vdots & & \ddots & \\ \mathbb{C}ov(X_p,X_1) & & & \mathbb{V}ar(X_p) \end{bmatrix}_{(p\times p)}$$

Properties:

(i) $\mathbb{V}\text{ar}\left(\boldsymbol{X}\right)$ is symmetric. And positive semi-definite.

(ii) $\mathbb{V}\text{ar}\left(\boldsymbol{X}-\boldsymbol{a}\right) = \mathbb{V}\text{ar}\left(\boldsymbol{X}\right)$

(iii) $\mathbb{V}\text{ar}\left(\boldsymbol{A}\boldsymbol{X}\right) = \boldsymbol{A}\mathbb{V}\text{ar}\left(\boldsymbol{X}\right)\boldsymbol{A}^T$ Important and useful

Proofs are left as exercises.

2.3 Properties of the OLS estimator $\hat{\beta}$

Consider the linear model for n observations and p parameters

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times p}{\boldsymbol{Z}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}, \quad \text{with} \quad \begin{aligned} \mathbb{E}\left[\boldsymbol{\epsilon}\right] &= \boldsymbol{0} \\ \mathbb{V}\text{ar}\left(\boldsymbol{\epsilon}\right) &= \sigma^2\boldsymbol{I}_n \end{aligned}$$

Assume that $\boldsymbol{Z}$ is of rank p, so that $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ exists.

Theorem 1. *Then*

(R0) The least squares estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}$

(R1) $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, i.e., $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$

(R2) $\mathbb{V}ar\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}$

(R3) $s^2 = \frac{1}{n-p}\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)$ is an unbiased estimator for σ^2

2.3.1 Proofs of Theorem 1

(R1) $\widehat{\boldsymbol{\beta}}$ is unbiased.

(R2) $\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$.

2.4 The null model

The **null model** fits a horizontal line to the data, i.e.

$$y_i = a + \epsilon_i \qquad i = 1, \ldots, n.$$

The null model is the simplest possible model. There is just one parameter in the model, a. What is $g(x)$ and Z?

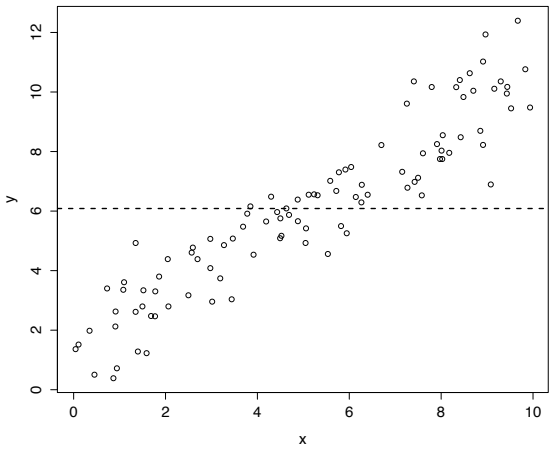


Figure 2.1: Some data with the fitted null model.

- To estimate a by least squares: minimise $S(a) = \sum_{i=1}^{n} (y_i - a)^2$

$$\hat{a} = \overline{y}$$

- This model is only appropriate for a dataset where y is unrelated to any of the input variables.

2.5 Linear transformations

It is possible to parameterise a model in more than one way. Two models that have the same fitted values (and hence residuals) for any vector of observations $\boldsymbol{y}$, can be considered to be the same model.

Suppose that we want to rewrite our model in terms of another parameter vector γ. Assume that $\gamma = \boldsymbol{A}\boldsymbol{\beta}$ ($\boldsymbol{A}$ is a $p \times p$ non-singular matrix) and so our model becomes $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{A}^{-1}\gamma + \boldsymbol{\epsilon}$, which has design matrix $\boldsymbol{Z}\boldsymbol{A}^{-1}$. Then:

$$\begin{aligned} \widehat{\gamma} &= \left(\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\right)^T\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\right)\right)^{-1}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\right)^T\boldsymbol{y} \\ &= \boldsymbol{A}\widehat{\boldsymbol{\beta}} \end{aligned}$$

and the fitted values are:

$$\begin{aligned} \widehat{\boldsymbol{y}} &= \boldsymbol{Z}\boldsymbol{A}^{-1}\widehat{\gamma} \\ &= \boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Z}\widehat{\boldsymbol{\beta}} \end{aligned}$$

i.e. fitted values (and thus residuals) are unchanged by a reparametrisation of the form $\gamma = \boldsymbol{A}\boldsymbol{\beta}$.

This is comforting!

2.6 Deviance and R^2

2.6.1 Deviance

Definition 6. $D = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}})$ *is called the model* ***deviance*** *or the* ***residual sum of squares*** *(ResidSS), as* $D = \sum_{i=1}^{n}\widehat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}$.

The deviance is a measure of model fit. Unfortunately, this depends on the scale of the y_is. However, we can standardise it using the deviance of the null model.

- The deviance of the null model is

$$D_0 = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

and is called the **total sum of squares** (TSS).

2.6.2 Decomposition of the total sum of squares

Theorem 2. *If the null model is nested within the full model, then*

$\sum_{i=1}^{n}(y_i - \overline{y})^2$ is called the total sum of squares

$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ is called the residual sum of squares (or Deviance)

$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ is called the regression sum of squares.

$$\sum_{i=1}^{n}(y_i - \overline{y}\,)^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y}\,)^2 \qquad (2.2)$$

$$\textit{total SoS} = \textit{residual SoS} + \textit{regression SoS}$$

This equation is called the **decomposition of the total sum of squares**.

Proof. First we show that if the design matrix contains a column of ones, then the residuals must sum to zero. From the normal equations,

$$\begin{aligned}
\boldsymbol{Z}^\top \boldsymbol{Z}\widehat{\boldsymbol{\beta}} &= \boldsymbol{Z}^\top \boldsymbol{y} \\
\Rightarrow \mathbf{0} &= \boldsymbol{Z}^\top (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) \\
\Rightarrow \mathbf{0} &= \boldsymbol{Z}^\top \widehat{\boldsymbol{\epsilon}}
\end{aligned}$$

and by considering the row of ones in $\boldsymbol{Z}^\top$, we find that the residuals must sum to zero for any model that has the null model nested within it.

Thus, $0 = \sum_{i=1}^{n} \widehat{\epsilon}_i = \sum_{i=1}^{n} (y_i - \widehat{y}_i)$.

Now,

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \overline{y}\,)^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y}\,)^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y}\,)^2 + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}\,)
\end{aligned}$$

but,

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}\,) &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)\widehat{y}_i - \sum_{i=1}^{n}(y_i - \widehat{y}_i)\,\overline{y} \\
&= \sum_{i=1}^{n}\widehat{y}_i(y_i - \widehat{y}_i) - \overline{y}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \\
&= \sum_{i=1}^{n}\widehat{y}_i(y_i - \widehat{y}_i) \qquad \text{– because the sum of residuals is zero} \\
&= \widehat{\boldsymbol{y}}^\top(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \\
&= (\boldsymbol{Z}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) \\
&= \widehat{\boldsymbol{\beta}}^\top(\boldsymbol{Z}^\top\boldsymbol{y} - \boldsymbol{Z}^\top\boldsymbol{Z}\widehat{\boldsymbol{\beta}}) \\
&= \widehat{\boldsymbol{\beta}}^\top \mathbf{0}_{p\times 1} \qquad \text{from the normal equations} \\
&= 0.
\end{aligned}$$

Hence,

$$\sum_{i=1}^{n}(y_i - \overline{y}\,)^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y}\,)^2.$$

□

□

2.6.3 R^2

Definition 7. *R^2 is defined to be*

Also called the coefficient of determination

$$\begin{aligned} R^2 &= 1 - \frac{\textit{Deviance(current model)}}{\textit{Deviance(null model)}} \\ &= 1 - \frac{D_1}{D_0} \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \end{aligned}$$

The deviance of the null model gives a baseline to compare the current model to. Not all models can be compared with the null model in this way. The comparison only makes sense when the null model is nested within the current model, i.e., when the model contains an intercept.

For example, the model $Y = \beta X$ cannot be compared to the null model $Y = \alpha$, but the model $Y = \alpha + \beta X$ can.

Properties of R^2:

(1) $R^2 = \dfrac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \dfrac{\text{Regression sum of squares}}{\text{Total sum of squares}},$

(2) $0 \leqslant R^2 \leqslant 1,$

Proof.

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

$$\text{total SoS} = \text{residual SoS} + \text{regression SoS}$$

Note that the deviance of the null model is equal to the total sum of squares,

$$\text{TotalSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

By applying the decomposition to the definition of R^2, we find (1).

To prove (2) note that sums of squares must be non-negative and so from the definition $R^2 \leqslant 1$ and (1) implies $R^2 \geqslant 0$. □

Interpretation

- We can use R^2 to compare the fit of models that are not nested, as long as the null model is nested within them both. For example, $Y = \alpha + \beta X$ can be compared with $Y = \alpha + \beta e^X$.
- R^2 can be interpretted as the proportion of the variation in the response that is absorbed by the model. By the decomposition of the total sum of squares, the 'left-over' variation in the response is represented in the residual sum of squares.

However, care needs to be taken. The deviance is **not** a good absolute measure, since if model M_1 is nested in M_2, then $D_{M_1} \geq D_{M_2}$.

thus even adding random variables into M_1 will improve the deviance and R^2

Because of this, the adjusted R-squared was proposed.

Definition 8. *The adjusted R^2 corrects R^2 to account for the number of variables and is defined to be*

$$R^2_{adj} = 1 - \frac{s^2(current)}{s^2(null)}.$$

Often this measure is preferred to R^2, as it takes into account p, the number of parameters in the current model, giving a preference for more parsimonious models.

With a little work, we can show that

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right).$$

2.6.4 An extra example

Consider the linear model with

$$\mathbb{E}[y_1] = \alpha \qquad \mathbb{E}[y_2] = \beta \qquad \mathbb{E}[y_3] = \alpha - \beta.$$

(a) Calculate the design matrix $\boldsymbol{Z}$ and the estimator $\widehat{\boldsymbol{\beta}}$.

(b) Given that $\boldsymbol{y} = (6, 9, 3)^T$, give an unbiased estimate for σ^2.

(c) What is the deviance?

2.7 Residuals and the hat matrix

$$\begin{aligned}\widehat{\boldsymbol{\epsilon}} &= \boldsymbol{y} - \widehat{\boldsymbol{y}} \\ &= \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{y} - \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\right)\boldsymbol{y} \\ &= (\boldsymbol{I}_n - \boldsymbol{P})\,\boldsymbol{y}, \quad \text{say,}\end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ is called the 'hat matrix'. Note that $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$ so $\boldsymbol{P}$ adds a hat!

Properties of the hat matrix:

(i) $\boldsymbol{P}^T = \boldsymbol{P},\ \boldsymbol{P}^2 = \boldsymbol{P}$ i.e. $\boldsymbol{P}$ is symmetric idempotent.

(ii) $(\boldsymbol{I}_n - \boldsymbol{P})$ is symmetric idempotent.

(iii) $\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}) = n - p = \operatorname{rank}(\boldsymbol{I}_n - \boldsymbol{P})$.

Proof. (i)

(ii)

(iii) $\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}) = \operatorname{tr}(\boldsymbol{I}_n) - \operatorname{tr}(\boldsymbol{P})$ and

$\operatorname{tr}(\boldsymbol{P}) = \operatorname{tr}\left(\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\right) = \operatorname{tr}(\boldsymbol{I}_p) = p$

(since $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$)

$\therefore \operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}) = n - p$

and rank = trace for any idempotent matrix.

{trace = sum of diagonal elements.}

□

2.8 Gauss-Markov Theorem

Theorem 3. *Let $\boldsymbol{y}$ be a random vector with*

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{Z}\boldsymbol{\beta}, \quad \mathbb{V}ar(\boldsymbol{y}) = \sigma^2\boldsymbol{I}_n, \quad \boldsymbol{Z} \text{ is } (n \times p) \text{ with rank } p.$$

Then $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ is the unique linear unbiased *estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ with minimum variance.*

We say that $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ is the BLUE estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. Note that the linearity here, is linearity in $\boldsymbol{y}$.

Proof. (i) Linearity:

(ii) Unbiasedness:

(iii) Best:

Let $\boldsymbol{b}^T\boldsymbol{y}$ be a second linear unbiased estimate of $\boldsymbol{a}^T\boldsymbol{\beta}$. Then

$$\begin{aligned} \boldsymbol{a}^T\boldsymbol{\beta} &= \mathbb{E}\left[\boldsymbol{b}^T\boldsymbol{y}\right] = \boldsymbol{b}^T\mathbb{E}\left[\boldsymbol{y}\right] = \boldsymbol{b}^T\boldsymbol{Z}\boldsymbol{\beta} \\ &\Rightarrow \left(\boldsymbol{a}^T - \boldsymbol{b}^T\boldsymbol{Z}\right)\boldsymbol{\beta} = 0 \qquad \forall\boldsymbol{\beta} \\ &\Rightarrow \boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{Z} \end{aligned} \tag{2.3}$$

Now

$$\mathbb{V}\mathrm{ar}\left(\boldsymbol{b}^T\boldsymbol{y}\right) = \boldsymbol{b}^T\mathbb{V}\mathrm{ar}\left(\boldsymbol{y}\right)\boldsymbol{b} = \sigma^2\boldsymbol{b}^T\boldsymbol{b}$$

and

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) &= \sigma^2\boldsymbol{a}^T\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{a} \\ &= \sigma^2\boldsymbol{b}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{b} \end{aligned}$$

So,

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left(\boldsymbol{b}^T\boldsymbol{y}\right) - \mathbb{V}\mathrm{ar}\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) &= \sigma^2\boldsymbol{b}^T\boldsymbol{b} - \sigma^2\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T\left(\boldsymbol{I}_n - \boldsymbol{P}\right)\boldsymbol{b} \\ &= \sigma^2\boldsymbol{b}^T\left(\boldsymbol{I}_n - \boldsymbol{P}\right)^2\boldsymbol{b} \end{aligned} \tag{2.4}$$

Let $\boldsymbol{d}^T = \boldsymbol{b}^T\left(\boldsymbol{I}_n - \boldsymbol{P}\right)^T$ then (2.4) becomes

$$\mathbb{V}\mathrm{ar}\left(\boldsymbol{b}^T\boldsymbol{y}\right) - \mathbb{V}\mathrm{ar}\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) = \sigma^2\boldsymbol{d}^T\boldsymbol{d} \geqslant 0. \tag{2.5}$$

i.e. $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ has minimum variance in the class of linear unbiased estimators.

(iv) Uniqueness

Let $\boldsymbol{b}^T\boldsymbol{y}$ be a 2nd linear unbiased estimator with

$$\mathbb{V}\mathrm{ar}\left(\boldsymbol{b}^T\boldsymbol{y}\right) = \mathbb{V}\mathrm{ar}\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right).$$

Then from (2.5), $\boldsymbol{d} = \mathbf{0}$

$$\begin{aligned}
\Rightarrow \boldsymbol{b}^T(\boldsymbol{I}_n - \boldsymbol{P}) &= \mathbf{0} \\
\Rightarrow \boldsymbol{b}^T &= \boldsymbol{b}^T\boldsymbol{P} \\
&= \boldsymbol{b}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T \\
&= \boldsymbol{a}^T\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T \quad \text{using (2.3)} \\
\Rightarrow \boldsymbol{b}^T\boldsymbol{y} &= \boldsymbol{a}^T\widehat{\boldsymbol{\beta}}
\end{aligned}$$

□

Corollary:

If $\boldsymbol{a}^T = (0, 0, \ldots, 1, 0, \ldots, 0)$ (1 in i^{th} position), then $\widehat{\beta}_i$ is the best linear unbiased estimate (BLUE) of β_i.

- These properties 'characterise' the least squares estimator, but non-linear or biased estimators may also have good properties.
- Note that no distributions have been assumed at any stage.
- Note also that we don't require the errors to be independent and identically distributed. We only need them to be uncorrelated and homoscedastic (constant variance).

2.8.1 Illustrative example:

Is y_i the best linear unbiased estimator for $\mathbb{E}[y_i]$?

2.9 Unbiased estimation of σ^2

Theorem 4. (R3) *Let* $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{Z\beta}$ *and* $\mathbb{V}ar(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$. *Then*

$$s^2 = \frac{1}{n-p}\left(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)$$

is an unbiased estimator of σ^2.

Before we prove the theorem, we need a lemma.

Lemma 1. *Let* $\boldsymbol{Y}$ *be an* $(n \times 1)$ *vector of random variables with* $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\Theta}$ *and* $\mathbb{V}ar(\boldsymbol{Y}) = \boldsymbol{\Sigma} = \left[(\sigma_{ij})\right]_{(n\times n)}$.

If $\boldsymbol{A} = \left[(a_{ij})\right]_{(n\times n)}$ *is a symmetric matrix, then*

$$\mathbb{E}\left[\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}\right] = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\Theta}^T\boldsymbol{A\Theta}.$$

Expectation of a quadratic form

Proof. [non-examinable]

$$(\boldsymbol{Y} - \boldsymbol{\Theta})^T \boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\Theta}) = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} - 2\boldsymbol{\Theta}^T\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{\Theta}^T\boldsymbol{A\Theta}$$

and

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}\right] &= \mathbb{E}\left[\sum_{i=1}^n\sum_{j=1}^n (Y_i - \theta_i)\, a_{ij}\, (Y_j - \theta_j)\right] + 2\mathbb{E}\left[\boldsymbol{\Theta}^T\boldsymbol{A}\boldsymbol{Y}\right] - \boldsymbol{\Theta}^T\boldsymbol{A\Theta} \\
&= \left(\sum_{i=1}^n\sum_{j=1}^n a_{ij}\mathbb{E}\left[(Y_i - \theta_i)(Y_j - \theta_j)\right]\right) + \boldsymbol{\Theta}^T\boldsymbol{A\Theta} \\
&= \sum_{i=1}^n\sum_{j=1}^n a_{ij}\sigma_{ij} + \boldsymbol{\Theta}^T\boldsymbol{A\Theta} \\
&= \operatorname{tr}(\boldsymbol{A\Sigma}) + \boldsymbol{\Theta}^T\boldsymbol{A\Theta}.
\end{aligned}$$

□

Proof of theorem. In Section 2.7, we showed that $\left(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}$. Then

$$\begin{aligned}
\left(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right) &= \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y} \\
&= \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}
\end{aligned}$$

using the properties of the hat matrix.

When we apply the lemma to $\boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}$, we get

$$\mathbb{E}\left[\boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}\right] = \sigma^2 \operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}) + (\boldsymbol{Z\beta})^T(\boldsymbol{I}_n - \boldsymbol{P})(\boldsymbol{Z\beta})$$

but tr $(\boldsymbol{I}_n - \boldsymbol{P}) = n - p$ (from section 2.7) and

$$\begin{aligned} \boldsymbol{\beta}^T \boldsymbol{Z}^T (\boldsymbol{I}_n - \boldsymbol{P}) \boldsymbol{Z}\boldsymbol{\beta} &= 0 \\ \{\text{since } \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right) &= \boldsymbol{0}_{p\times p}\} \\ \therefore \mathbb{E}\left[\boldsymbol{y}^T (\boldsymbol{I}_n - \boldsymbol{P}) \boldsymbol{y}\right] &= (n-p)\sigma^2 \\ \Rightarrow \mathbb{E}\left[s^2\right] &= \sigma^2 . \end{aligned}$$

□

2.10 Some useful matrix algebra

Recall:

1. $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$,
2. $\frac{\partial}{\partial\boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial}{\partial\beta_1} \\ \vdots \\ \vdots \\ \frac{\partial}{\partial\beta_p} \end{bmatrix}$,
3. $\frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}^T\boldsymbol{a}\right) = \boldsymbol{a}$
4. $\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}\right) &= (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{\beta} \\ &= 2\boldsymbol{A}\boldsymbol{\beta} \quad \text{when } \boldsymbol{A} \text{ is symmetric.} \end{aligned}$
5. $\text{rank}(A) =$ the number of linearly independent (LI) columns of A = number of LI rows of A.
6. $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(A^TA)$
7. If P and Q are conformable non-singular matrices, then $\text{rank}(PAQ) = \text{rank}(A)$.
8. A is idempotent if $A^2 = A$.
9. A symmetric matrix A is positive definite if $\boldsymbol{x}^\top A\boldsymbol{x} > 0$ for every non-zero column vector $\boldsymbol{x}$. It is positive semi-definite if $\boldsymbol{x}^\top A\boldsymbol{x} \geq 0$. Note that variance matrices must by definition be positive semi-definite.
10. We can diagonalise any real symmetric matrix Σ as

$$\Sigma = ADA^\top$$

where A is an orthonormal matrix with $A^\top A = AA^\top = I$ and D is a diagonal matrix consisting of eigenvalues of Σ.

3

Normal linear models

We now make the additional assumption that $\epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. This class of models is known as the class of **normal linear models**. Equivalently,

So far we have not made any assumptions concerning the distribution of the errors ϵ_i

$$y_i \overset{\text{iid}}{\sim} N\left(z_i^\top \boldsymbol{\beta}, \sigma^2\right) \quad i = 1, \ldots, n.$$

In matrix form the normal linear model is written as

$z_i^\top$ is the i^{th} row of $\boldsymbol{Z}$.

$$\boldsymbol{y} \sim N_n\left(\boldsymbol{Z\beta}, \sigma^2 \boldsymbol{I}_n\right),$$

i.e. $\boldsymbol{y}$ follows a multivariate normal distribution, where $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution in n dimensions with mean vector $\boldsymbol{\mu}$ $(n \times 1)$ and $n \times n$ covariance matrix $\boldsymbol{\Sigma}$.

3.1 Distribution theory for $\hat{\beta}$

We can now strengthen Theorem 1, our main result about the distribution of $\widehat{\boldsymbol{\beta}}$.

Theorem 5. *For the normal linear model*

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \epsilon$$

where $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right), \ i = 1, \ldots, n$, *then,*

(R4) $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y}$ *is the maximum likelihood estimator (MLE) of* $\boldsymbol{\beta}$

(R5) $\widehat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1}\right)$

(R6) $\dfrac{\widehat{\beta}_i - \beta_i}{s\sqrt{d_{ii}}} \sim t_{n-p}$, *where* d_{ii} *is the* i^{th} *diagonal element of* $\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1}$.

3.1.1 Proofs

Theorem (R4). *$\widehat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.*

Proof. The likelihood is

$$L\left(\boldsymbol{y};\boldsymbol{\beta},\sigma^2\right) = \frac{\left|\sigma^2\boldsymbol{I}_n\right|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}-\boldsymbol{Z\beta})^\top(\boldsymbol{y}-\boldsymbol{Z\beta})\right]\right\}$$

and the log likelihood is

$$\ell\left(\boldsymbol{y};\boldsymbol{\beta},\sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Z\beta})^\top(\boldsymbol{y}-\boldsymbol{Z\beta}). \tag{3.1}$$

To obtain the MLE's we must solve

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}} = \mathbf{0} \qquad \text{and} \qquad \frac{\partial\ell}{\partial\sigma^2} = 0.$$

However $\dfrac{\partial\ell}{\partial\boldsymbol{\beta}} = \mathbf{0}$ gives the normal equations (Eq. 2.1) and so the MLE for $\boldsymbol{\beta}$ must be $\widehat{\boldsymbol{\beta}}$ irrespective of the value of σ^2. □

SETTING

$$\frac{\partial\ell}{\partial\sigma^2} = 0$$

gives

$$\begin{aligned} 0 &= \frac{-n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^\top\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right) \\ \Rightarrow \widehat{\sigma}^2 &= \frac{1}{n}\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^\top\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right) \\ \text{i.e. the MLE } \widehat{\sigma}^2 &= \frac{1}{n}\times\text{Deviance} = \frac{(n-p)}{n}s^2. \end{aligned}$$

Result R3 showed that

$$s^2 = \frac{1}{n-p}\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)^\top\left(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)$$

is an unbiased estimator of σ^2. Therefore, the MLE for σ^2 must be a biased estimator:

$$\begin{aligned} \mathbb{E}[\widehat{\sigma}^2] &= \mathbb{E}\left(\frac{n-p}{n}s^2\right) \\ &= \frac{n-p}{n}\sigma^2 \\ &\neq \sigma^2. \end{aligned}$$

Consequently, instead of using the MLE to estimate σ^2, we usually use the estimator s^2 instead.

Theorem (R5). *Let* $\boldsymbol{y} \sim N_n\left(\boldsymbol{Z\beta}, \sigma^2 \boldsymbol{I}_n\right)$, *then*
$\widehat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1}\right)$,

Proof.

□

If σ^2 is known then we can use (R5) to provide basic confidence intervals for β. However, σ^2 is rarely known, and so we must estimate it. We then must use (R6) to find confidence intervals.

3.2 Distributional properties of $\hat{\beta}$ and s^2

Suppose $\boldsymbol{X} \sim N_n\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}\right)$ where $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix.

Lemma 2. $\boldsymbol{AX} \sim N_q\left(\boldsymbol{A\theta}, \boldsymbol{A\Sigma A}^\top\right)$, *where* $\boldsymbol{A}$ *is a* $q \times n$ *matrix.*

Proof. See G12PMM.

□

Matrix square roots are not unique. As $\boldsymbol{\Sigma}$ is symmetric, we can diagonalise it $\boldsymbol{\Sigma} = \boldsymbol{QDQ}^\top$ and use $\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{QD}^{-\frac{1}{2}}$

If $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\theta})^\top$, then $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, where $\Sigma^{-\frac{1}{2}}$ is the matrix square root of Σ^{-1}.

Lemma 3. $\left(\boldsymbol{X} - \boldsymbol{\theta}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X} - \boldsymbol{\theta}\right) \sim \chi^2_n$.

See Section 3.6 for a reminder of the definition of some common distributions.

Proof. Let $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\theta})^\top$. Then

$$\begin{aligned}\boldsymbol{Y}^\top \boldsymbol{Y} &= \left(\boldsymbol{X} - \boldsymbol{\theta}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{X} - \boldsymbol{\theta}\right) \\ &= \sum_{i=1}^{n} Y_i^2 \sim \sum_{i=1}^{n} N(0,1)^2 \sim \chi^2_n.\end{aligned}$$

□

Before proving (R6) we shall require

Theorem 6. *Let* $\boldsymbol{y} \sim N_n\left(\boldsymbol{Z\beta}, \sigma^2 \boldsymbol{I}_n\right)$, *then*

(i) $\frac{1}{\sigma^2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^\top \boldsymbol{Z}^\top \boldsymbol{Z}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2_p$,

(ii) $\dfrac{(n-p)\,s^2}{\sigma^2} \sim \chi^2_{n-p}$,

Tells us the distribution of the unbiased estimator of σ^2.

(iii) $\widehat{\boldsymbol{\beta}}$ *and* s^2 *are independent.*

Proof. (i) Note that

$$\frac{(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \mathbf{Z}^\top \mathbf{Z}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} = \left(\widehat{\boldsymbol{\beta}}-\mathbb{E}[\widehat{\boldsymbol{\beta}}]\right)^\top \left[\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})\right]^{-1} \left(\widehat{\boldsymbol{\beta}}-\mathbb{E}[\widehat{\boldsymbol{\beta}}]\right)$$

by (R1) and (R5). Lemma 3 then implies that the left hand side has a χ^2_p distribution.

□

(ii) and (iii) These proofs are not particularly difficult, but require several pages of algebra. They are included in the appendix to this chapter for completeness, but will **not be examined**.

□

3.3 Single parameter distributions

We are now in a position to prove (R6), namely that

$$\frac{\widehat{\beta}_i - \beta_i}{s\sqrt{d_{ii}}} \sim t_{n-p},$$

where d_{ii} is the ith diagonal element of $(\mathbf{Z}^\top \mathbf{Z})^{-1}$.

We shall prove the stronger result which gives the distribution of a linear function of the least squares estimator.

Proposition 2. *Let*

$$T = \frac{\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{a}^\top \boldsymbol{\beta}}{s\sqrt{\boldsymbol{a}^\top \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \boldsymbol{a}}}$$

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

Once we have proved the proposition, we then arrive at R6 by setting $\boldsymbol{a}^\top = (0, \ldots, 1, \ldots, 0)$, a vector of zeros with 1 in the ith position.

Proof.

$$\widehat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\right) \tag{3.2}$$

and

$$\frac{(n-p)\, s^2}{\sigma^2} \sim \chi^2_{n-p} \tag{3.3}$$

are independent (Thm 6).

Consider $\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}}$ where $\boldsymbol{a}$ is a p-vector of known constants. Then[1]

[1] Recall Lemma 2.

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

and so

$$U = \frac{\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{a}^\top \boldsymbol{\beta}}{\sigma\sqrt{\boldsymbol{a}^\top \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \boldsymbol{a}}} \sim N(0, 1).$$

From (3.3),

$$V = \frac{s}{\sigma} \sim \sqrt{\frac{1}{(n-p)}\chi^2_{n-p}}$$

where U and V are independent.

Hence

$$T = \frac{U}{V} \sim t_{n-p}$$

from the definition of the t-distribution, □

Finally, note that $\boldsymbol{a}^\top (Z^\top Z)^{-1} a = d_{ii}$ the ith diagonal element of $(Z^\top Z)^{-1}$.

3.4 Confidence intervals

We proved

$$\frac{\widehat{\beta}_i - \beta_i}{s\sqrt{d_{ii}}} \sim t_{n-p},$$

where d_{ii} is the ith diagonal element of $\left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1}$. This immediately gives a 100$(1-\alpha)$% confidence interval (C.I.) for β_i:

Note that $\text{std.error}(\widehat{\beta}_i) = s\sqrt{d_{ii}}$.

$$\widehat{\beta}_i \pm t_{n-p}\,(1-\alpha/2)\; s\sqrt{d_{ii}}$$

Using Proposition 2, we get the more general result that a 100$(1-\alpha)$% confidence interval for $\boldsymbol{a}^\top \boldsymbol{\beta}$ is given by

$$\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} \pm t_{n-p}(1-\tfrac{\alpha}{2}) s \sqrt{\boldsymbol{a}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{a}}.$$

3.4.1 A simple example

Data:

x	4.6	5.1	4.8	4.4	5.9	4.7	5.1	5.2	4.9	5.1
y	87.1	93.1	89.8	91.4	99.5	92.1	95.5	99.3	98.9	94.4

Fitting a simple linear regression model with R gives output:

```
> x <- c(4.6, 5.1, 4.8, 4.4, 5.9, 4.7, 5.1, 5.2, 4.9, 5.1)
> y <- c(87.1, 93.1, 89.8, 91.4, 99.5, 92.1, 95.5, 99.3, 98.9, 94.4)
> fit <- lm (y ~ x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-4.1966 -1.7792 -0.2677  1.3135  5.3823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   57.240     12.495   4.581   0.0018 **
x              7.404      2.501   2.960   0.0181 *
---
Signif. codes:  0 *** 0.001 **  0.01 * 0.05 . 0.1   1

Residual standard error: 3.1 on 8 degrees of freedom
Multiple R-squared: 0.5227,     Adjusted R-squared: 0.4631
F-statistic: 8.762 on 1 and 8 DF,  p-value: 0.01815
```

What is the fitted regression line?

Given that

$$(Z^\top Z)^{-1} = \begin{pmatrix} 16.25 & -3.24 \\ -3.24 & 0.65 \end{pmatrix}$$

Calculate a 95% equi-tailed confidence interval for the gradient?

You will need to use statistical tables

```
> confint(fit)
                2.5 %   97.5 %
(Intercept) 28.427314 86.05237
x            1.635831 13.17146
```

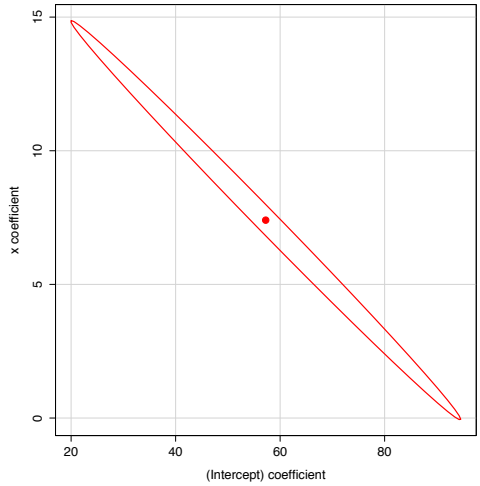


Figure 3.1: You can visualise joint confidence intervals using
`library(car)`
`confidenceEllipse(fit)`

3.5 Estimation and prediction

Imagine that we have fit a linear model and now wish to apply it to some new combination of the input variates, x_0. We can:

1. Estimate the expected value of the response for x_0,
2. Predict the value of a new observation with input variates x_0.

At first glance estimation and prediction appear to be the same - in both cases we get $z_0^\top \widehat{\boldsymbol{\beta}}$ as our estimate/prediction, where $z_0^\top$ is the row

of the design matrix that corresponds to the input variates $\boldsymbol{x}_0$. However, the key difference arises in the variability of the estimator/predictor.

3.5.1 Estimation

We want to estimate $\mathbb{E}[y_0]$: the expected value of the response with input variates $\boldsymbol{x}_0$. Our model says that $\mathbb{E}[y_0] = z_0^\top \boldsymbol{\beta}$, and by the Gauss-Markov Theorem, $z_0^\top \widehat{\boldsymbol{\beta}}$ is the unique unbiased linear estimator of $z_0^\top \boldsymbol{\beta}$ with minimum variance. From Section 3.3, a $100(1-\alpha)\%$ confidence interval for $\mathbb{E}[y_0]$ is given by

$$z_0^\top \widehat{\boldsymbol{\beta}} \pm t_{n-p}(1-\tfrac{\alpha}{2}) s \sqrt{z_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} z_0}.$$

That is, the variance of the estimation error is:

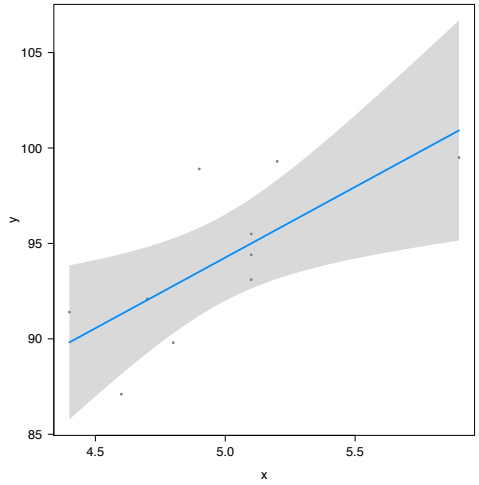


Figure 3.2: You can visualise the estimation intervals using
`library(visreg)`
`visreg(fit)`

3.5.2 Prediction

This time we want to estimate (predict) the actual value y_0 rather than its expectation $\mathbb{E}[y_0]$, and so our model is $y_0 = z_0^\top \boldsymbol{\beta} + \epsilon_0$, where $\epsilon_0 \sim N(0, \sigma^2)$ is independent of $\epsilon_1, \ldots, \epsilon_n$.

Although our prediction is still $z_0^\top \widehat{\boldsymbol{\beta}}$, the variance in the prediction error is

which is clearly strictly greater than the estimation error.
Now, the prediction error $y_0 - z_0^\top \widehat{\boldsymbol{\beta}}$ has a normal distribution, and so a $100(1-\alpha)\%$ predictive interval for y_0 is given by

$$z_0^\top \widehat{\boldsymbol{\beta}} \pm t_{n-p}(1-\tfrac{\alpha}{2}) s \sqrt{1 + z_0^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} z_0}.$$

3.5.3 Simple example ctd.

Suppose we are interested in the value of y at $x = 5$. Then the row of the design matrix for this x value would be $\boldsymbol{z}^\top =$

What is the estimated value of $\mathbb{E}[y]$ and the predicted value of y?

Calculate a 95% confidence interval for $\mathbb{E}[y]$

What is a 95% predictive interval for y

```
> newdata <- data.frame(x=5)
> predict(fit, newdata, interval = "confidence", level=0.95)  # Confidence interval for the mean respor
       fit      lwr      upr
1 94.25807 91.99462 96.52153
> predict(fit, newdata, interval = "prediction", level=0.95)  #  prediction interval of the response
       fit      lwr      upr
1 94.25807 86.75991 101.7562
```

The predictive interval is wider than the confidence interval for the expected value, as the predictive interval includes extra variation generated by the error for this observation, in addition to the variation from the random vector $\widehat{\boldsymbol{\beta}}$ which is present in both intervals.

3.6 Definitions of some common probability distributions

(1) If $X_i \overset{iid}{\sim} N(0,1)$ for $i = 1, \ldots, d$, then

$$Q = \sum_{i=1}^{d} X_i^2 \sim \chi_d^2.$$

(2) If $U \sim N(0,1)$ and $V \sim \chi^2_d$ where U and V are independent then

$$Y = \frac{U}{\sqrt{V/d}} \sim t_d.$$

(3) If $A \sim \chi^2_a$ and $B \sim \chi^2_b$ are independent then

$$F = \frac{A/a}{B/b} \sim F_{a,b}.$$

These relations *define* these three distributions.

3.7 Proof of Theorem 6 - non-examinable

We will now prove

$$\frac{(n-p)\,s^2}{\sigma^2} \sim \chi^2_{n-p}$$

First we need a lemma.

Lemma 4. *$\boldsymbol{I} - \boldsymbol{P}$ has $n-p$ eigenvalues of 1, and p eigenvalues of 0.*

Proof. Suppose $\boldsymbol{x}$ is an eigenvector with eigenvalue λ. Then

$$\lambda \boldsymbol{x}^\top \boldsymbol{x} = \boldsymbol{x}^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{x}^\top (\boldsymbol{I} - \boldsymbol{P})^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x} = \lambda^2 \boldsymbol{x}^\top \boldsymbol{x}$$

Thus $\lambda(\lambda - 1) = 0$, and so all the eigenvalues are 0 or 1.

$\boldsymbol{I} - \boldsymbol{P}$ is symmetric, so we can diagonalise it and write

$$\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$$

where $\boldsymbol{D}$ is a diagonal matrix of eigenvalues and $\boldsymbol{A}$ is an orthonormal matrix with $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}$.

We saw previously that $\mathrm{rank}(\boldsymbol{I} - \boldsymbol{P}) = n - p$ and thus

$$\mathrm{rank}(\boldsymbol{D}) = \mathrm{rank}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top) = \mathrm{rank}(\boldsymbol{I} - \boldsymbol{P}) = n - p.$$

As $\boldsymbol{D}$ is a diagonal matrix, it must thus contain exactly $n-p$ non-zero terms and p zero terms along its diagonal. □

Proof.

$$\begin{aligned}
\frac{(n-p)s^2}{\sigma^2} &= \frac{RSS}{\sigma^2} \\
RSS &= \widehat{\boldsymbol{\epsilon}}^\top \widehat{\boldsymbol{\epsilon}} \\
&= \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} \\
&= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^\top (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\
&= \boldsymbol{\epsilon}^\top (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\epsilon}
\end{aligned}$$

See Section 2.7 and the proof of Theorem 4

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

Recall that $\boldsymbol{I}-\boldsymbol{P}$ is symmetric. We can thus write $\boldsymbol{I}-\boldsymbol{P}=\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{\top}$.

$$\begin{aligned}\boldsymbol{\epsilon}^{\top}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\epsilon} &= \boldsymbol{\epsilon}\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{\top}\boldsymbol{\epsilon}\\ &= \boldsymbol{Z}^{\top}\boldsymbol{D}\boldsymbol{Z}\\ &= \sum_{i=1}^{n} d_i Z_i^2\\ &= \sum_{i=1}^{n-p} Z_i^2\end{aligned}$$

where $\boldsymbol{Z}=\boldsymbol{A}^{\top}\boldsymbol{\epsilon}$. Note that because $\boldsymbol{A}$ is orthogonal $\boldsymbol{Z}\sim N_n(\boldsymbol{0},\sigma^2\boldsymbol{I})$, and hence

$$\frac{(n-p)s^2}{\sigma^2}=\frac{\boldsymbol{\epsilon}^{\top}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\epsilon}}{\sigma^2}\sim\chi^2_{n-p}.$$

□

Lemma 5. *Suppose* $\boldsymbol{X}\sim N_n(\boldsymbol{\theta},\boldsymbol{\Sigma})$ *and let* $\boldsymbol{U}=\boldsymbol{A}\boldsymbol{X}$ *and* $\boldsymbol{V}=\boldsymbol{B}\boldsymbol{X}$. *If* $\mathbb{C}ov(\boldsymbol{U},\boldsymbol{V})=\boldsymbol{0}$, *then* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are independent.*

Moreover

(a) $\boldsymbol{U}^{\top}\boldsymbol{U}$ and $\boldsymbol{V}^{\top}\boldsymbol{V}$ are independent,

(b) $\boldsymbol{U}$ and $\boldsymbol{V}^{\top}\boldsymbol{V}$ are independent.

Proof. Let

$$\boldsymbol{W}=\begin{pmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{pmatrix}=\begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{pmatrix}\boldsymbol{X}$$

Thus $\boldsymbol{W}$ has a multivariate normal distribution with variance matrix

$$\mathbb{V}\mathrm{ar}(\boldsymbol{W})=\begin{pmatrix}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top} & \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\top}\\ \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^{\top} & \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^{\top}\end{pmatrix}$$

Thus, if

$$\mathbb{C}\mathrm{ov}(\boldsymbol{U},\boldsymbol{V})=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\top}=\boldsymbol{0}$$

the result follows. □

Finally, lets prove that $\widehat{\boldsymbol{\beta}}$ and s^2 are independent.

Proof. Let $\boldsymbol{U}=\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{V}=\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ then $\boldsymbol{U}$ and $\boldsymbol{V}^{\top}\boldsymbol{V}$ are independent by Lemma 5 since

Note that $\boldsymbol{U}=\boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{V}=\boldsymbol{B}\boldsymbol{y}$

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[\widehat{\boldsymbol{\beta}},\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right] &= (\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\mathbb{C}\mathrm{ov}\left(\boldsymbol{y},(\boldsymbol{I}_n-\boldsymbol{P})\boldsymbol{y}\right)\\ &= \sigma^2(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}(\boldsymbol{I}_n-\boldsymbol{P})^{\top}\\ &= \sigma^2((\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}-(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top})\\ &= \boldsymbol{0}_{p\times n}.\end{aligned}$$

Hence, $\widehat{\boldsymbol{\beta}}$ and s^2 are independent.

□

4

Hypothesis testing

4.1 Hypothesis testing reminder

Suppose that we have a null hypothesis H_0 represented by a completely specified model and that we wish to test this hypothesis using data $X_1, \ldots, X_n$. We proceed as follows

1. Assume H_0 is true.
2. Find a test statistic $T(X_1, \ldots, X_n)$ for which large values indicate departure from H_0.
3. Calculate the theoretical sampling distribution of T under H_0.
4. The observed value $T_{obs} = T(x_1, \ldots, x_n)$ of the test statistic is compared with the distribution of T under H_0.
   - Using the Neyman-Pearson approach we reject H_0 if $T_{obs} > c$. Here c is chosen so that $\mathbb{P}(T \geq c | H_0) = \alpha$ where α is the size of the test, i.e., $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) = \alpha$.
   - Under the Fisherian approach we compute the p-value $p = \mathbb{P}(T \geq T_{obs} | H_0)$ and report it. This represents the strength of evidence against H_0.

4.1.1 Simple regression hypothesis tests

Consider testing

$$\begin{aligned} & H_0 : \beta_i = 0 \\ \text{vs} \quad & H_1 : \beta_i \neq 0 \end{aligned}$$

at the 100α% level. The natural test statistic is $T = \frac{\widehat{\beta}_i}{\text{std.error}(\widehat{\beta}_i)}$, and under the null hypothesis we know the distribution of T:

$$T = \frac{\widehat{\beta}_i}{\text{std.error}(\widehat{\beta}_i)} = \frac{\widehat{\beta}_i}{s\sqrt{d_{ii}}} \sim t_{n-p}.$$

We reject H_0 if $|T_{obs}| > t_{n-p}(1 - \frac{\alpha}{2})$, or if the p-value is less than α.

R reports the p-value of the test.

I've included a table of quantiles of the t-distribution in the appendix.

4.1.2 Simple example from Ch3 continued

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1966 -1.7792 -0.2677  1.3135  5.3823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   57.240     12.495   4.581   0.0018 **
x              7.404      2.501   2.960   0.0181 *
---
Signif. codes:  0 *** 0.001 **  0.01 * 0.05 . 0.1   1

Residual standard error: 3.1 on 8 degrees of freedom
Multiple R-squared: 0.5227,     Adjusted R-squared: 0.4631
F-statistic: 8.762 on 1 and 8 DF,  p-value: 0.01815
```

(a) Is the intercept of the fitted line equal to zero?

i.e. test $H_0 \;:\; \underset{\text{intercept}}{a = 0}$ vs $H_1 : a \neq 0$

(b) Consider the quadratic model $y_i = a + bx_i + cx_i^2 + \epsilon_i$,

The quadratic regression model can be fit using R:

```
> fit2 <- lm(y~x+I(x^2))     # Note the use of I(x^2) to separate terms
> summary(fit2)

Call:
```

```
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.8841 -2.0468 -0.1909  1.9633  5.0651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.157    127.851  -0.095    0.927
x             34.550     49.819   0.694    0.510
I(x^2)        -2.637      4.832  -0.546    0.602

Residual standard error: 3.246 on 7 degrees of freedom
Multiple R-squared: 0.5422,     Adjusted R-squared: 0.4114
F-statistic: 4.145 on 2 and 7 DF,  p-value: 0.06492
```

Is c significantly different from zero?

i.e. is the quadratic term worth including in the model?

Hypotheses: $H_0 : c = 0 \quad H_1 : c \neq 0.$

4.1.3 Important note:

These T statistics are not independent. In the quadratic model none of the parameters are significant, but once γ is removed from the model, the intercept and the slope become significant.

4.2 F-test for the General Linear Hypothesis

Consider the normal linear model

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Z}$ is $n \times p$ and $\boldsymbol{\epsilon} \sim N_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right)$.

Suppose we want to test — the so called general linear hypothesis

$$H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$$

vs

$$H_1 : \boldsymbol{\beta} \text{ general},$$

where $\boldsymbol{A}$ is a $q \times p$ matrix of rank q and $\boldsymbol{c}$ is a q-vector.

Example restrictions:

- $A = (1\ 0\ 0\ \ldots\ 0)$ and $c = 0$, gives $H_0 : \beta_1 = 0$.
- $A = (0\ 1\ 1\ \ldots\ 1)$ and $c = 0$, gives $H_0 : \beta_2 = \beta_3 = \ldots = 0$.[1]
- $A = (1\ 1\ 1\ \ldots\ 1)$ and $c = 2$, gives $H_0 : \sum \beta_i = 2$.

[1] So we would be testing whether M_1 is an improvement over the null model

We can calculate the least squares estimator of $\boldsymbol{\beta}$ under any linear constraint.

Lemma 6. *Under $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ the least squares estimator of $\boldsymbol{\beta}$ is:*

$$\widehat{\boldsymbol{\beta}}_H = \widehat{\boldsymbol{\beta}} + \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\left[\boldsymbol{A}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\left(\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right).$$

Proof. We will use Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_q)^T$. The objective is to minimise $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.

$$\begin{aligned} g(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \left(\boldsymbol{\beta}^T\boldsymbol{A}^T - \boldsymbol{c}^T\right)\boldsymbol{\lambda} \\ \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{Z}^T\boldsymbol{y} + 2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)\boldsymbol{\beta} + \boldsymbol{A}^T\boldsymbol{\lambda} \end{aligned}$$

At the point $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_H$, $\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$. Solving this gives

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_H &= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y} - \frac{1}{2}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\boldsymbol{\lambda} \\ &= \widehat{\boldsymbol{\beta}} - \frac{1}{2}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\boldsymbol{\lambda}, \end{aligned}$$

however $\boldsymbol{A}\widehat{\boldsymbol{\beta}}_H = \boldsymbol{c}$, and so

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{A}\widehat{\boldsymbol{\beta}} - \frac{1}{2}\boldsymbol{A}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\boldsymbol{\lambda} \\ \Rightarrow -\frac{1}{2}\boldsymbol{\lambda} &= \underbrace{\left[\boldsymbol{A}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\right]^{-1}}_{q \times q \text{ with rank } q}\left(\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right). \end{aligned}$$

Substituting $\boldsymbol{\lambda}$ into the equation for $\widehat{\boldsymbol{\beta}}_H$ completes the proof of the lemma. □

Theorem 7. General Linear Hypothesis Test

Let $D_1 = (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})$ *be the deviance of the larger model.*
Let $D_0 = (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_H)^T(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_H)$ *be the deviance under* H_0.

Then:

(1) $D_0 - D_1 = \left(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)^T \left[\boldsymbol{A}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\left(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)$

(R7) *(2)* $F = \dfrac{(D_0 - D_1)/q}{D_1/(n-p)} \sim F_{q,n-p}.$

Proof. Proof of (1) follows by substituting the formula for $\widehat{\boldsymbol{\beta}}_H$ derived in the lemma (Exercise).

To prove (2) first note that from (1),

$$\frac{D_0 - D_1}{\sigma^2} = \left(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)^T \left(\boldsymbol{A}\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^T\right)^{-1}\left(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right).$$

Under H_0, $\boldsymbol{A}\widehat{\boldsymbol{\beta}} \sim N_q\left(\boldsymbol{c}, \boldsymbol{A}\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^T\right)$ from Lemma 2. Thus, by Lemma 3

$$\frac{D_0 - D_1}{\sigma^2} \sim \chi^2_q.$$

Recall from Theorem 6 part (ii)

$$\frac{(n-p)\,s^2}{\sigma^2} = \frac{D_1}{\sigma^2} \sim \chi^2_{n-p}.$$

Hence from the definition of an F-distribution,

$$\begin{aligned}\frac{(D_0 - D_1)/(\sigma^2 q)}{D_1/\sigma^2(n-p)} &= \frac{(D_0 - D_1)/q}{D_1/(n-p)} \\ &\sim \frac{\chi^2_q/q}{\chi^2_{n-p}/(n-p)} \sim F_{q,n-p}.\end{aligned}$$

as required. □

4.2.1 Comments:

- The model under H_0 has $p - q$ parameters (p parameters under H_1 with q restrictions/constraints).
- When a smaller model can be expressed as a simplification of a larger model by setting $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$, then the smaller model is said to be *nested* within the larger model.
- The F-test with statistic

$$F = \frac{(D_0 - D_1)/q}{D_1/(n-p)} \qquad \begin{array}{l} D_0 \text{ – deviance of smaller model} \\ D_1 \text{ – deviance of larger model}\end{array}$$

is only appropriate for comparing *nested* models.

4.3 F test for the existence of regression

We want to test for the existence of regression. In other words, is there statistical evidence that supports the use of the linear model $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{Z\beta}$ over the use of the null model $\mathbb{E}(\boldsymbol{y}) = \mathbf{1}_{n\times 1}\beta_0$?

That is, we want to compare the models

$$\begin{aligned} \text{M0:} \quad \boldsymbol{y} &= \beta_0 \mathbf{1}_{n\times 1} + \boldsymbol{\epsilon} \\ \text{M1:} \quad \boldsymbol{y} &= \boldsymbol{Z\beta} + \boldsymbol{\epsilon}. \end{aligned}$$

If we write $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}^* \end{bmatrix}$ then the hypotheses for the test for the existence of regression become

H_0 : M0 applies $(\boldsymbol{\beta}^* = \mathbf{0})$,

H_1 : M1 applies ($\boldsymbol{\beta}^*$need not equal $\mathbf{0}$).

This test is called testing for the existence of regression

Under the normality assumptions $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ we can apply the result (R7) from Theorem 7. We can see that the constraint in H_0 can be considered to be

$$A\boldsymbol{\beta} = c$$

where

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \text{ is } (p-1)\times p \qquad c = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ is } (p-1)\times 1$$

Then

$$\begin{aligned} D_1 &= (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) \\ &= \sum (y_i - \hat{y}_i)^2 = \text{ResidSS} \\ D_0 &= (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_H)^T(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_H) \\ &= \sum (y_i - \bar{y})^2 = \text{TotalSS} \end{aligned}$$

By Theorem 2, we can see that

$$D_0 - D_1 = \sum (\hat{y}_i - \bar{y})^2 = \text{RegrSS}.$$

Thus, by results (R7) in Theorem 7 We have

$$\begin{aligned} F &= \frac{(D_0 - D_1)/(p-1)}{D_1/(n-p)} \\ &= \frac{\text{RegrSS}/(p-1)}{\text{ResidSS}/(n-p)} \\ &= \frac{\text{RegrMS}}{\text{ResidMS}} \end{aligned}$$

RegrMS is the mean regression sum-of-squares taking into account the degrees of freedom, i.e.,

$$\text{RegrMS} = \frac{\text{RegrSS}}{p-1},$$

and similarly

$$\text{ResidMS} = \frac{\text{ResidSS}}{n-p}.$$

and $F \sim F_{p-1,n-p}$ by (R7).

So we reject H_0 at the 100α% level if $F > F_{p-1,n-p}(1-\alpha)$.

4.3.1 ANOVA table

An ANOVA table is a handy way of presenting this information:

Source	d.f.	SS	MS	F
Regression	$p-1$	$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$	$\frac{\text{RegrSS}}{p-1}$	$\frac{\text{RegrMS}}{\text{ResidMS}}$
Residual	$n-p$	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$\frac{\text{ResidSS}}{n-p}$	
Total	$n-1$	$\sum_{i=1}^{n}(y_i - \overline{y})^2$		

The R command for finding the ANOVA table is `anova(fit)`.

Consider the simple toy example from Ch3.

```
> x <- c(4.6, 5.1, 4.8, 4.4, 5.9, 4.7, 5.1, 5.2, 4.9, 5.1)
> y <- c(87.1, 93.1, 89.8, 91.4, 99.5, 92.1, 95.5, 99.3, 98.9, 94.4)
> fit <- lm (y ~ x)
> deviance(fit)
[1] 76.87474
> deviance(fitNull)
[1] 161.069
```

Calculate the F statistic:

What do you conclude?

F-tables are provided in the appendix.

```
> summary(fit)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.1966 -1.7792 -0.2677  1.3135  5.3823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   57.240     12.495   4.581   0.0018 **
x              7.404      2.501   2.960   0.0181 *
---
Signif. codes:  0 *** 0.001 **  0.01 * 0.05 . 0.1   1

Residual standard error: 3.1 on 8 degrees of freedom
Multiple R-squared:  0.5227,Adjusted R-squared:  0.4631
F-statistic: 8.762 on 1 and 8 DF,  p-value: 0.01815

> anova(fit)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 84.194  84.194  8.7617 0.01815 *
Residuals  8 76.875   9.609
---
Signif. codes:  0 *** 0.001 **  0.01 * 0.05 . 0.1   1
```

4.4 F-tests for comparing nested models

Now consider testing two nested models. Consider partitioning the model

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

into two parts:

$$\mathbf{Z} = (\mathbf{Z}_A, \mathbf{Z}_B) \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B \end{bmatrix}$$

where

$\boldsymbol{\beta}_A$ is $(p-q) \times 1$, $\boldsymbol{Z}_A$ is $n \times (p-q)$

$\boldsymbol{\beta}_B$ is $q \times 1$, $\boldsymbol{Z}_B$ is $n \times q$

This gives us two models

- Reduced model A (includes intercept):

$$\boldsymbol{y} = \boldsymbol{Z}_A\boldsymbol{\beta}_A + \boldsymbol{\epsilon}$$

- Full model B:

$$\boldsymbol{y} = \boldsymbol{Z}_A\boldsymbol{\beta}_A + \boldsymbol{Z}_B\boldsymbol{\beta}_B + \boldsymbol{\epsilon}$$

Note that model A is nested in model B, and that we need to fix q parameters to reduce B to A.

We want to test

$$H_0 : \boldsymbol{\beta}_B = \mathbf{0} \qquad \text{vs} \qquad H_1 : \boldsymbol{\beta}_B \text{ arbitrary.}$$

Result R7 says that for nested linear models A and B,

$$\frac{(D_A - D_B)/q}{D_B/(n-p)} \sim F_{q,n-p} \text{ under } H_0$$

This is how we nearly always think of R7, rather than counting rows of a constraint matrix.

where q is the number of parameters we need to constrain to get from model B to model A. We reject $H_0 : \boldsymbol{\beta}_B = \mathbf{0}$ in favour of $H_1 : \boldsymbol{\beta}_B \neq \mathbf{0}$ if F is larger than $F_{q,n-p}(1-\alpha)$ for a 100α% level test.

The command for this test in R is `anova(fitA, fitB)` which gives us a table of the form

	Resid d.f.	RSS	Df	Sum of Sq	F	$\mathbb{P}(>F)$
Model A	$n-p+q$	D_A				
Model B	$n-p$	D_B	q	$D_A - D_B$	$\frac{(D_A-D_B)/q}{D_B/(n-p)}$	$\mathbb{P}(F^{obs} > F_{q,n-p})$

Returning to the simple example, lets see if the quadratic model is a significant improvement over the straight line.

```
> fit2 <- lm(y~x+I(x^2))
> deviance(fit2)
[1] 73.73828
```

Write down the two hypotheses, and carry out an F-test

```
> anova(fit, fit2)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 76.875
2      7 73.738  1    3.1365 0.2977 0.6022
```

4.5 A worked ANOVA example

Consider an agricultural experiment to determine the effect of 6 different fertilisers on crop yield

Trial	Fert A	Fert B	Fert C	Fert D	Fert E	Fert F	
1	14.5	13.5	11.5	13	15	12.5	
2	12	10	11	13	12	13.5	
3	9	9	14	13.5	8	14	
4	6.5	8.5	10	7.5	7	8	
$\sum_j$	42.0	41.0	46.5	47.0	42.0	48.0	$\sum_{ij} = 266.5$

The response is the yield, and the covariate is which fertiliser was used, which is a discrete factor with 6 levels.

Let y_{ij} be the yield from the j^{th} trial using the i^{th} fertiliser. Then an appropriate model to test whether the fertiliser used affects yield would be

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{4.2}$$

and we would test

$$H_0 : \mu_i = \mu \ \forall \ i \qquad \text{vs} \qquad H_1 : \ \ \mu_i \text{ arbitrary}$$

Testing of this form is known as one-way analysis of variance, as we have a single discrete factor. To carry out an F-test, we first need to calculate the parameters under both models. It is very easy to see that under H_0

$$\hat{\mu} = \overline{y}_{\cdot\cdot} = \frac{1}{24}\sum_{ij} y_{ij}$$

and under H_1

$$\hat{\mu}_i = \overline{y}_{i\cdot} = \frac{1}{4}\sum_{j=1}^{4} y_{ij}$$

Other parameterisations such as

$$y_{ij} = \begin{cases} \mu + \epsilon_{ij} \text{ if } i = A \\ \mu + \alpha_i + \epsilon_{ij} \text{ otherwise} \end{cases} \tag{4.1}$$

are also used, and we would test if $\alpha_i = 0 \ \forall \ i$. This is simply a linear reparameterization of Equation 4.2 and so it doesn't matter which we use.

The R command lm(yield $\sim$ fert) would fit (4.1), whereas lm(yield $\sim$ fert-1) would fit model (4.2).

Next we must calculate the two deviances

$$D_0 = \sum (y_{ij} - \bar{y}_{\cdot\cdot})^2$$

$$D_1 = \sum (y_{ij} - \bar{y}_{i\cdot})^2$$

Finally, we can calculate

$$F = \frac{(D_0 - D_1)/q}{D_1/(n-p)}$$

and compare this with a $F_{q,n-p}$ random variable.

5

Model validation and improvement

Before fitting a model, we should plot the data in an exploratory data analysis. This will help suggest sensible models and highlight difficult data points. **After** fitting a model, regression diagnostics should be used to check whether our modelling assumptions are valid.

- Is the assumed mean function $\boldsymbol{Z\beta}$ a good choice?
- Are the errors normally distributed?
- Do they have constant variance?
- Are any of the observations wrong? (outliers)
- Are any of the data points more influential on the model fit than others? (high leverage)

We've seen the use of R^2 and adjusted-R^2 for assessing model fit, but using a single numerical summary can be misleading. Francis Anscombe constructed 4 datasets to warn about the use of simple statistics such as R^2. They illustrate the importance of visually examining the data before assuming a particular type of relationship.

See the case study!

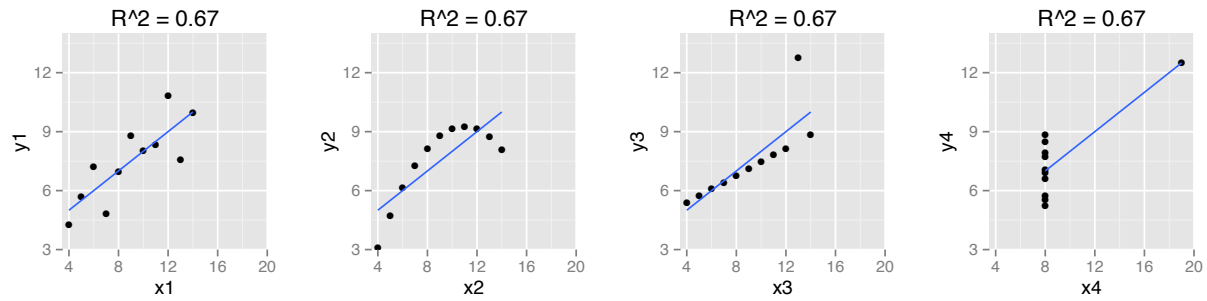


Figure 5.1: Each of these datasets has the same mean (of x and y), variance (of x and y), line of best fit ($y = 3 + 0.5x$) and R^2.

If we find the modelling assumptions are violated, then there are various approaches for improving the model, including transformations, adding higher order modelling terms, and weighted least squares. These diagnostic and corrective techniques can greatly extend the practical application of linear models. Careful investigation of data and model is often the difference between a crude mechanical data

analysis[1] and a careful nuanced analysis that leads to meaningful interpretations and conclusions. This is a big topic, and we are only going to scratch the surface of available techniques.

[1] See `www.automaticstatistician.com` - what is 'crude and automated' is improving all the time!

5.1 Remedies: transformations and weighted least squares

5.1.1 Transformations

By transforming either X or Y, we can often find better model fits. A particularly useful family of transformations are the power transformations

$$X \to X^p$$

where we usually consider values of p between -2 and 3. Another useful transformation is

$$X \to \log X$$

which we informally consider corresponds to $p = 0$.

The Box-Cox family of transformations is

$$X \to \frac{X^p - 1}{p}$$

which has the benefit a preserving the direction of X (which is reversed when p is negative), and for which $\lim_{p \downarrow 0} \frac{X^p - 1}{p} = \log X$. However, in practice it is usually simpler to just work with raw powers instead.

- If X contains negative values, we can instead use the transformation

  $$X \to (X + s)^p$$

  where s is called a start.
- The ratio of smallest X to largest X is less than about 5, a power transformation will not have much effect. We can shift the data towards zero by using a negative value of s so that the transformation does have some effect.

Transforming skewness

Skewed distributions can cause problems as many of the values tend to be clustered together. This can make some observations in the long tail wrongly appear to be outliers, and can hide outliers in the body of the distribution. The effect of the power transformations is to spread out either high or low values and can rectify skewness. Descending the ladder of powers towards $\log X$ can correct a positive skew by pulling in the right tail, and conversely, ascending the ladder of powers towards X^3 can correct a negative skew.

Transforming non-linearity

Transformations of X and Y can be used to make simple[2] monotone non-linear relationships between X and Y appear more linear.

[2] Simple here means the direction of curvature doesn't change.

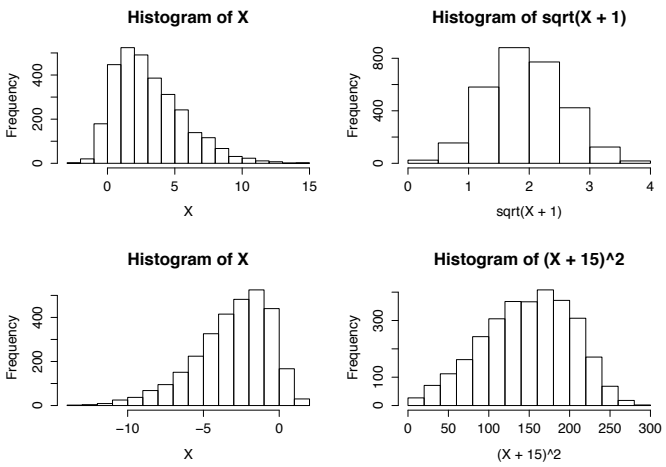


Figure 5.2: Top row: A positive skew can be removed by a power transformation with $p < 1$. Bottom row: a negative skew can often be removed with a power transformation with $p > 1$.

Mosteller and Tukey's bulging rule can be used to suggest which way we need to transform either X or Y up or down the ladder of powers in order to correct a non-linear relationship. Consider the following UN data on infant mortality rates in 207 countries.

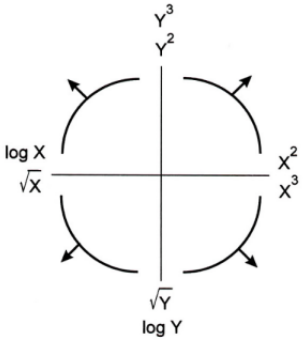


Figure 5.3: Mosteller and Tukey's bulging rule

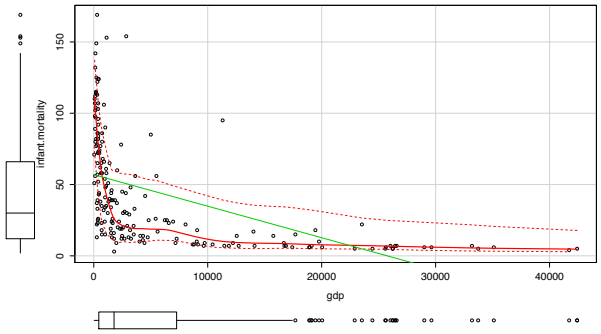


Figure 5.4: Infant mortality rate per 1000 live births versus gdp in US dollars for the 207 countries in the UN.

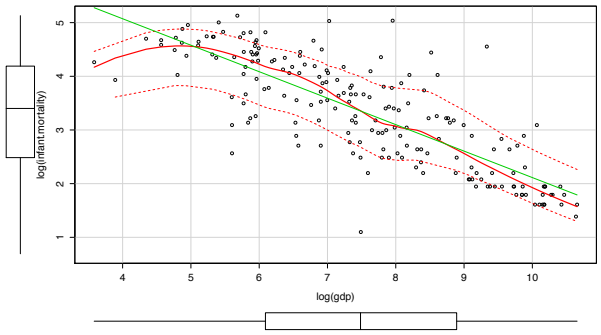


Figure 5.5: Log Infant mortality rate versus log(gdp). Notice how the transformations has corrected the non-linear relationship and corrected the skew in the two distributions.

THE CHOICE OF TRANSFORMATION can be made on a trial-and-error basis, and we usually only try a small number of powers such as $p = -1, -\frac{1}{2}, \frac{1}{2}, 1, \frac{3}{2}, 2, 3$ and $\log X$ and eyeball the result to see which performs best. Generally we prefer interpretable transformations.

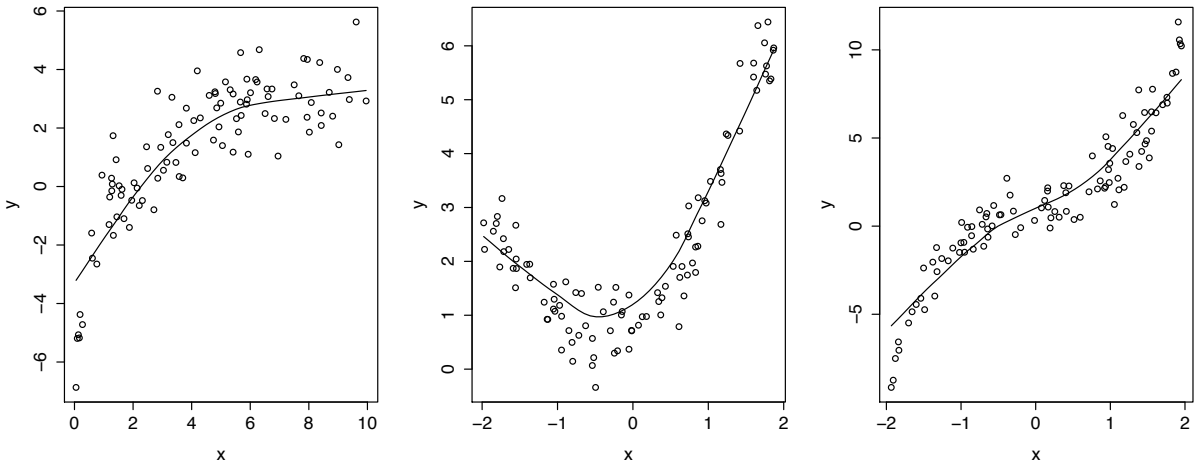
Transforming non-constant variance

A common violation of the linear regression assumptions, is to find that the variance of the random error depends on the value of y, with larger values of y having higher variance than smaller values (i.e. heteroscedastic errors). We can often transform to constant variance by transforming down Y down the ladder of powers towards $\log(y)$.

5.1.2 Including higher order terms

Although we can often correct simple monotone non-linear relationships by transforming either X or Y, we cannot correct complex relationships or non-monotone relationships. In this case, it can be useful to include higher order modelling terms such as X^2, X^3, $\log X$ etc.

Below are three plots showing a simple monotone relationship that can be corrected by transforming either x or y, a simple non-monotone relationship that can't be corrected by transformation, and a complex monotone relationship that can't be corrected by transformation.



5.1.3 Weighted least squares estimation

Occasionally we may wish to relax the assumption that $\mathbb{V}\mathrm{ar}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$ and consider the more general model in which

$$\mathbb{V}\mathrm{ar}\,(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$$

where $\boldsymbol{V}$ is assumed to be known. However, as we will see, any weighted least squares model can be transformed to an ordinary least squares model[3].

$\boldsymbol{V}$ is symmetric positive definite $\Rightarrow \boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}^T$, where $\boldsymbol{R}$ is a $n \times n$ square root matrix[4].

[3] Which is why we focussed on OLS throughout this module.

[4] Matrix square roots are not uniquely defined. Recall that we can diagonalise $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^\top$, and so we could take $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{D}^{'\frac{1}{2}}$ as the matrix square root.

Consider the model $\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$, where $E[\boldsymbol{\epsilon}] = \boldsymbol{0}$ and $\mathbb{V}\mathrm{ar}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$. If we pre-multiply by $\boldsymbol{R}^{-1}$ we have

$$\boldsymbol{R}^{-1}\boldsymbol{y} = \boldsymbol{R}^{-1}\boldsymbol{Z\beta} + \boldsymbol{R}^{-1}\boldsymbol{\epsilon}$$

i.e. $\boldsymbol{y}' = \boldsymbol{Z}'\boldsymbol{\beta} + \boldsymbol{\epsilon}'$, where

$$\begin{aligned} \boldsymbol{y}' &= \boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}' &= \boldsymbol{R}^{-1}\boldsymbol{Z} \end{aligned}$$

and

Therefore the model $\boldsymbol{y}' = \boldsymbol{Z}'\boldsymbol{\beta} + \boldsymbol{\epsilon}'$ is an ordinary least squares regression model. Thus, the least squares estimator for $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}$:

This is called the generalised least squares estimator.

Properties:

$$\begin{aligned} E\left[\widehat{\boldsymbol{\beta}}\right] &= \boldsymbol{\beta} \quad \text{(unbiased)} \\ \mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}\right) &= \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1}. \end{aligned}$$

5.2 Diagnostic plots

When we constructed the model, we made some assumptions about the random errors.

1. uncorrelated,

2. have equal variance,
3. have zero mean.

   In addition, we also assumed

4. The errors are normally distributed.

when carrying out hypothesis tests. Once we have fitted the model we can examine the residuals to see if these assumptions were acceptable or not.

5.2.1 Residual plots

Proposition 3. *If the model is true, then*

$$\mathbb{C}ov(\widehat{\epsilon}_i, \widehat{y}_i) = 0.$$

Proof.

□

We get a visual indication of whether these assumptions are true by examining a plot of the residuals $\widehat{\epsilon}_i$ against the fitted values $\widehat{y}_i$.

- If the regression model is correct, then the residual plots should look like null plots.
- If the variance depends on the fitted value (which is not uncommon, especially increasing variance as $\widehat{y}$ increases), this will show up as a funnel or megaphone shape in the residual plot.

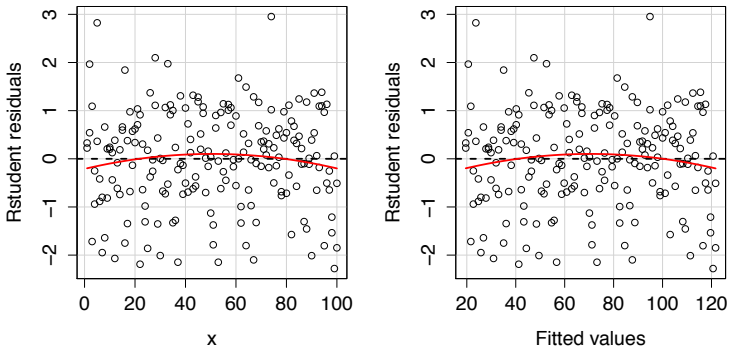


Figure 5.6: An acceptable residual plot is a null plot - a band of points with no discernible trend between the residual and the fitted value or between the residual and the covariates.

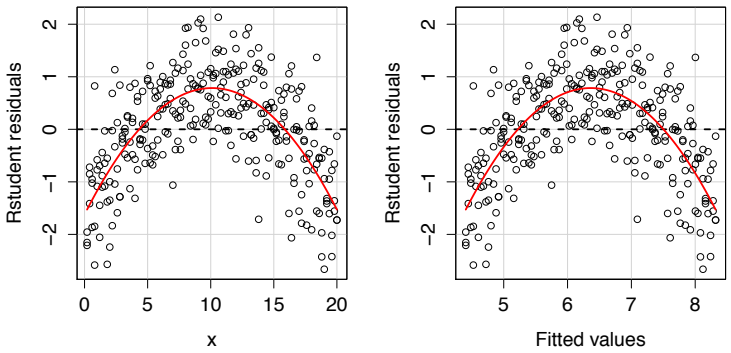


Figure 5.7: There is correlation in the residuals, but the variance seems to be constant. Because the trend is non-monotonic we cannot use a simple transformation to correct this, so we'd need to fit a more complex model, such as a quadratic model.

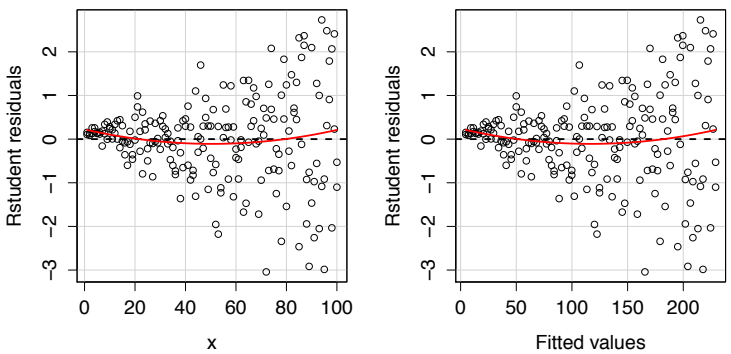


Figure 5.8: A 'right-opening megaphone' residual plot. The variance is not constant between observations. The best way to correct this is with a transformation of the y variable. In this example $y' = \log(y)$ is the best transformation. These plots are very common in practice. For example when measuring lengths or distances: short distances can be measured very accurately, but longer distances are harder to get quite so accurate.

For multiple linear regression it can also be informative to plot the residuals against each of the input variables. There should be no dependence left between the residuals and the input variables. If there is, then you may need to transform the input variable or include high order terms such as quadratic terms.

Note that if $dim(x) > 1$ then if will be hard to spot the problem from a scatter-plot matrix, but that the problem may still appear in the residuals plot.

5.2.2 Checking normality: QQ-plots

We have assumed $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$. A QQ-plot can be used to check this assumption and to spot outliers. A QQ-plot plots quantiles from one distribution against quantiles from another. If these quantiles match, the distributions are the same.

Fortunately, the F-test is quite robust to departures from normality, i.e. the test results are only moderately affected by a broad class of departures from normality. However, we still should examine our statistics for normality.

Note that for normal linear models, If the normality assumption holds, then the externally Studentised residuals t_i follow the t-distribution with $n - p - 1$ degrees of freedom. Let $d^{(1)} \ldots d^{(n)}$ be the *rank ordered* standardised residuals, i.e. $d^{(1)} = \min(t_1, t_2, \ldots, t_n)$, $d^{(n)} = \max(t_1, t_2, \ldots, t_n)$. Let $D^{(i)} = E[d^{(i)}]$, the expected values of the $d^{(i)}$'s, where the expectations are obtained under the t_{n-p-1} distribution. The $D^{(i)}$'s are called the *normal scores*. A QQ-plot of $d^{(i)}$ vs $D^{(i)}$ can then be used to check the validity of the normal assumption.

And become increasingly similar to the standard normal distribution as $n - p$ increases

Or if $n - p$ is large we can just use a standard normal distribution.

If the data approximately match the straight line representing the reference t-distribution, then the assumption that the data come from the normal distribution is validated.

If the data depart from the straight line (there will always be some natural variability), then the assumption of normality is called into question.

You can use the qqPlot command in the car package to plot QQ-plots in R.

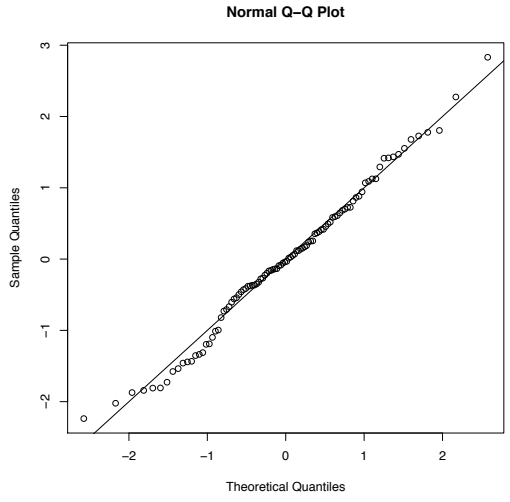


Figure 5.9: This QQ plot looks good - the points mostly fall on or near the diagonal line. There is no evidence to suggest that the residuals are not normally distributed.

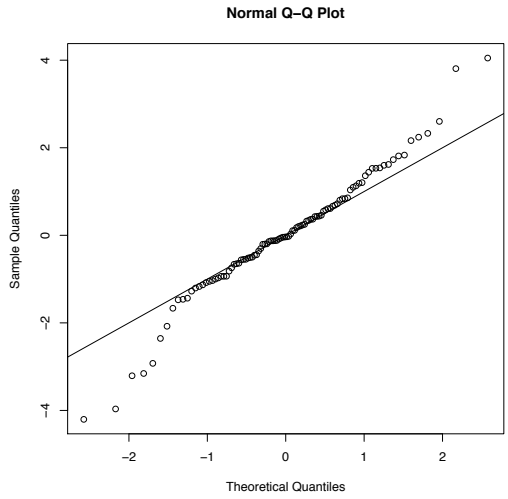


Figure 5.10: This plot suggests the normal distribution is not a good fit to the data. In particular, we can see that the tails of the data are heavier than the tails of the normal distribution. We could try to correct this with a transformation of the data, such as taking the logarithm of the response.

5.2.3 Component-Plus-Residual Plots

For simple linear regression (i.e., a single covariate), a scatter plot of y vs x shows the shape the model should take. For multiple linear regression (several covariates), then a scatter plot matrix can fail to illustrate the correct form of the model, as the effect of the other covariates is hidden. Component-plus-residual plots, also called partial-residual plots, can be useful for illustrating the relationship between y and each of the covariates.

Although they don't always work. See the Fox text book for an in-depth discussion.

Define the partial residual for the j^{th} explanatory variable to be

$$\widehat{\epsilon}_i^{(j)} = \widehat{\epsilon}_i + \beta_j Z_{ij}$$

This adds back the linear component of the partial relationship between Y and X_j, which may include an unmodeled nonlinear component. We then plot $\widehat{\epsilon}^{(j)}$ vs X_j. The slope of this curve will be β_j (by construction), but it may also highlight non-linear relationships as well. By adding a non-parametric smoother over the top, as well as the line of best fit, we can get an idea of how to improve the model by seeing how the smooth departs from the straight line.

See the case study for an example of their use.

5.3 Unusual and influential data

We now turn our attention to problems with the data itself. Unusual data are problematic as they can unduly influence the results of any analysis.

Definition 9. *An* **outlier** *is an observation y_i that is not near its fitted value $\widehat{y}_i$, i.e. an observation with an unusually large residual.*

Definition 10. High leverage *points, are observations (x_i, y_i) that have a large effect on the fitted regression model.*

Typically, high leverage points are observations with covariates x_i far from the other observations. Although we could say that x_i are outliers from the rest of the data, we only use the term outlier for when y_i differs from $\widehat{y}_i$.

Beware that some books give slightly different definitions.

If an outlier is also a high-leverage point, then it will greatly influence the model fit, which might mean that we are not able to detect that it is an outlier. For this reason, it is important to examine closely any high leverage points.

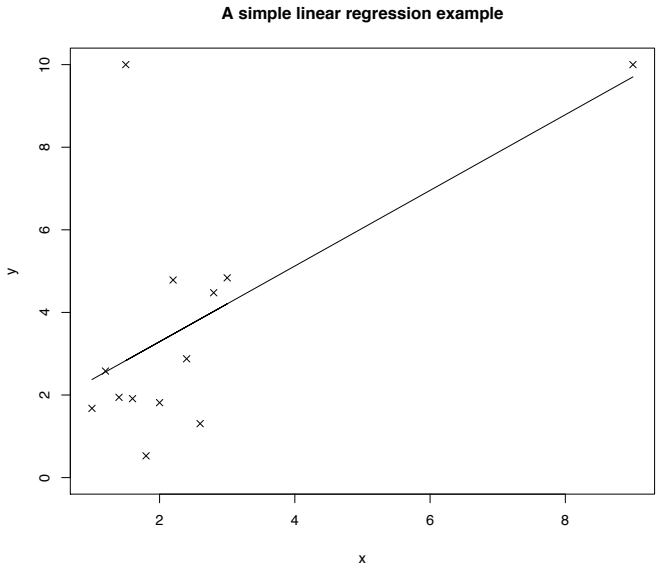


Figure 5.11: Which points are outliers, and which are high leverage points?

Points that are outliers and have high leverage have high **influence** on the regression coefficients - it is these points that we need to be particularly careful about.

See the Davis data in the case study.

5.3.1 Assessing Leverage

Recall that

$$\begin{aligned}\widehat{\boldsymbol{y}} &= \boldsymbol{Z}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{P}\boldsymbol{y}\end{aligned}$$

and so $\widehat{y}_i = p_{ii}y_i + \sum_{j\neq i} p_{ij}y_j$, where p_{ii} is the i^{th} hat-value.

Definition 11. *The* i^{th} **leverage** *is defined to be* p_{ii}

Use the R command `hatvalues(fit)` to find the leverages.

It determines the effect of the ith observation on the ith fitted value.

Proposition 4. *The variance of the* i^{th} *residual is*

$$\mathbb{V}ar(\widehat{\epsilon}_i) = \sigma^2\left[(\boldsymbol{I}_n - \boldsymbol{P})\right]_{ii} = \sigma^2(1 - p_{ii}).$$

It can also be shown that $\frac{1}{n} \leqslant p_{ii} \leqslant 1$ (exercise). Therefore as p_{ii} approaches 1 then the variance of the i^{th} residual tends to zero, and so

whatever the value of y_i, the fitted line will go through it.

Proof. We saw in Section 2.7 that

$$\widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{y}$$

where $\boldsymbol{P} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ is the hat-matrix . The variance-covariance matrix of the residuals is given by

since $(\boldsymbol{I}_n - \boldsymbol{P})$ (and also $\boldsymbol{P}$) is symmetric idempotent. In scalar form $\mathbb{V}\text{ar}(\widehat{\epsilon}_i) = \sigma^2(1 - [\boldsymbol{P}]_{ii}) = \sigma^2(1 - p_{ii})$, for $i = 1, \ldots, n$. □

A rule of thumb is that any observation with $p_{ii} > 2p/n$ should be highlighted as a **high-leverage** point.

Example:

For simple linear regression we have that

$$\begin{aligned} p_{ii} &= \left[\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\right]_{ii} \\ &= \begin{bmatrix} 1 & x_i \end{bmatrix} (\boldsymbol{Z}^T\boldsymbol{Z})^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \end{aligned}$$

This implies that the points with the highest leverages are those that are furthest from $\bar{x}$.

5.3.2 Detecting outliers

The scale of the response variable $\boldsymbol{y}$ is arbitrary, and therefore so is the scale of the raw residuals. One way to standardise the residuals is to divide by their standard error.

We've just shown that

$$\mathbb{V}\text{ar}(\boldsymbol{\epsilon}) = \sigma^2(\boldsymbol{I} - \boldsymbol{P})$$

Definition 12. *The* **standardised residuals** *are*

$$r_i = \frac{\widehat{\epsilon}_i}{std.error(\widehat{\epsilon}_i)} = \frac{\widehat{\epsilon}_i}{s\sqrt{1 - p_{ii}}}.$$

Large r_i suggest that y_i is an outlier.

These are produced by R with the command `rstandard(fit)`.

The standardised residuals use the variance estimate s^2, which has been calculated using the entire data set. If there is an outlier, this will skew the estimate of σ^2, and so dividing by s^2 may not identify the point as a residual. To get around this, we can remove the i^{th} point from the model and calculate an estimate of σ^2 without this point.

Definition 13. *The* **Studentized residuals** *are*

$$t_i = \frac{\widehat{\epsilon}_i}{s_{(i)}\sqrt{1-p_{ii}}}$$

where $s^2_{(i)}$ is the unbiased estimate of σ^2 computed with the ith observation removed from the data.

The externally Studentized residuals are given in R by the function `rstudent(fit)`

It can be shown that $t_i \sim t_{n-k-1}$, and hence a value of $|t_i| > 2$ is generally considered to be large.

5.3.3 Influence

Informally, we can think of **influence** as the combination of being a high-leverage point and an outlier. The simplest way to assess the influence of the i^{th} data point is to remove it from the analysis, and see how much the regression coefficient estimates change.

Definition 14. *The* **Cook's distance**,

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{Z}^T \boldsymbol{Z} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{ps^2}$$

is a measure of the influence that a data point has on all of the fitted responses. It compares $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}_{(i)}$, where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the fitted value when the i^{th} observation is ignored.

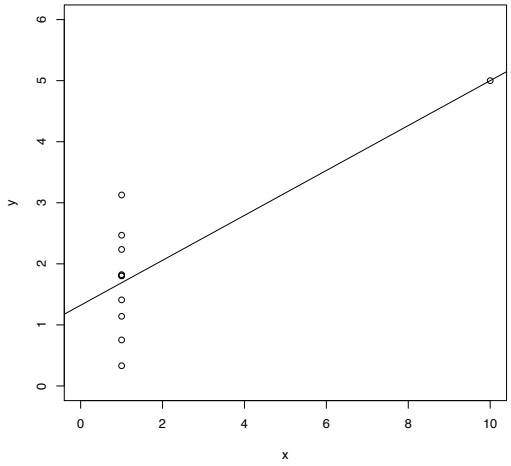


Figure 5.12: An extreme example. The fitted model depends entirely on a single point. If we move this point, the fitted line moves with it, and if we were to delete it, we could no longer fit a line.

This can be thought of as defining the distance from $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}_{(i)}$, taking the variance of $\widehat{\boldsymbol{\beta}}$ into account.

R2 stated that $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) = \sigma^2(Z^TZ)^{-1}$

It can be shown that

$$\begin{aligned} D_i &= \frac{(\widehat{\boldsymbol{y}}_{(i)} - \widehat{\boldsymbol{y}})^T (\widehat{\boldsymbol{y}}_{(i)} - \widehat{\boldsymbol{y}})}{ps^2} \\ &= \frac{r_i^2}{p} \times \left(\frac{p_{ii}}{1-p_{ii}} \right) \\ &= \text{"outlyingness"} \times \text{"leverage"} \end{aligned}$$

Cook's distances can be found in R using `cooks.distance(fit)`.

where $\widehat{\boldsymbol{y}}_{(i)}$ is the fitted response from the model which excludes the ith observation. Cases with large D_i are ones whose deletion will lead to substantial changes in the analysis.

What values for D_i are considered to be large? There are several rough guidelines, the simplest is simply to look for values with $D_i \geq 1$.

HOWEVER, THE BEST PRACTICE is to plot the D_i values for each observation, and to see if D_i for one or two observations are significantly larger than D_i for the rest of the observations. An attractive alternative is to plot the Studentised residuals t_i against the hat values p_{ii}, and look for observations which are big.

The `influencePlot` command in the `car` package produces this plot, with point size proportional to D_i.

5.3.4 Should we discard unusual data?

Outliers and influential data should not be ignored, but nor should they simply be deleted. It may be, as in the Davis data in the case study, that there are "bad" data points that can either be corrected or deleted. However, "good" observations that are unusual may help to suggest aspects of the model that are incorrect, or which we've over-looked.

The important thing is to investigate why an observation is unusual.

6

Model selection

A common statistical problem is to be given a large dataset consisting of n observations, $y_1, \ldots, y_n$ and a $n \times k$ matrix, X, of covariates, and to be asked to find a good model for predicting y. Often the number of covariates k can be large, and we wish to find a smaller subset of p, which can be used in a linear model to predict y. It is usually undesirable to include all k covariates, as models that are too complex are often over-fit, and give inaccurate predictions.

Model selection (often called 'variable selection' or more generally 'data mining') is the process of finding only the most important predictors. Application areas include:

- Social sciences (e.g. predicting crime hotspots)
- Marketing (e.g. Nectar card scheme).
- Pharmacology (e.g. drug discovery)
- Epidemiology (e.g. predicting disease incidence from genetic markers)

The most important criterion for including predictors in a model is the analyst's/experimenter's expert knowledge of the area under study.

We should always ask ourselves whether the model is plausible? i.e. does it make practical sense?

These techniques are usually applied when the focus is on prediction rather than explanation. If two explanatory variables are highly correlated, then from a predictive point of view it doesn't matter which we include in the model.

6.0.5 The model hierarchy

Definitions

- The **null** model is where we fit just a constant mean function to the data.
- The **minimal** model is the simplest model consistent with *known* features of the experiment, the data and the underlying theory.
- The **maximal** model is the most complex model worth considering in the analysis.

- The **saturated** model is the model with n parameters (the same number as data points).

In the saturated model, $\mathbf{Z}$ is $n \times n$ and (we assume) of full rank, so $\mathbf{Z}^{-1}$ exists.

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y} = \mathbf{Z}^{-1}\boldsymbol{y}$$

and so

$$\widehat{\boldsymbol{y}} = \mathbf{Z}\widehat{\boldsymbol{\beta}} = \mathbf{Z}\mathbf{Z}^{-1}\boldsymbol{y} = \boldsymbol{y},$$

i.e. the saturated model fits exactly.

Why is it a bad idea to use the saturated model for prediction?

In judging the adequacy of the *current* model we must be aware of the model hierarchy:

$$\left\{\begin{array}{l}\text{null}\\ \text{minimal}\\ \text{current}\\ \text{maximal}\\ \text{saturated}\end{array}\right.$$

6.1 Automated Variable Selection

F-tests are useful for comparing nested models, but can't be used in more general cases. Moreover, care needs to be taken when doing repeated hypothesis tests: the type I error rate is no longer α if we perform n independent hypothesis tests each at significance level α! Moreover, for small samples F-tests may lack power leading us to fail to reject a false null, which is completely different from confirming H_0. Further, for large n, we shall nearly always reject H_0.

The fallacy of affirming the consequent

Instead of using statistical significance for comparing two models, we can use brute force model selection techniques that take advantage of modern computer power, and use a computer to search through many different possible models to pick the best. While this is quick, if used improperly, it can lead to poor choices being made.

Variable selection methods have been described like doing carpentry with a chain saw: you can get a lot work done quickly, but you may end up doing more harm than good.

We need a criterion to optimize in order to choose the "best" model. The criteria are all a balance between requiring goodness of fit, while penalising for complexity. Models that are too complex are often over-fit[1], and thus have poor performance on prediction tests. Various different measures have been proposed, but we focus on three:

Parsimony

> 'Everything should be made as simple as possible, but not simpler.' Albert Einstein

In statistics, the most parsimonious model is the simplest one that still adequately fits the data. In practice this means the model with the fewest parameters. If a model is unnecessarily complex, the precision of estimation and prediction decreases.

[1] A model is over-fit if it describes the random noise rather than the underlying relationship.

(i) Adjusted R^2. This is intuitively reasonable, but has no theoretical justification.

(ii) Mallow's C_p

$$C_p = \frac{\text{ResidSS(current model)}}{s^2} + 2p - n,$$

where

$$s^2 = \frac{\text{ResidSS(full model)}}{n-k}$$

is the unbiased estimator of the variance of the full model. To interpret C_p, note that if a subset model with p covariates fits well, then

$$\mathbb{E}(RSS(p)) = (n-p)\sigma^2$$

so that $C_p \approx p$. Whereas if important predictor variables have

been omitted from the model, then $RSS(p)$ is an estimate of $(n-p)\sigma^2$ plus a positive term, so that $C_p > p$. A good model therefore has C_p either around or less than p. Minimising C_p for models of a given size minimises the residual sum of squares and thus maximises R^2.

(iii) The Akaike information criterion (AIC) is defined to be

$$AIC = -2\log(L) + 2P$$

and the Bayesian information criterion (BIC) is defined to be

$$BIC = -2\log(L) + P\log n$$

where $\log(L)$ is the log-likelihood calculated at the maximized likelihood estimate of β and σ^2, and P is the number of parameters that needed to be estimated. By noting that $\hat{\sigma}^2 = \frac{RSS}{n}$, we can show that for linear models

$$AIC \propto n\log(RSS) + 2p$$
$$BIC \propto n\log(RSS) + p\log n$$

The theoretical motivation for the AIC and BIC are complex, but they can both be seen as trade-offs between the goodness of fit of the model (measured by $-2\log(L)$) and the model complexity P. There are many other ICs that have been proposed.

Only relative values of the AIC/BIC are of interest - we compare the AIC/BIC for two different models and choose the model with the lowest value.

The AIC and BIC can be calculated in R using the command `AIC(fit)` and `BIC(fit)`.

6.2 Best subsets regression

Best subsets regression is a method of variable selection in which all possible regressions are performed and the best models for each number of parameters are suggested. The statistician then chooses the most appropriate model (or models) from the subset presented, based on the values of measures like adjusted-R^2, Mallow's C_p, or the AIC. If we are using C_p, we aim to find the **simplest model with** $C_p \leqslant p$ (if this isn't possible report model(s) with lowest C_p).

Note that, if we have k covariates, then there are 2^k possible models. This will quickly become too large to do an exhausitve search, and so instead we can use methods such as....

6.3 Stepwise regression

If the number of input variables is very large, then stepwise regression is often used for variable selection instead of best subsets regression.

In stepwise regression, we start from an initial model and then add or remove predictors based on the criterion value of the model that includes that variable. We choose to add or remove the variable that leads to the model with the smallest value of the criterion. The advantage of this approach, is that instead of trying 2^k different models, we only need to try k different models at each stage. The disadvantage is that we are not considering all possible models. For example if two input variates are only informative in combination then neither will ever be added to the model.

The special case of stepwise regression where we start from the minimal model and only add predictors is called forward regression, and the special case where we start from the full model and only remove predictors is called backward regression.

Warning:

- Automatic variable selection proceedures like best subsets regression and stepwise regression are not perfect.
- They are no substitute for thinking about which predictors should be important.
- They may include variables that just happen to explain the data by chance, and are therefore useless for prediction.

For example if you generated 10 columns of random numbers and used them as predictors, automatic variable selection methods will probably find that one or more of them 'explains' some of the variation in the response, even though each column is equally useful (ie. completely useless) for prediction.

6.4 Ridge regression

Recall that theorem R0 required that $Z^\top Z$ was invertible, and that this was the case if $\mathrm{rank}(Z) = p$. What happens if $n < p$? Then we have more columns than rows so the columns can not be linearly independent and so $\mathrm{rank}(Z) < p$, and hence $Z^\top Z$ is not invertible. We can also find that $(Z^\top Z)^{-1}$ does not exist when $n > p$ if some of the columns of Z are close to being colinear.

$\mathrm{rank}(Z)$ is the number of linearly independent columns of Z, which equals the number of linearly independent rows

One solution to this problem is to use ridge regression. Ridge regression is like ordinary least squares regression, except we add a penalty term to constrain the size of the parameter.

We choose $\boldsymbol{\beta}$ to minimise

$$\begin{aligned} S_r(\boldsymbol{\beta}) &:= \sum_{i=1}^{n}(y_i - z_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \\ &= (\boldsymbol{y} - Z\boldsymbol{\beta})^\top (\boldsymbol{y} - Z\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \end{aligned} \tag{6.1}$$

The second term $\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$ is a penalty term that penalises values of $\boldsymbol{\beta}$ that are large. The parameter λ is a complexity parameter which controls how strongly large values of $\boldsymbol{\beta}$ are punished. We only allow $\lambda \geq 0$.

The ridge regression estimator, denoted $\widehat{\boldsymbol{\beta}}_r$, is the value of $\boldsymbol{\beta}$ which minimises $S_r(\boldsymbol{\beta})$. Note that if λ is zero, this is equivalent to least-squares regression and we find $\widehat{\boldsymbol{\beta}}_r = \widehat{\boldsymbol{\beta}}$. As λ grows, the penalty for large $\boldsymbol{\beta}$ values grows. In the limit $\lambda = \infty$ the optimal solution is to take $\widehat{\boldsymbol{\beta}}_r = 0$. We can see that the effect of the penalty term is to shrink the parameter estimates towards 0. Note also that the estimator $\widehat{\boldsymbol{\beta}}_r$ is now a function of λ.

Adding a penalty term to the sum of squares is called regularisation and is a very powerful approach for finding good parameter estimates in over-parametrized models.

Proposition 5.

$$\widehat{\boldsymbol{\beta}}_r = (Z^\top Z + \lambda I)^{-1} Z^\top \boldsymbol{y}$$

minimizes Equation (6.1).

Note that $Z^\top Z + \lambda I$ can be made to be invertible by increasing the size of λ sufficiently, regardless of the value of Z.

Proof. Very similar to the proof of R0. □

We can think of ridge regression as shrinking the parameters towards zero.

Proposition 6.

$$||\widehat{\boldsymbol{\beta}}_r||_2 \leq ||\widehat{\boldsymbol{\beta}}||_2$$

Proof. Suppose not. Then

$$\begin{aligned} S_r(\widehat{\boldsymbol{\beta}}_r) &= (\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_r)^\top(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_r) + \lambda||\widehat{\boldsymbol{\beta}}_r||_2^2 \\ &> (\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}) + \lambda||\widehat{\boldsymbol{\beta}}||_2^2 \\ &= S_r(\widehat{\boldsymbol{\beta}}) \end{aligned}$$

which is a contradiction as $\widehat{\boldsymbol{\beta}}_r$ minimises S_r.

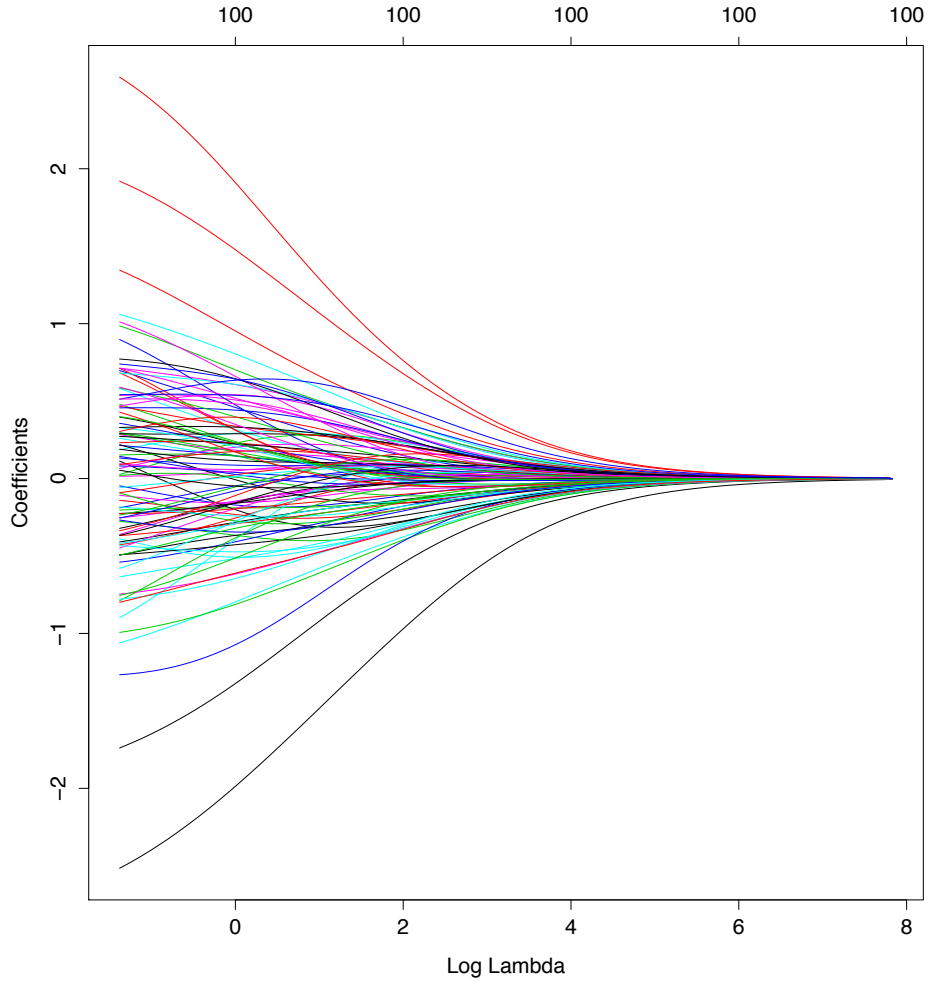
□

6.4.1 Why is ridge regression useful?

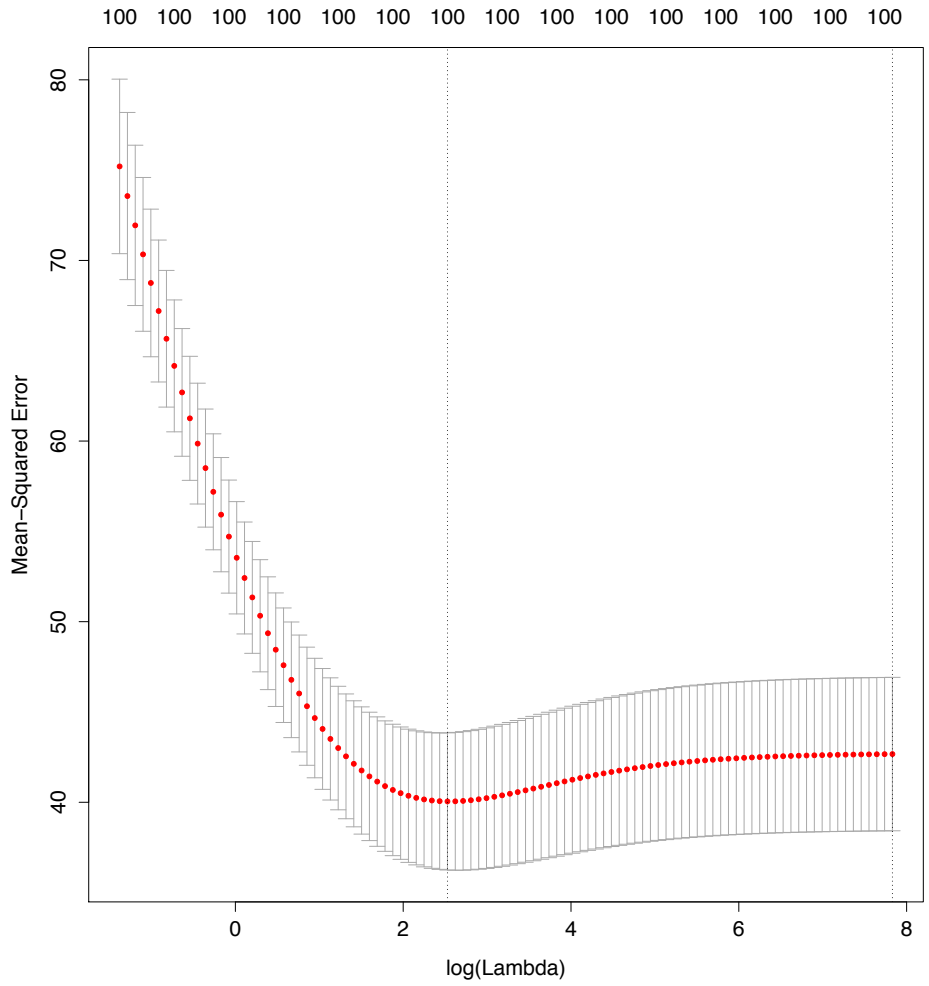
Retaining a subset of regressors and discarding the rest, as we did in stepwise or best susbsets regression, produces a model that is easily interpretable and has possibly lower prediction error than the true model. However, because variables are either retained or discarded, it often exhibits high variance. Shrinkage methods, such as ridge regression, are more continuous, and so don't suffer as much from high variance.

Ridge regression adds an L_2 penalty $\lambda||\widehat{\boldsymbol{\beta}}||_2^2$ to the sum of squares. Other forms of penalisation are popular (and are actively being researched) and can have different effects. In particular, models fit using the L_1 norm are have particularly attractive properties.

```
> install.package('glmnet')
> # install package first time it is used - requires an internet connection
> library(glmnet)
>
> ## create some data
> n=150 # number of data point
> p=100 # number of covaoriates in X
> ptrue=10  # number of covariates that have any effect
> x=matrix(rnorm(n*p),n,p) # covariate matrix
> beta=rnorm(ptrue) # generate random true parameter value
>
> # generate observations
> y= x[,1:ptrue] %*% beta + rnorm(n,0,5)
>
> ### Fit ordinary least squares regression
> OLS <- lm(y ~  x)
>
> ### Now do ridge regression
> ridge = glmnet(x, y, alpha=0)  # alpha=0 gives ridge regression
> # Other values of alpha give different regularisation penalties
>
> plot(ridge, xvar='lambda')
```




```
> # 10 fold cross-validation to find the best value of lambda
> cvridge=cv.glmnet(x,y, alpha=0)
>
> # How the prediction error varies with lambda
> plot(cvridge)
>
```




```
> cvridge$lambda.min # value of lambda that minimizes the CV error
[1] 12.52552
>
> cvridge$lambda.1se
[1] 2516.669
> # largest value of lambda that gives a CV error within
> # 1standard deviation of the minimum
>
>
> ### create some test data
> xnew = matrix(rnorm(n*p),n,p)
> ynew =xnew[,seq(ptrue)] %*% beta + rnorm(n,0,5)
>
> ## predict at new x values
> OLS.prediction  = predict(OLS, data.frame(xnew))
> ridge.prediction1 = predict(cvridge, xnew, s = "lambda.1se")
> ridge.prediction2 = predict(cvridge, xnew, s = "lambda.min")
>
```

```
> ## calculate the prediction mean square error
> OLS.mse = mean((OLS.prediction - ynew)^2)
> ridge1.mse = mean((ridge.prediction1 - ynew)^2)
> ridge2.mse = mean((ridge.prediction2 - ynew)^2)
>
> print(OLS.mse)
[1] 69.11146
> print(ridge1.mse)
[1] 42.33062
> print(ridge2.mse)
[1] 36.55833
>
>
> ## Do the same with step-wise regression
> OLS2 <- lm(y ~ ., data.frame(x))
> stepfit <- step(OLS2)

 [Output omitted - as its several pages long]

 > stepfit

Call:
lm(formula = y ~ X1 + X2 + X4 + X7 + X8 + X10 + X11 + X12 + X13 +
    X15 + X18 + X20 + X30 + X34 + X35 + X37 + X38 + X47 + X49 +
    X51 + X53 + X56 + X64 + X65 + X66 + X68 + X69 + X71 + X72 +
    X78 + X81 + X84 + X88 + X90 + X93 + X94 + X95 + X97 + X98,
    data = data.frame(x))

Coefficients:
(Intercept)           X1           X2           X4           X7
     0.2180      -2.1981       1.4949      -0.8936      -2.1013
         X8          X10          X11          X12          X13
     2.4690       0.5780      -1.3254       0.6264      -0.7887
        X15          X18          X20          X30          X34
    -1.1453       0.6247       0.8945      -0.4747      -1.3311
        X35          X37          X38          X47          X49
    -0.6589      -0.8051       1.7640      -0.7050      -0.6686
        X51          X53          X56          X64          X65
     1.2445       1.1418       1.2341       0.9172       0.9905
        X66          X68          X69          X71          X72
     0.5608       0.9082      -0.7909       0.4558       0.5351
```

```
      X78          X81          X84          X88          X90
   1.1435       0.9946      -0.8425       0.7150       0.5983
      X93          X94          X95          X97          X98
  -0.7096       0.8366      -0.7435      -0.7579      -0.9696
```

```
> step.prediction  = predict(stepfit, data.frame(xnew))
> step.mse = mean((step.prediction - ynew)^2)
> print(step.mse)
[1] 64.79377
```

So in this case, OLS gives the worst model (as measured by the prediction error), followed by the model found by stepwise linear regression, and ridge regression by far the best prediction error.

Notes

1. `glmnet` standardizes all the covariates so that they are on the same scale, so that the estimated parameters are of similar size and are penalized similarly. Coefficients are always returned on the original scale though.
2. Usually, the intercept term is not penalized. This term (the mean) isn't seen as adding complexity, and so it makes no sense to shrink it. By default, `glmnet` adds an intercept to the model and does not apply a penalty term to the size of the intercept.
3. Ridge regression is not strictly a model-selection method, as it doesn't leave variables out it just shrinks the parameter estimates. However, as we've just seen, it is useful for finding models that predict well.

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

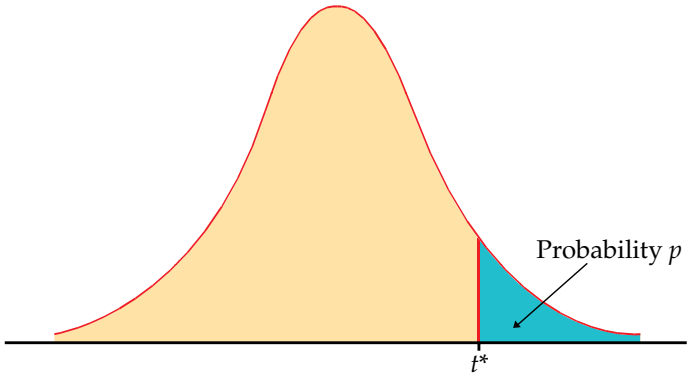


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Table of critical values for the F distribution (for use with ANOVA):

How to use this table:

There are two tables here. The first one gives critical values of F at the p = 0.05 level of significance. The second table gives critical values of F at the p = 0.01 level of significance.
1. Obtain your F-ratio. This has (x,y) degrees of freedom associated with it.
2. Go along x columns, and down y rows. The point of intersection is your critical F-ratio.
3. If your obtained value of F is equal to or larger than this critical F-value, then your result is significant at that level of probability.
An example: I obtain an F ratio of 3.96 with (2, 24) degrees of freedom.
I go along 2 columns and down 24 rows. The critical value of F is 3.40. My obtained F-ratio is larger than this, and so I conclude that my obtained F-ratio is likely to occur by chance with a p<.05.

Critical values of F for the 0.05 significance level:

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
31	4.16	3.31	2.91	2.68	2.52	2.41	2.32	2.26	2.20	2.15
32	4.15	3.30	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.29	2.89	2.66	2.50	2.39	2.30	2.24	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11

36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.15	2.10
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08
40	4.09	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
41	4.08	3.23	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.07
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.07
43	4.07	3.21	2.82	2.59	2.43	2.32	2.23	2.16	2.11	2.06
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04
47	4.05	3.20	2.80	2.57	2.41	2.30	2.21	2.14	2.09	2.04
48	4.04	3.19	2.80	2.57	2.41	2.30	2.21	2.14	2.08	2.04
49	4.04	3.19	2.79	2.56	2.40	2.29	2.20	2.13	2.08	2.03
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
51	4.03	3.18	2.79	2.55	2.40	2.28	2.20	2.13	2.07	2.02
52	4.03	3.18	2.78	2.55	2.39	2.28	2.19	2.12	2.07	2.02
53	4.02	3.17	2.78	2.55	2.39	2.28	2.19	2.12	2.06	2.02
54	4.02	3.17	2.78	2.54	2.39	2.27	2.19	2.12	2.06	2.01
55	4.02	3.17	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01
56	4.01	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.05	2.01
57	4.01	3.16	2.77	2.53	2.38	2.26	2.18	2.11	2.05	2.00
58	4.01	3.16	2.76	2.53	2.37	2.26	2.17	2.10	2.05	2.00
59	4.00	3.15	2.76	2.53	2.37	2.26	2.17	2.10	2.04	2.00
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
61	4.00	3.15	2.76	2.52	2.37	2.25	2.16	2.09	2.04	1.99
62	4.00	3.15	2.75	2.52	2.36	2.25	2.16	2.09	2.04	1.99
63	3.99	3.14	2.75	2.52	2.36	2.25	2.16	2.09	2.03	1.99
64	3.99	3.14	2.75	2.52	2.36	2.24	2.16	2.09	2.03	1.98
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98
66	3.99	3.14	2.74	2.51	2.35	2.24	2.15	2.08	2.03	1.98
67	3.98	3.13	2.74	2.51	2.35	2.24	2.15	2.08	2.02	1.98
68	3.98	3.13	2.74	2.51	2.35	2.24	2.15	2.08	2.02	1.97
69	3.98	3.13	2.74	2.51	2.35	2.23	2.15	2.08	2.02	1.97
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
71	3.98	3.13	2.73	2.50	2.34	2.23	2.14	2.07	2.02	1.97
72	3.97	3.12	2.73	2.50	2.34	2.23	2.14	2.07	2.01	1.97
73	3.97	3.12	2.73	2.50	2.34	2.23	2.14	2.07	2.01	1.96
74	3.97	3.12	2.73	2.50	2.34	2.22	2.14	2.07	2.01	1.96
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96
76	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96
77	3.97	3.12	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.96
78	3.96	3.11	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.95
79	3.96	3.11	2.72	2.49	2.33	2.22	2.13	2.06	2.00	1.95
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
81	3.96	3.11	2.72	2.48	2.33	2.21	2.13	2.06	2.00	1.95
82	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	2.00	1.95
83	3.96	3.11	2.72	2.48	2.32	2.21	2.12	2.05	2.00	1.95
84	3.96	3.11	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.95
85	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94

86	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94
87	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94
88	3.95	3.10	2.71	2.48	2.32	2.20	2.12	2.05	1.99	1.94
89	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
91	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.98	1.94
92	3.95	3.10	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.94
93	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
94	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
95	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
96	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
97	3.94	3.09	2.70	2.47	2.31	2.19	2.11	2.04	1.98	1.93
98	3.94	3.09	2.70	2.47	2.31	2.19	2.10	2.03	1.98	1.93
99	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.98	1.93
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.98	1.93

Critical values of F for the 0.01 significance level:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052.19	4999.52	5403.34	5624.62	5763.65	5858.97	5928.33	5981.10	6022.50	6055.85
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.93	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.52	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.90	3.81
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.02	3.84	3.71	3.60	3.51
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.77	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.05	3.73	3.50	3.33	3.20	3.09	3.01
30	7.56	5.39	4.51	4.02	3.70	3.47	3.31	3.17	3.07	2.98
31	7.53	5.36	4.48	3.99	3.68	3.45	3.28	3.15	3.04	2.96
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93

33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86
37	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83
39	7.33	5.19	4.33	3.84	3.53	3.31	3.14	3.01	2.90	2.81
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
41	7.30	5.16	4.30	3.82	3.50	3.28	3.11	2.98	2.88	2.79
42	7.28	5.15	4.29	3.80	3.49	3.27	3.10	2.97	2.86	2.78
43	7.26	5.14	4.27	3.79	3.48	3.25	3.09	2.96	2.85	2.76
44	7.25	5.12	4.26	3.78	3.47	3.24	3.08	2.95	2.84	2.75
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74
46	7.22	5.10	4.24	3.76	3.44	3.22	3.06	2.93	2.82	2.73
47	7.21	5.09	4.23	3.75	3.43	3.21	3.05	2.92	2.81	2.72
48	7.19	5.08	4.22	3.74	3.43	3.20	3.04	2.91	2.80	2.72
49	7.18	5.07	4.21	3.73	3.42	3.20	3.03	2.90	2.79	2.71
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70
51	7.16	5.05	4.19	3.71	3.40	3.18	3.01	2.88	2.78	2.69
52	7.15	5.04	4.18	3.70	3.39	3.17	3.01	2.87	2.77	2.68
53	7.14	5.03	4.17	3.70	3.38	3.16	3.00	2.87	2.76	2.68
54	7.13	5.02	4.17	3.69	3.38	3.16	2.99	2.86	2.76	2.67
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66
56	7.11	5.01	4.15	3.67	3.36	3.14	2.98	2.85	2.74	2.66
57	7.10	5.00	4.15	3.67	3.36	3.14	2.97	2.84	2.74	2.65
58	7.09	4.99	4.14	3.66	3.35	3.13	2.97	2.84	2.73	2.64
59	7.09	4.98	4.13	3.66	3.35	3.12	2.96	2.83	2.72	2.64
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
61	7.07	4.97	4.12	3.64	3.33	3.11	2.95	2.82	2.71	2.63
62	7.06	4.97	4.11	3.64	3.33	3.11	2.94	2.81	2.71	2.62
63	7.06	4.96	4.11	3.63	3.32	3.10	2.94	2.81	2.70	2.62
64	7.05	4.95	4.10	3.63	3.32	3.10	2.93	2.80	2.70	2.61
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61
66	7.04	4.94	4.09	3.62	3.31	3.09	2.92	2.79	2.69	2.60
67	7.03	4.94	4.09	3.61	3.30	3.08	2.92	2.79	2.68	2.60
68	7.02	4.93	4.08	3.61	3.30	3.08	2.91	2.79	2.68	2.59
69	7.02	4.93	4.08	3.60	3.30	3.08	2.91	2.78	2.68	2.59
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
71	7.01	4.92	4.07	3.60	3.29	3.07	2.90	2.77	2.67	2.58
72	7.00	4.91	4.07	3.59	3.28	3.06	2.90	2.77	2.66	2.58
73	7.00	4.91	4.06	3.59	3.28	3.06	2.90	2.77	2.66	2.57
74	6.99	4.90	4.06	3.58	3.28	3.06	2.89	2.76	2.66	2.57
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57
76	6.98	4.90	4.05	3.58	3.27	3.05	2.88	2.76	2.65	2.56
77	6.98	4.89	4.05	3.57	3.27	3.05	2.88	2.75	2.65	2.56
78	6.97	4.89	4.04	3.57	3.26	3.04	2.88	2.75	2.64	2.56
79	6.97	4.88	4.04	3.57	3.26	3.04	2.87	2.75	2.64	2.55
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
81	6.96	4.88	4.03	3.56	3.25	3.03	2.87	2.74	2.63	2.55
82	6.95	4.87	4.03	3.56	3.25	3.03	2.87	2.74	2.63	2.55

83	6.95	4.87	4.03	3.55	3.25	3.03	2.86	2.73	2.63	2.54
84	6.95	4.87	4.02	3.55	3.24	3.03	2.86	2.73	2.63	2.54
85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54
86	6.94	4.86	4.02	3.55	3.24	3.02	2.85	2.73	2.62	2.53
87	6.94	4.86	4.02	3.54	3.24	3.02	2.85	2.72	2.62	2.53
88	6.93	4.86	4.01	3.54	3.23	3.01	2.85	2.72	2.62	2.53
89	6.93	4.85	4.01	3.54	3.23	3.01	2.85	2.72	2.61	2.53
90	6.93	4.85	4.01	3.54	3.23	3.01	2.85	2.72	2.61	2.52
91	6.92	4.85	4.00	3.53	3.23	3.01	2.84	2.71	2.61	2.52
92	6.92	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.61	2.52
93	6.92	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.60	2.52
94	6.91	4.84	4.00	3.53	3.22	3.00	2.84	2.71	2.60	2.52
95	6.91	4.84	4.00	3.52	3.22	3.00	2.83	2.70	2.60	2.51
96	6.91	4.83	3.99	3.52	3.21	3.00	2.83	2.70	2.60	2.51
97	6.90	4.83	3.99	3.52	3.21	2.99	2.83	2.70	2.60	2.51
98	6.90	4.83	3.99	3.52	3.21	2.99	2.83	2.70	2.59	2.51
99	6.90	4.83	3.99	3.52	3.21	2.99	2.83	2.70	2.59	2.51
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50