Statistical Modelling I
MATH 50011

Dr Riccardo Passeggeri
Imperial College London

Spring 2022

Contents

1 Statistical Models **5**

1.1 Notation 5

1.2 Parametric Statistical Models 5

1.3 Using Models 7

2 Point Estimation **9**

2.1 Properties of Estimators 10

2.1.1 Bias 10

2.1.2 Standard Error 11

2.1.3 Mean Square Error 12

3 The Cramér-Rao Lower Bound **15**

4 Asymptotic Properties **18**

5 Maximum Likelihood Estimation **24**

5.1 Properties of Maximum Likelihood estimators 28

5.1.1 MLEs are functionally invariant 28

5.1.2 Large sample properties 30

5.2 Sketch of proofs 32

6 Confidence Regions **35**

6.1 Construction of Confidence Intervals 38

6.2 Asymptotic confidence intervals 40

6.3 Simultaneous Confidence Intervals/Confidence Regions 42

7 Hypothesis Tests **43**

7.1 Power of a Test 45

7.2 p-value 47

7.3 Connection between tests and confidence intervals 48

8 Likelihood Ratio Tests **50**

9 Linear Models with Second Order Assumptions **56**

9.1 Simple Linear Regression 56

9.2 Matrix Algebra 58

9.3 Review of rules for E, cov for random vectors 60

9.4 Linear Model 63

9.5 Identifiability 66

9.6 Least Squares Estimation 67

9.7 Properties of Least Squares Estimation . . . 68
9.8 Projection Matrices . . . 71
9.9 Residuals, Estimation of the variance . . . 73

10 Linear Models with Normal Theory Assumptions 80

10.1 Distributional Results . . . 80
10.1.1 The Multivariate Normal Distribution . . . 80
10.1.2 Distributions derived from the Multivariate Normal . . . 83
10.1.3 Some Independence Results . . . 85
10.2 The Linear Model with Normal Theory Assumptions . . . 88
10.3 Confidence Intervals, Tests for one-dimensional quantities . . . 89
10.4 The F-Test . . . 91
10.5 Confidence Regions . . . 95

11 Diagnostics, Model Selection, Extensions 98

11.1 Outliers . . . 98
11.2 Leverage . . . 99
11.3 Cook's Distance . . . 100
11.4 Residual Plots . . . 101
11.5 Distributional Checks . . . 101
11.6 Weighted Least Squares . . . 104
11.7 Under/overfitting . . . 105
11.8 Model Selection . . . 106
11.9 Generalized Linear Models (GLMs) . . . 107

Background and Scope

You have now had one or more modules introducing the key concepts of probability and statistics. In this module, we delve more deeply into selected topics and apply statistical methods to real data.

Broadly speaking, statistics is all about science. Science is about proving things, which requires demonstrating the validity of hypotheses.

Hypotheses don't just pop into existence from thin air, though. A common first step in scientific investigations is thus hypothesis generation. This is typically driven by **observational studies** in which an association between events or objects is noted in existing populations or systems. We must be careful not to ascribe cause and effect from an observational study unless we have systematically ruled out all confounding factors (i.e. other objects/events that explain the observed relationship).

Once we have a scientific hypothesis, we are in a position to refine its statement or confirm its truth. To do this, the gold standard is to design a **controlled experiment** in which we directly intervene on a well-defined study population or system.

Statistics can aid the process of scientific inquiry when we can translate our scientific hypothesis into statements or comparisons about one or more numbers, and we know how we would answer the scientific question if we had knowledge of the entire population.

The primary goal of this module is to introduce the theory and application of linear models, which include and substantially generalise the simple linear regression models you will have previously encountered. In service of this goal, we will also derive further results that apply to other models. This preamble is not the last time we consider the role of statistics in science.

1 Statistical Models

The methods we consider deal with the analysis of relationships between measurements collected on groups of subjects or objects.

When a measurement is expected to vary in response to other variables, we call it a **response**, **outcome** or **dependent variable** (these terms are exchangeable in the literature).

The measurements that are believed to lead to a change in the response are called **covariates**, **explanatory variables**, **predictor variables**, or **independent variables**. These terms are not exhaustive. In economics, the terms **endogenous** (dependent variable) and **exogenous** (independent variable) are common.

We must be careful when making causal statements about the relationships between variables we have called the response and predictor.

In this module, we will consider settings with just one response variable, but allow there to be possibly many predictor variables. At all times, we model measurements as samples from some random process.

1.1 Notation

The typical convention of denoting random variables by uppercase italic letters such as X, Y, Z and observed data values by corresponding lowercase letters such as x, y, z will be followed. Similarly, greek letters such as $\theta, \mu, \sigma, \beta$ are used to denote parameters. An estimator or estimate of a parameter will typically be denoted by placing a decoration such as a "hat" on the corresponding letter, such as $\hat{\theta}, \hat{\mu}, \hat{\sigma}, \hat{\beta}$. At all times, exceptions to these conventions will be made clear by context.

1.2 Parametric Statistical Models

In the following, we suppose that we observe some data y. We interpret this as a realisation of some random object Y.

A *statistical model* is a collection of probability distribtions $\{P_\theta : \theta \in \Theta\}$ on a given sample space. The set Θ of all possible parameter values is called the *parameter space*.

In this module, $\Theta \subseteq \mathbb{R}^p$, so that we consider *parametric models*. A semiparametric or nonparametric model involves parameters that belong to more general sets (e.g. function spaces).

Since we interpret the data y as realisation of a random variable Y, a *statistical model* can be seen as a specification of the distribution of Y up to an unknown parameter θ.

A statistical model (or a parametrization) is generally required to be such that distinct parameter values give rise to distinct distributions. In particular, a statistical model is called *identifiable* if the mapping $\theta \mapsto P_\theta$ is one-to-one, that is $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ must hold for all $\theta_1, \theta_2 \in \Theta$.

Often, the data $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ is a vector and $Y = (Y_1, \dots, Y_n)$ is a random vector. In this case the statistical model is the specification of the joint distribution of $Y_1, \dots, Y_n$ up to some unknown parameter θ.

If $Y_1, \dots, Y_n$ are iid (independent and identically distributed) then $Y_1, \dots, Y_n$ are called a *random sample* (or just *sample*). However, this terminology is not universally accepted, hence it is always a good practice to explicitely mention that the observations are iid.

> **Example 1**
> At a fair, a competition is held where participants guess the weight of an ox based only on visual inspection. The organizer of the competition has weighed the ox and knows that the animal weighs exactly 543.4 kg (about 1198 lbs). Suppose that participants are equally likely to overestimate or underestimate the weight, and that extremely poor guesses are unlikely. We can then reasonably model the guess y (measured in kg) of a given participant as the realization of a normally distributed random variable $Y \sim N(543.4, \sigma^2)$ for some $\sigma > 0$. The collection of n guesses $Y_1, \dots, Y_n$ is a random sample from the $N(543.4, \sigma^2)$ distribution.

In many situations, $Y_1, \dots, Y_n$ are independent but do not have the same distribution. For instance, the distribution of $Y_1, \dots, Y_n$ may depend on (nonrandom) values $x_1, \dots, x_n$. The x_i's are an example of covariates.

> **Example 2**
> - In a clinical trial, the survival times with a new and an old treatment are compared. The study enrolls n patients, and the treatments are randomly

allocated. The outcome for the ith patient is $Y_i =$ survival time. As covariate for the ith patient, we may use

$$x_i = \begin{cases} 1, & \text{patient receives new treatment} \\ 0, & \text{patient receives old treatment} \end{cases}$$

- Do taller people have a higher income?

  $Y_i =$ income, $x_i =$ height, $i = 1, \ldots, n$

  Model: $Y_i = \beta_0 + x_i \beta_1 + \epsilon_i$, $i = 1, \ldots, n$, $\epsilon_i \sim N(0, \sigma^2)$ iid, $\theta = (\beta_0, \beta_1, \sigma^2)$, $\Theta = \mathbb{R}^2 \times [0, \infty)$

Other commonly used distributions in statistics also give rise to parametric families. For instance, if we assume that there exists $\lambda_0 > 0$ such that a random count Y (e.g. the number of named storms in a given year) has a Poisson distribution with mean λ_0, then the corresponding parametric model is the set of distributions $\{\text{Poisson}(\lambda) : \lambda > 0\}$. We will see later how the model that we assume can help us choose between estimators for a quantity of interest.

1.3 Using Models

Having formulated a model, we can then draw inferences from the sample to answer our scientific question. Often, the first step is to estimate any unknown parameters required to answer our scientific question. This step is often called "fitting the model," though we stress that this step is really *fitting the data to the model*.

A *point estimate* is convenient, because it provides a single number (or vector) to answer our scientific question. However, we will often be interested in using some combination of *hypothesis tests* and *confidence intervals* to help make decisions about our scientific question in a manner that accounts for random variation in our sample.

When we adopt a model, we must accept that it will not perfectly reflect reality. However, we still want the model to be useful! A model should

- ...agree with observed data reasonably well.
- ...be reasonably simple (no more parameters than necessary).

- ...be easy to interpret, e.g. parameters should have a practical meaning.

In service of the above aims, we might conduct sensitivity analyses that try to answer the question, "Is the model adequate for the data?" If the answer to this is "No," then we should refine the model. This may become an iterative procedure, as we will discuss at the end of the module. It is important that we clearly document this process in any analysis, to preserve the validity of our primary results.

2 Point Estimation

How can we estimate unknown parameters based on observed data $y_1, \dots, y_n$? Using our model, we view y_i a realisation of the random variables Y_i for $i = 1, \dots, n$.

A function of observable random variables is called *statistic*. *Any* statistic *could* be used to estimate θ.

Definition 1
Let t be a statistic. The value $t(y_1, \dots, y_n)$ is called the **estimate** of θ. The random variable $T = t(Y_1, \dots, Y_n)$ is called an **estimator** of θ.

There can be many candidate estimates for a parameter.

Example 3
Model: $Y_1, \dots, Y_n \sim N(\mu, 1)$ iid, $\mu \in \mathbb{R}$ unknown.
Even in this simple situation μ can be estimated in several ways:

- sample mean: $\bar{y} = \frac{1}{n} \sum y_i$
- sample median: $y_{((n+1)/2)}$ if n is odd and $(y_{(n/2)} + y_{(n/2+1)})/2$ if n is even, where $y_{(1)} < \cdots < y_{(n)}$ is the ordered sample
- trimmed mean: discard the highest and lowest k observed y_i before computing the mean
- ...

For the estimate $t(y_1, \dots, y_n) = \bar{y}$ the corresponding estimator is $T(Y_1, \dots, Y_n) = \bar{Y} = \frac{1}{n} \sum Y_i$.
Note: T is a r.v. Its distribution may depend on $\theta = \mu$.
Here: $T \sim N(\mu, 1/n)$.
To prove that T has the above distribution: Show that the sum of independent normal random variables is normally distributed and compute the mean and variance of T.

Remark We judge how good t is by looking at the properties of T.

2.1 Properties of Estimators

We assume that we have specified a parametric model for the data. That means that we have a set of possible distributions, indexed by a parameter $\theta \in \Theta$, that could have generated the observations. Whenever we are working out properties of estimators (such as probabilities, expected values or variances), we have to indicate with respect to which of these distributions we are doing our calculations. To do this we usually write the parameter as subscript, e.g. we write

$$P_\theta(T \in \mathcal{A}), \quad \mathsf{E}_\theta(T), \quad \mathsf{Var}_\theta(T).$$

2.1.1 Bias

The bias of an estimator is the difference between its expected value and the true value. More precisely:

Definition 2
Let T be an estimator for $\theta \in \Theta \subset \mathbb{R}$. The bias of T is defined by $\mathsf{bias}_\theta(T) = \mathsf{E}_\theta(T) - \theta$. If $\mathsf{bias}_\theta(T) = 0$ for all $\theta \in \Theta$, we say that T is unbiased for θ.

Example 4
Let $X \sim \mathsf{Binomial}(n, p)$. Here, $p \in [0, 1]$ is the unknown parameter and n is known.

Consider the two estimators $S = X/n$ and $T = \frac{X+1}{n+2}$ for p.

$\forall p$: $\mathsf{bias}_p\, S = \mathsf{E}_p(S - p) = \dfrac{1}{n}\,\mathsf{E}_p\, X - p = 0$. Thus S is unbiased for p.

$$\mathsf{bias}_p\, T = \mathsf{E}_p(T - p) = \frac{\mathsf{E}_p\, X + 1}{n+2} - p = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}.$$

Thus T is a biased estimator for p. This example shows that the bias may depend on the model parameters.

More generally, if the paramter space is higher dimensional, e.g. $\boldsymbol{\Theta} \subset \mathbb{R}^k$ for some $k > 1$, we may only be interested in $g(\boldsymbol{\theta})$, where $g : \boldsymbol{\Theta} \to \mathbb{R}$ is some function. The obvious extension of the definition of bias to this situation is $\mathsf{bias}_{\boldsymbol{\theta}}(T) = \mathsf{E}_{\boldsymbol{\theta}}(T) - g(\boldsymbol{\theta})$.

Example 5
$Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ independently with unknown parameter $\boldsymbol{\theta} = (\mu, \sigma^2) \in \boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$. If we are only interested in the mean μ, we may use $g(\boldsymbol{\theta}) = \mu$.

Example 6
For a random sample $Y_1, \dots, Y_n$ with $\mathrm{E}\, Y_i = \mu$ and $\mathrm{Var}\, Y_i = \sigma^2$ with μ, σ^2 unknown:

- $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is unbiased for $\mu = \mathrm{E}\, Y$.
  Indeed,
  $$\mathrm{E}(\bar{Y}) = \mathrm{E}(\frac{1}{n} \sum_{i=1}^{n} Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\, Y_i = \mathrm{E}\, Y = \mu$$
- $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ is unbiased for $\sigma^2 = \mathrm{Var}\, Y$.
  Indeed, $\sum (Y_i - \bar{Y})^2 = \sum_i Y_i^2 - \frac{1}{n} \underbrace{\sum_{i,j} Y_i Y_j}_{\sum_{i=j} \cdots + \sum_{i \neq j} \cdots} = (1 - \frac{1}{n}) \sum_i Y_i^2 - \frac{1}{n} \sum_{i \neq j} Y_i Y_j$
  and hence, $\mathrm{E}\, s^2 = \frac{1}{n-1} \Big(\frac{n-1}{n} \sum_i \mathrm{E}\, Y_i^2 - \frac{1}{n} \sum_{i,j} \underbrace{\mathrm{E}\, Y_i Y_j}_{= \mathrm{E}\, Y_i \, \mathrm{E}\, Y_j} \Big) = \mathrm{E}\, Y^2 - (\mathrm{E}\, Y)^2 = \sigma^2$

Thus, $(\bar{Y}, s^2)$ is an unbiased estimator of (μ, σ^2).

However, in general: $\bar{Y}^2$ is not unbiased for μ^2 and s is not unbiased for σ (see Problem Sheet).

Remark T unbiased for θ does not imply $h(T)$ unbiased for $h(\theta)$.

2.1.2 Standard Error

The bias measures how far the center of the sampling distribution of T is from the true parameter θ. A natural measure of spread for an estimator is its standard deviation.

Definition 3
Let T be an estimator for $\theta \in \Theta \subset \mathbb{R}$. The standard error of T is the standard deviation of the sampling distribution of T. Therefore, $\mathrm{SE}_\theta(T) = \sqrt{\mathrm{Var}_\theta(T)}$.

We will later see that many confidence intervals and test statistics involve the (approximate) standard error of an estimator T.

Example 7
Let $X_1, \ldots, X_n$ be i.i.d. with mean $\mathrm{E}(X_1) = \mu$, variance $\mathrm{Var}(X_1) = \sigma^2$ and define $\bar{X} = n^{-1} \sum_{i=1}^n X_i$. The standard error of the sample mean is $\mathrm{SE}(\bar{X}) = \sigma/\sqrt{n}$. We estimate this by $\mathrm{SE}(\bar{X}) = \sqrt{S^2}/\sqrt{n}$ where S^2 is the unbiased sample variance.

2.1.3 Mean Square Error

Definition 4
Let T be an estimator for $\theta \in \Theta \subset \mathbb{R}$. The mean square error of T is defined as

$$\mathrm{MSE}_\theta(T) = \mathrm{E}_\theta(T - \theta)^2.$$

Lemma 1
$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + (\mathrm{bias}_\theta(T))^2$

Proof Problem sheet (though this should be revision!).

Remark The MSE is a good criterion for selection an estimator as it takes both the bias and the variance of an estimator into account.

The following example shows that a biased estimator can outperform an unbiased estimator.

Example 8
$X \sim$ Binomial(n,p). n is known, the unknown parameter is $p \in [0,1]$.

Consider the two estimators $S = X/n$ and $T = \frac{X+1}{n+2}$.

We already know that S is unbiased for p and that $\text{bias}_p\, T = \frac{1-2p}{n+2}$.

Thus $\text{MSE}_p(S) = \text{Var}_p\, S = \frac{1}{n^2}\,\text{Var}_p\, X = p(1-p)/n$.

Furthermore, $\text{Var}_p\, T = \frac{1}{(n+2)^2}\,\text{Var}_p\, X = \frac{np(1-p)}{(n+2)^2}$ and thus $\text{MSE}_p(T) = \text{Var}_p(T) + \text{bias}_p(T)^2 = \frac{np(1-p)+(1-2p)^2}{(n+2)^2}$.

For $p = 0$ and $p = 1$, $\text{MSE}_p(T) = \frac{1}{(n+2)^2} > 0 = \text{MSE}_p(S)$
However, for $p = \frac{1}{2}$, $\text{MSE}_{\frac{1}{2}}(T) = \frac{n}{4(n+2)^2} < \frac{n}{4n^2} = \frac{1}{4n} = \text{MSE}_{\frac{1}{2}}(S)$

Since $\text{MSE}_p(T)$ and $\text{MSE}_p(S)$ are quadratic in p, this implies that

$$\exists 0 < p_1 < p_2 < 1$$

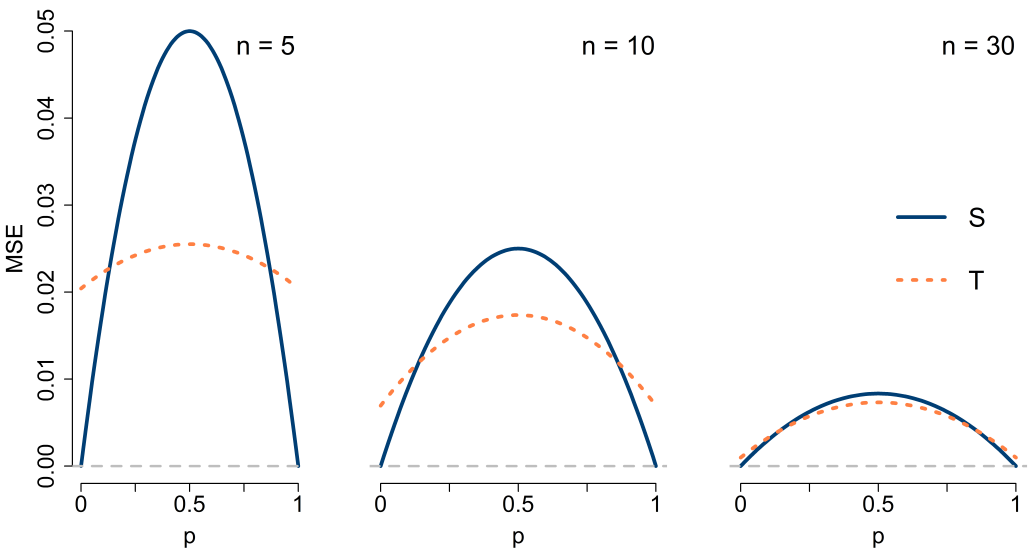
such that

$$\forall p \in (p_1, p_2) : \text{MSE}_p(T) < \text{MSE}_p(S)$$

and

$$\forall p \in [0, p_1) \cup (p_2, 1], \text{MSE}_p(T) > \text{MSE}_p(S).$$

The figure below illustrates this behavior for $n = 5, 10, 30$:


n = 5
n = 10
n = 30
MSE
0.00
0.01
0.02
0.03
0.04
0.05
0
0.5
1
p
S
T

3 The Cramér-Rao Lower Bound

When we choose between estimators, we would ideally know how the best possible estimator would perform. Unfortunately, minimizing criteria such as MSE across all parameter values is typically impossible. This means that we must restrict our attention to a set of reasonable estimators before we can discuss optimality. The set of unbiased estimators is one sensible such choice.

In fact, under weak conditions, we can derive a lower bound on the variance of an unbiased estimator (and hence its MSE). Regularity conditions will not be looked at in detail in this course - see separate course on Statistical Theory. We will only consider the case of unbiased estimators in which the parameter space is one-dimensional. There exist generalisations to biased estimators and to higher-dimensional situations.

Theorem 1 (Cramér-Rao lower bound)
Suppose $T = T(X)$ is an unbiased estimator for $\theta \in \Theta \subset \mathbb{R}$ based on $X = (X_1, \ldots, X_n)$ with joint pdf $f_\theta(x)$. Under mild regularity conditions,

$$\mathrm{Var}_\theta(T) \geq \frac{1}{I(\theta)},$$

where

$$I(\theta) = \mathrm{E}_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log f_\theta(X) \right\}^2 \right]$$

is the *Fisher information* of the sample.

The Fisher information can also be written as

$$I(\theta) = -\mathrm{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

Remark Suppose $X_1, \ldots, X_n$ are a random sample. Then

$$f_\theta(x) = \prod_{i=1}^{n} f_\theta^{(1)}(x_i),$$

where $x = (x_1, \ldots, x_n)$ and $f_\theta^{(1)}$ is the pdf/pmf of a single observation. This implies

$$I_f(\theta) = -\mathrm{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right) = \sum_{i=1}^{n} -\mathrm{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta^{(1)}(X_i) \right) = n I_{f^{(1)}}(\theta).$$

Thus for a random sample, the Fisher information is proportional to the sample size.

Example 9 (Sample Proportions)
$X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$ independently, $\theta \in \Theta = (0, 1)$.
Want to compute $I_f(\theta)$. We are dealing with a random sample thus $I_f(\theta) = nI_{f^{(1)}}(\theta)$.

Then $f_\theta^{(1)}(0) = \mathsf{P}_\theta(X_1 = 0) = 1 - \theta$, $f_\theta^{(1)}(1) = \mathsf{P}_\theta(X_1 = 1) = \theta$. Putting this together gives

$$f_\theta^{(1)}(x) = \theta^x(1 - \theta)^{1-x}.$$

Then

$$\frac{\partial}{\partial\theta} \log f_\theta^{(1)}(x) = \frac{\partial}{\partial\theta}(x \log\theta + (1 - x)\log(1 - \theta)) = \frac{x}{\theta} - \frac{1 - x}{1 - \theta}$$

and

$$\left(\frac{\partial}{\partial\theta}\right)^2 \log f_\theta^{(1)}(x) = -\frac{x}{\theta^2} - \frac{1 - x}{(1 - \theta)^2}$$

Hence,

$$\begin{aligned} I_{f^{(1)}}(\theta) &= -\mathsf{E}_\theta[\left(\frac{\partial}{\partial\theta}\right)^2 \log f_\theta^{(1)}(X_1)] = \frac{\mathsf{E}_\theta X_1}{\theta^2} + \frac{1 - \mathsf{E}_\theta X_1}{(1 - \theta)^2} \\ &= \frac{\theta}{\theta^2} + \frac{1 - \theta}{(1 - \theta)^2} = \frac{1}{\theta} + \frac{1}{1 - \theta} = \frac{1}{\theta(1 - \theta)} \end{aligned}$$

Thus $I_f(\theta) = nI_{f^{(1)}}(\theta) = \frac{n}{\theta(1-\theta)}$.

Hence, for any unbiased estimator T for θ,

$$\mathsf{Var}_\theta\, T \geq \frac{(1 - \theta)\theta}{n}.$$

Consider $S = \frac{1}{n}\sum_{i=1}^n X_i$.

$$\mathsf{Var}(S) = \frac{1}{n^2} n \,\mathsf{Var}\, X_1 = \frac{1}{n}\theta(1 - \theta).$$

Hence, S has minimal variance among all unbiased estimators for θ.

Proof (Sketch of the derivation of the Rao-Cramer bound.)
Using the Cauchy-Schwarz inequality $[(\mathsf{E}\, YZ)^2 \leq \mathsf{E}\, Y^2\, \mathsf{E}\, Z^2$ for square integrable $Y, Z]$,

$$\begin{aligned}\mathsf{Var}_\theta(T) I_f(\theta) &= \mathsf{E}_\theta[(T - \mathsf{E}_\theta\, T)^2]\, \mathsf{E}_\theta[(\frac{\partial}{\partial\theta} \log f_\theta(X))^2] \\ &\geq \left(\mathsf{E}_\theta\left[(T - \mathsf{E}_\theta\, T)\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right)^2\end{aligned}$$

$$\begin{aligned}\mathsf{E}_\theta\left[(T - \mathsf{E}_\theta\, T)\frac{\partial}{\partial\theta}\log f_\theta(X)\right] &= \mathsf{E}_\theta\left[(T - \mathsf{E}_\theta\, T)\frac{\frac{\partial}{\partial\theta} f_\theta(X)}{f_\theta(X)}\right] \\ &= \int (T(x) - \mathsf{E}_\theta\, T))\, \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ &= \int T(x)\frac{\partial}{\partial\theta} f_\theta(x) dx - \int \mathsf{E}_\theta\, T \frac{\partial}{\partial\theta} f_\theta(x) dx \\ &= \frac{\partial}{\partial\theta}\int T(x) f_\theta(x) dx - \mathsf{E}_\theta\, T \frac{\partial}{\partial\theta}\int f_\theta(x) dx \\ &= \frac{\partial}{\partial\theta}\,\mathsf{E}_\theta(T) - 0 \\ &= \frac{\partial}{\partial\theta}\theta = 1\end{aligned}$$

Thus,

$$\mathsf{Var}_\theta(T) \geq \frac{1}{I_f(\theta)}$$

Proof (Sketch of proof that the two formulation of the Fisher information are the same)
We need to show $\mathsf{E}_\theta[(\frac{\partial}{\partial\theta}\log f_\theta(X))^2] = -\,\mathsf{E}_\theta[\left(\frac{\partial}{\partial\theta}\right)^2 \log f_\theta(X)]$.

4 Asymptotic Properties

Many estimators do not admit straightforward comparison at a fixed sample size n. This may be the case if the bias and variance involve complex calculations, or are not even available in closed form. We can often simplify our comparisons by using approximations based on large sample (asymptotic) theory. Thus, we now compare the performance of *sequences* of estimators as the sample size n increases.

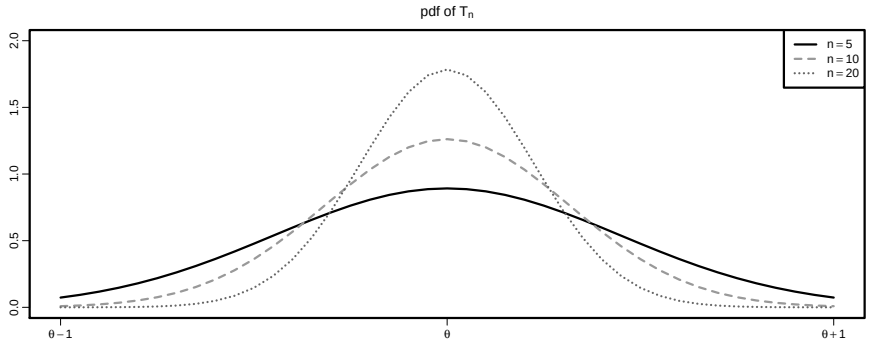
To motivate what large sample properties are to be evaluated, consider the familiar mean of a gaussian random sample.

Example 10
$Y_1, Y_2, \cdots \sim N(\theta, 1)$ independent, $\theta \in \mathbb{R}$.
Consider $T_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as estimator for θ. Then

$$\mathsf{E}\, T_n = \mathsf{E}\, \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\, Y_i = \theta$$

$$\mathsf{Var}\, T_n = \frac{1}{n^2}\,\mathsf{Var} \sum_{i=1}^{n} Y_i = \frac{1}{n^2}\sum_{i=1}^{n} \mathsf{Var}\, Y_i = \frac{1}{n}$$

As the sum of independent normally distributed r.v. is normally distributed, this implies $T_n \sim N(\theta, \frac{1}{n})$.



As n increases, the estimator T_n becomes more precise, it fluctuates less around the true value θ.

A formal concept for an estimator that gets "perfect" as $n \to \infty$ is the following.

Definition 5
A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called *(weakly) consistent* if for all $\theta \in \Theta$:

$$T_n \xrightarrow{\mathsf{P}_\theta} g(\theta) \quad (n \to \infty)$$

Recall that $T_n \xrightarrow{\mathsf{P}_\theta} g(\theta) \quad (n \to \infty)$ denotes convergence in probability and is defined by:

$$\forall \epsilon > 0 : \quad \lim_{n\to\infty} \mathsf{P}_\theta(|T_n - g(\theta)| < \epsilon) = 1$$

Usually: T_n depends only on $Y_1, \dots, Y_n$.

Lemma 2 (Portmanteau Lemma)
Let X and X_n, $n \in \mathbb{N}$, be real valued random variables. The following statements are equivalent:

- $X_n \xrightarrow{d} X$, as $n \to \infty$,
- $E[f(X_n)] \to E[f(X)]$as $n \to \infty$ for all bounded and continuous funtions $f : \mathbb{R} \to \mathbb{R}$.

A consistent estimator has a high probability of being close to the true parameter value. Here, closeness is defined as the euclidean distance. Showing consistency via the definition can be tedious! Hence, we will introduction a simple *sufficient* condition for consistency.

Definition 6
A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ for $g(\theta)$ is called *asymptotically unbiased* if for all $\theta \in \Theta$:

$$\mathsf{E}_\theta(T_n) \to g(\theta) \quad (n \to \infty)$$

Lemma 3
Suppose (T_n) is asymptotically unbiased for $g(\boldsymbol{\theta})$ and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\mathrm{Var}_{\boldsymbol{\theta}}(T_n) \to 0 \quad (n \to \infty).$$

Then (T_n) is consistent for $g(\boldsymbol{\theta})$.

Proof Recall Markov's inequality: $\mathrm{P}(|X| \geq a) \leq \frac{\mathrm{E}|X|}{a}$ for $a > 0$.

[Proof: $a\,\mathrm{I}(|X| \geq a) \leq |X|$. Hence, $\mathrm{P}(|X| \geq a) = \mathrm{E}\,\mathrm{I}(|X| \geq a) \leq \frac{1}{a}\mathrm{E}|X|$]

Applying Markov's inequality, we have

$$\forall \epsilon > 0: \quad \mathrm{P}_\theta(|T_n - g(\theta)| \geq \epsilon) = \mathrm{P}_\theta((T_n - g(\theta))^2 \geq \epsilon^2) \leq \frac{\mathrm{E}_\theta(T_n - g(\theta))^2}{\epsilon^2} = \frac{\mathrm{MSE}_\theta(T_n)}{\epsilon^2}$$
$$= \frac{1}{\epsilon^2}(\underbrace{\mathrm{Var}_\theta\, T_n}_{\to 0} + (\underbrace{\mathrm{E}_\theta\, T_n - g(\theta)}_{\to 0})^2)$$

Example 11 (Sample Proportions)
$X_i \sim \mathrm{Bernoulli}(\theta)$, $\theta \in \Theta = [0, 1]$.
$T_n(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$
$\left.\begin{array}{l} E_\theta\, T_n(X_1, \ldots, X_n) = \theta \quad \forall \theta \in \Theta \\ \mathrm{Var}_\theta(T_n) = \frac{1}{n}\theta(1-\theta) \to 0 \quad (n \to \infty) \end{array}\right\} \implies T_n$ consistent

Consistency is only a minimal requirement for an estimator. To derive valid hypothesis tests and confidence intervals we need the sampling distribution of our estimators.

Example 12
Returning to the example at the beginning of the subsection where $Y_1, Y_2, \cdots \sim N(\theta, 1)$. We have shown $T_n \sim N(\theta, \frac{1}{n})$. Rephrasing this gives

$$\sqrt{n}(T_n - \theta) \sim N(0, 1)$$

The distribution of many estimators cannot be computed as nicely as in the above example. However, the distribution of many estimators can be approximated by a normal distribution. The following property holds true for many estimators.

Definition 7
A sequence T_n of estimators for $\theta \in \mathbb{R}$ is called *asymptotically normal* if

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$$

for some $\sigma^2(\theta)$.

The following important result (that you have encountered before) demonstrates that sample averages are asymptotically normal.

Theorem 2 (Central Limit Theorem)
Let $Y_1, \ldots, Y_n$ be iid random variables with $\mathrm{E}\, Y_i = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$. Then the sequence $\sqrt{n}(\bar{Y} - \mu)$ converges in distribution to a $N(0, \sigma^2)$ distribution.

Example 13 (Sample Proportions)
$Y_1, Y_2, \cdots \sim$ Bernoulli(θ), independent, $\theta \in \Theta = [0, 1]$.
Consider the estimator $T_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ for θ.
Using the Central Limit Theorem,

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \theta) \xrightarrow{d} N(0, \theta(1-\theta)).$$

Thus T_n is asymptotically normal. [Recall: $Var(Y_1) = \theta(1-\theta)$]

Remark Under some mild regularity conditions, the standard error of an asymptotically normal estimator T_n can be approximated by $\mathrm{SE}_\theta(T_n) \approx \sigma(T_n)/\sqrt{n}$.

While we are often interested in estimators other than the sample mean, many estimators can be closely related to such averages. The following result, stated without proof, is typically called *Slutsky's lemma*.

Lemma 4 (Slutsky)
Let X_n, X and Y_n be random variables (or vectors). If $X_n \to_d X$ and $Y_n \to_p c$ for a constant c, then

(i) $X_n + Y_n \to_d X + c$;

(ii) $Y_n X_n \to_d cX$;

(iii) $Y_n^{-1} X_n \to_d c^{-1} X$ provided $c \neq 0$.

Example 14 (Sample Proportions)
Suppose that $X \sim \text{Binomial}(n, p)$ where p is unknown. The estimator sequence $T_n = \frac{X+1}{n+2}$ is asymptotically normal:

$$\sqrt{n}(T_n - p) \to_d N(0, p(1-p)).$$

To see this, write T_n as the sum of $A_n(X/n) + B_n$ for scalar sequences A_n and B_n such that $A_n \to 1$ and $B_n \to 0$ as $n \to \infty$. To finish your proof, apply Slutsky's lemma twice.

What does this suggest about the use of T_n and X/n as estimators of p for large sample sizes?

Another important tool for deriving the sampling distribution of more complex estimators is the so-called delta method.

Theorem 3 (Delta Method)
Suppose that T_n is an asymptotically normal estimator of θ with

$$\sqrt{n}(T_n - \theta) \to_d N(0, \sigma^2(\theta)).$$

Let $g : \Theta \to \mathbb{R}$ be a differentiable function with $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(T_n) - g(\theta)) \to_d N(0, g'(\theta)^2 \sigma^2(\theta)).$$

Example 15 (Sample Odds)
Recall that the odds of an event are defined as $P(A)/(1-P(A))$. Let $Y_1, \ldots, Y_n$ be iid Bernoulli(p) random variables. We consider estimating the odds that $Y_i = 1$.

We have that $X \sim \text{Binomial}(n, p)$. Let $S_n = X/n$. We know that

$$\sqrt{n}(S_n - p) \to_d N(0, p(1-p)).$$

The sample odds are $T_n = \frac{S_n}{1-S_n} = g(S_n)$ for the function $g(s) = s/(1-s)$. The first derivative of g is $g'(s) = (1-s)^{-2}$. By the delta method

$$\sqrt{n}(T_n - p/(1-p)) = \sqrt{n}(g(S_n) - g(p)) \to_d N(0, (1-p)^{-3}p).$$

Theorem 4 (Continuous Mapping Theorem)
Let $k, m \in \mathbb{N}$ and let X and X_n, $n \in \mathbb{N}$, be $\mathbb{R}^k$-valued random variables (i.e. random vectors of dimension k). Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a function continuous at every point of a set C such that $P(X \in C) = 1$.

- If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$, as $n \to \infty$,
- If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$, as $n \to \infty$,
- If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$, as $n \to \infty$.

5 Maximum Likelihood Estimation

One common method of finding an estimator for $\boldsymbol{\theta}$ is maximum likelihood. This method is very widely applicable, and variations on maximum likelihood remain important for current research.

Recall that the main idea behind maximum likelihood estimation is to find the parameter value for which the observed data is most "likely" and report this as our point estimate. For discrete distributions, we can directly consider probabilities based on the mass function.

Example 16
Suppose we observe the number of successes of 10 independent trials (outcomes: 1=success, 0=failure) and we know that the success probability θ is an element of $\Theta = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$.

Model: $X \sim \text{Binomial}(10, \theta)$, $\theta \in \Theta$.

The probability of a given outcome is

$$P_\theta(X = x) = \binom{10}{x} \theta^x (1-\theta)^{10-x}, \quad x = 0, \ldots, 10.$$

The following table shows the probability for all outcomes for all possible parameter values, i.e. $\mathrm{P}_\theta(X = x)$. The model specifies every row of the table.

	x=0	x=1	x=2	x=3	x=4	x=5	x=6	x=7	x=8	x=9	x=10
$\theta = 0$	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\theta = 0.25$	0.06	0.19	0.28	0.25	0.15	0.06	0.02	0.00	0.00	0.00	0.00
$\theta = 0.5$	0.00	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.00
$\theta = 0.75$	0.00	0.00	0.00	0.00	0.02	0.06	0.15	0.25	0.28	0.19	0.06
$\theta = 1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

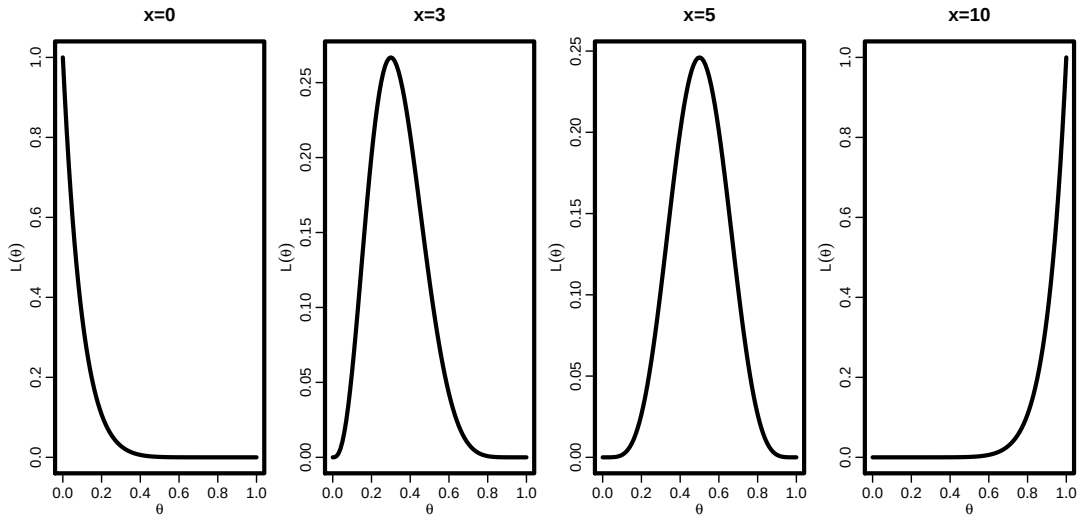
In any given experiment we observe a specific outcome x. The maximum likelihood estimate is the value that maximises $P_\theta(X = x)$. For example, if we observe $x =$ then the maximum likelihood estimate is $\hat{\theta} = 0.25$, if we observe $x = 4$ the maximum likelihood estimate is $\hat{\theta} = 0.5$.

Example 17
We now make the previous more realistic and use the parameter space $\Theta = [0, 1]$. Suppose we observe $x = 5$ successes. The function

$$L : \Theta \to [0, \infty), L(\theta) = P_\theta(X = 5) = \binom{10}{5}\theta^5(1-\theta)^5$$

is called the likelihood function. The maximum likelihood estimate is the maximiser of L, which in this case turns out to be $\hat{\theta} = 0.5$.

The following figure contains the likelihood function for several different outcomes x.



In general, suppose we observe the random object $\mathbf{Y}$ with realisation $\mathbf{y}$. The likelihood function is

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y}) = \begin{cases} \mathrm{P}(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}), & \text{discrete data} \\ f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}), & \text{absolutely continuous data.} \end{cases}$$

Thus the likelihood function is the *joint* pdf/pmf of the observed data considered as function of the unknown parameter.

Consider the case of a random sample, i.e. $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $Y_1, \ldots, Y_n$ are iid. If

the model specifies that Y_i has pdf $f(\cdot;\boldsymbol{\theta})$ then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta}).$$

Definition 8
A *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$ is an estimator $\hat{\boldsymbol{\theta}}$ s.t.

$$L(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}).$$

Usually, the MLE is well defined. However, one can construct situations in which it does not exist or is not unique.

Example 18 (Poisson distribution)
$X_1, \dots, X_n \sim$Poisson(θ) independent, $\theta > 0$. Then

$$L(\theta) = \mathrm{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P_\theta(X_i = x_i) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta}$$

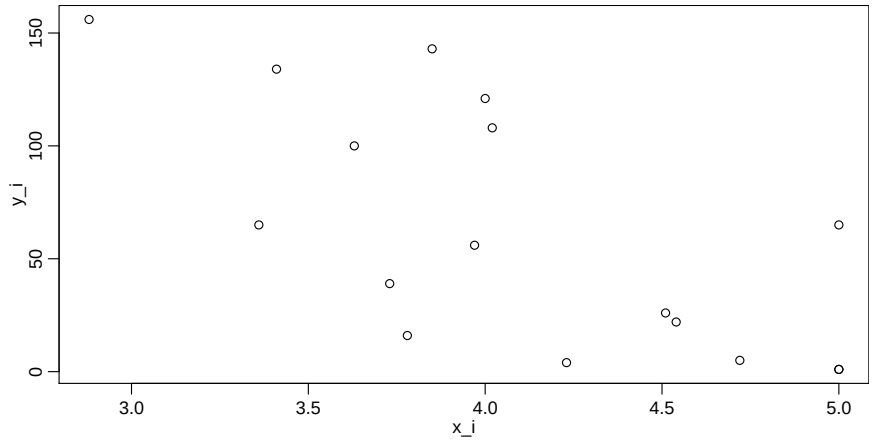
and $\log L(\theta) = \sum_{i=1}^{n}[x_i \log(\theta) - \log(x_i!) - \theta]$. Differentiating and equation to 0 gives

$$\frac{\partial}{\partial\theta} \log L(\theta) = \sum_{i=1}^{n}[x_i/\theta - 1] = \frac{1}{\theta}\sum_{i=1}^{n} x_i - n \stackrel{!}{=} 0.$$

This gives $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as candidate for a maximum likelihood estimator. You need to check that $\hat{\theta}$ is actually a maximiser - we omit this here. Once this is done we know that $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the maximum likelihood estimate and $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimator.

Example 19 (Survival of Leukemia Patients)
This data contains the survival time y_i (in weeks) and $x_i = \log_{10}$(initial white blood cell count) for 17 leukemia patients.

x_i	3.36	2.88	3.63	3.41	3.78	4.02	4	4.23	3.73	3.85	3.97	4.51	4.54	5	5	4.72	5
y_i	65	156	100	134	16	108	121	4	39	143	56	26	22	1	1	5	65



We choose to treat $x_1, \ldots, x_n$ as constants (we model Y_i conditional on the observed x_i). Suppose we use the model

$$Y_i = \alpha \exp(\beta(x_i - \bar{x}))\epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon_i \sim Exp(1)$ iid and $\boldsymbol{\theta} = (\alpha, \beta)^T \in (0, \infty) \times \mathbb{R}$.

To work out the MLE: First, note that $Y_i \sim Exp(\lambda_i)$ with $\lambda_i = \frac{1}{\alpha \exp[\beta(x_i - \bar{x})]}$. Furthermore, the pdf of Y_i is $f_{Y_i}(y_i) = \lambda_i e^{-\lambda_i y_i}$. Hence,

$$L(\theta) = \prod_i \lambda_i e^{-\lambda_i y_i} = \frac{1}{\alpha^n} e^{-\beta \sum (x_i - \bar{x}) - \frac{1}{\alpha} \sum y_i \exp[-\beta(x_i - \bar{x})]}$$

$$\log L(\theta) = -n \log \alpha - \beta \sum (x_i - \bar{x}) - \frac{1}{\alpha} \sum y_i \exp(-\beta(x_i - \bar{x}))$$

Differentiate wrt α, β and solve numerically (or optimise numerically). This can be done in R by running the code:

```
> dat <- read.csv("leuk.dat")
> g <- function(par){
      -sum(log(dexp(dat$y,1/(par[1]*exp(par[2]*(dat$x-mean(dat$x)))))))
}
> fit <- optim(g,par=c(1,1),hessian=TRUE)
> fit$par

[1] 51.086326 -1.110557
```

Remark In fact, this data set is a bit unrealistic. Survival time is observed for all patients. In practice such a study may take a long time. At the time the study is evaluated, for many patients it will only be known that they did not die before a given time (their observation time is said to be *censored*). The third year module on *Survival Models* covers many methods appropriate for censored data.

The following example demonstrates that MLEs can fail to be unbiased.

Example 20 (MLEs are not necessarily unbiased)
$Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ iid with μ and σ^2 unknown. Then the MLEs are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ (Check). $\hat{\mu}$ is unbiased but $\hat{\sigma}^2$ is not:

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n}\mathrm{E}(\sum(Y_i - \bar{Y})^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

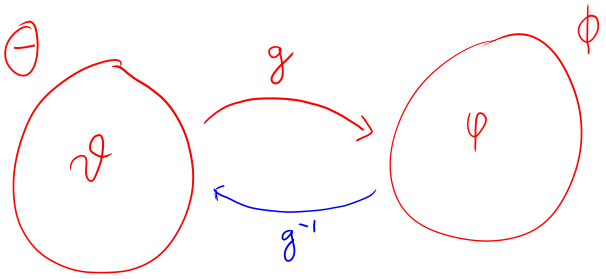
It is straightforward to derive an MLE, but we need to understand why we would prefer to use the MLE over other estimators (especially if they may be biased). We will see in the next section that the MLE is asymptotically well behaved.

5.1 Properties of Maximum Likelihood estimators

5.1.1 MLEs are functionally invariant

In this section we show that MLEs are invariant under reparametrisations of the parameter space. More precisely:

If g is a bijective function and if $\hat{\theta}$ is an MLE of θ, then $\hat{\phi} = g(\hat{\theta})$ is an MLE of $\phi = g(\theta)$.



This can be seen as follows: Suppose $g : \Theta \to \Phi$. Since g is bijective, g has an inverse, which we denote by g^{-1}.
We will denote the likelihood under the new parametrisation by $\tilde{L}$. It is now (using the notation for the discrete observation case, assuming that the observation is x),

$$\tilde{L}(\phi) = \mathsf{P}_\phi(X = x) = \mathsf{P}_{g^{-1}(\phi)}(X = x) = L(g^{-1}(\phi))$$

To see that $\hat{\phi} = g(\hat{\theta})$ maximises $\tilde{L}$: For all $\phi \in \Phi$,

$$\tilde{L}(\hat{\phi}) = L(g^{-1}(\hat{\phi})) = L(g^{-1}(g(\hat{\theta})) = L(\hat{\theta}) \geq L(g^{-1}(\phi)) = \tilde{L}(\phi)$$

Thus $\hat{\phi}$ maximises $\tilde{L}$.

> **Example 21**
> $Y_1, \ldots, Y_n \sim N(\theta, 1)$ iid, $\theta \in \Theta = \mathbb{R}$
> The MLE of θ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
>
> What is the MLE of $\phi = \theta + 2$?
> $\phi = \theta + 2 =: g(\theta)$. Clearly, $g : \Theta \to \mathbb{R} = \Phi$ is bijective
> Hence, $\hat{\phi} := g(\hat{\theta}) = \hat{\theta} + 2 = \frac{1}{n}\sum_{i=1}^{n} Y_i + 2$ is the MLE of ϕ.

Remark What if g is not bijective?

Recall: Let $f : A \to B$ be a function. It is injective iff $\forall a_1, a_2, \in A : f(a_1) = f(a_2) \implies a_1 = a_2$. It is surjective iff $\forall b \in B \exists a \in A : f(a) = b$. It is bijective if it is both injective and surjective.

If ϕ is **not surjective** then there are ϕs that are not in the range of g, implying that for these parameter values no model is defined. It does not really make sense to speak

of the likelihood of these parameter values, so one should set the likelihood for these to the lowest possible value (which is 0). Doing this, one can easily argue that the invariance of the MLE under the transformation induced by g is retained.

If ϕ is **not injective** then knowing ϕ does not uniquely identify the parameter θ or the model. One way to define the likelihood on Φ is the "induced" likelihood function

$$\tilde{L} : \mathbb{R} \to \mathbb{R}, \tilde{L}(\phi) = \sup\{L(\theta) : g(\theta) = \phi\}.$$

This gives every $\phi \in \Phi$ the highest likelihood of all θ that g maps onto it.

With this definition the invariance is retained: If $\hat{\theta}$ is the MLE then $\hat{\phi} = g(\hat{\theta})$ maximises the induced likelihood function $\tilde{L}$.

Example 22
(continued from previous example)
Consider $g : \mathbb{R} \to [0, \infty) = \Phi, g(\theta) = \theta^2$ is not bijective.
Then

$$\begin{aligned}\tilde{L}(\phi) &= \sup\{L(\theta) : g(\theta) = \phi\} = \sup\{L(\theta) : \theta^2 = \phi\} \\ &= \sup\{L(\theta) : \theta \in \{-\sqrt{\phi}, \sqrt{\phi}\}\} \\ &= \max(L(-\sqrt{\phi}), L(\sqrt{\phi}))\end{aligned}$$

5.1.2 Large sample properties

In most cases, MLEs are defined as the solution to a system of equations. A natural question is then, "What happens as $n \to \infty$?"

Theorem 5
Let $X_1, X_2, \ldots$ be iid observations with pdf (or pmf) $f_\theta(x)$, where $\theta \in \Theta$ and Θ is an open interval. Let $\theta_0 \in \Theta$ denote the true parameter. Under regularity conditions (e.g. $\{x : f_\theta(x) > 0\}$ does not depend on θ), the following holds:

(i) There exists a *consistent* sequence $(\hat{\theta}_n)_{n\in\mathbb{N}}$ of maximum likelihood estimators. [$\hat{\theta}_n$ is an MLE based on $X_1, \ldots, X_n$].

(ii) Suppose $(\hat{\theta}_n)_{n\in\mathbb{N}}$ is a consistent sequence of MLEs. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_0))^{-1}),$$

where $I_f(\theta) = \mathrm{E}_\theta[(\frac{\partial}{\partial\theta}\log f_\theta(X))^2]$ is the *Fisher Information* of a sample of size 1.

Remark (i) implies: if the MLE is unique (for every n) then the sequence of MLEs is consistent.

Example 23
$X_1, X_2, \ldots \sim$ Bernoulli(θ) independently, $\theta \in (0,1)$.
We already shown that $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the unique MLE based on $X_1, \ldots, X_n$. Furthermore, we know $I_f(\theta) = \frac{1}{\theta(1-\theta)}$. Thus,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \theta_0(1-\theta_0)).$$

[We have previously shown this directly using the central limit theorem).]

In other words, the distribution of $\hat{\theta}$ can be approximated by an $N(\theta_0, \dfrac{\theta_0(1-\theta_0)}{n})$ distribution.

Remark The above theorem has a limiting distribution that depends on $I_f(\theta_0)$, which would not be known in practical situations. To use this result, we need to estimated $I_f(\theta_0)$.

In an iid sample, $I_f(\theta_0)$ can be estimated by

- $I_f(\hat{\theta})$
- $\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial}{\partial\theta}\log(f(x_i;\theta))|_{\theta=\hat{\theta}}\right)^2$
- $-\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial}{\partial\theta}\right)^2 \log(f(x_i;\theta))|_{\theta=\hat{\theta}}$

These estimators are often consistent, i.e. will converge to $I_f(\theta_0)$ in probability.

Remark Where it is required, the standard error of an asymptotically normal MLE $\hat{\theta}_n$ can be approximated by $\mathrm{SE}(\hat{\theta}_n) = \sqrt{\hat{I}_n^{-1}}/\sqrt{n}$. Here, $\hat{I}_n$ can be any of the estimators for $I_f(\theta_0)$ listed above.

Remark Multivariate version: Suppose $\Theta \subset \mathbb{R}^k$ is an open set and $\hat{\boldsymbol{\theta}}_\mathbf{n}$ is the MLE based on n observations. Then the equivalent result to the above theorem is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\mathbf{n} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, (I_f(\theta_0))^{-1}),$$

where $\boldsymbol{\theta}_0$ denotes the true parameter and $I_f(\boldsymbol{\theta})$ is the *Fisher Information Matrix* given by

$$I_f(\theta) := \mathrm{E}_\theta[(\nabla \log f(X;\boldsymbol{\theta}))^T(\nabla \log f(X;\boldsymbol{\theta}))] = -\mathrm{E}_{\boldsymbol{\theta}}[\nabla^T \nabla \log f(X;\boldsymbol{\theta})]$$

[∇ denotes the gradient wrt $\boldsymbol{\theta}$, i.e. the row vector of partial derivatives wrt the elements of $\boldsymbol{\theta}$.) Convergence in distribution for random vectors is defined as follows: Let $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ be random vectors of dimension k. Then $X_n \xrightarrow{d} X (n \to \infty)$ if $\mathrm{P}(\mathbf{X}_n \leq \mathbf{z}) \to \mathrm{P}(\mathbf{X} \leq \mathbf{z})$ $(n \to \infty)$ $\forall \mathbf{z} \in \mathbb{R}^k$ at which $\mathbf{z} \mapsto \mathrm{P}(\mathbf{X} \leq \mathbf{z})$ is continuous.]

5.2 Sketch of proofs

Proof (Sketch of the proof of (i) of Theorem 5:) The likelihood is $L(\theta) = \prod_{i=1}^n f_\theta(X_i)$. Let

$$S_n(\theta) := \frac{1}{n}\log L(\theta) = \frac{1}{n}\sum_{i=1}^n \underbrace{\log f_\theta(X_i)}_{=:Z_i}$$

$\hat{\theta}$ maximises $L(\theta) \Leftrightarrow \hat{\theta}$ maximises $S_n(\theta)$.

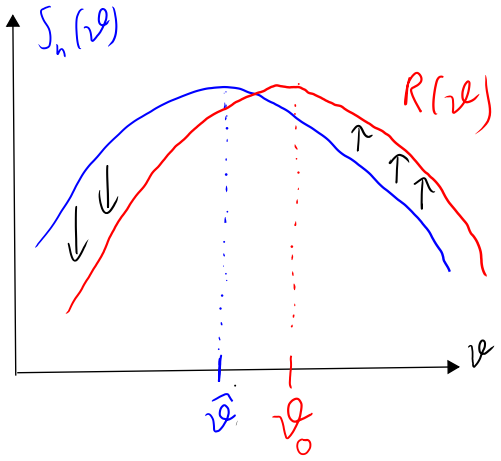
$Z_1, Z_2, \ldots$ are iid. Thus, $\forall\theta$, by the law of large numbers,

$$S_n(\theta) \to \mathrm{E}_{\theta_0}(Z_1) = \mathrm{E}_{\theta_0}[\log f_\theta(X_1)] =: R(\theta).$$

We next show that θ_0 maximises $R(\theta)$. Indeed, $\forall\theta$,

$$\begin{aligned} R(\theta) - R(\theta_0) &= \mathrm{E}_{\theta_0}[\log(\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)})] \leq \mathrm{E}_{\theta_0}[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} - 1] \quad (\text{using } \forall z > 0 : z - 1 \geq \log(z)) \\ &= \int \left(\frac{f_\theta(x)}{f_{\theta_0}(x)} - 1\right) f_{\theta_0}(x)dx = \int f_\theta(x)dx - \int f_{\theta_0}(x)dx = 1 - 1 = 0. \end{aligned}$$

Thus, we known that $S_n \to R$ pointwise, $\hat{\theta}_n$ maximises $S_n(\theta)$ and θ_0 maximises $R(\theta)$. This indicates that $\hat{\theta}_n \to \theta_0$. [For a formal proof see books on mathematical statistics.]



Proof (Sketch of the proof of (ii) of Theorem 5:) The MLE $\hat{\theta}$ satisfies $\frac{\partial}{\partial\theta} \log L(\theta)|_{\theta=\hat{\theta}} = 0$. Hence,

$$0 = \frac{1}{\sqrt{n}} \frac{\partial}{\partial\theta} \log L(\theta)|_{\theta=\hat{\theta}} = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i;\theta)}_{=:g(\theta)} |_{\theta=\hat{\theta}} = g(\hat{\theta}) \overset{\text{Taylor}}{=} g(\theta_0) + \frac{1}{\sqrt{n}} g'(\tilde{\theta}) \sqrt{n}(\hat{\theta} - \theta_0)$$

for some $\tilde{\theta}$ between $\hat{\theta}$ and θ_0. Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{g(\theta_0)}{-\frac{1}{\sqrt{n}} g'(\tilde{\theta})}.$$

For all θ, by the weak law of large numbers,

$$-\frac{1}{\sqrt{n}} g'(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\frac{\partial}{\partial\theta}\right)^2 \log f(X_i;\theta)}_{=:\eta_i} = -\frac{1}{n} \sum_{i=1}^{n} \eta_i \overset{P_{\theta_0}}{\to} -\mathrm{E}[\left(\frac{\partial}{\partial\theta}\right)^2 \log f(X_1;\theta)] = I_f(\theta)$$

as $\eta_1, \eta_2, \ldots$ are iid.

One can show that $\hat{\theta} \overset{P}{\to} \theta_0$ implies $\tilde{\theta} \overset{P}{\to} \theta_0$ and that

$$-\frac{1}{\sqrt{n}} g'(\tilde{\theta}) \overset{P}{\to} I_f(\theta_0).$$

Want to use CLT for getting the asymptotic distribution of $g(\theta_0)$: We have $g(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$, with $Y_i = \frac{\partial}{\partial\theta} \log f(X_i;\theta)|_{\theta=\theta_0}$,

$$\mathrm{E}_{\theta_0} Y_i = \mathrm{E}_{\theta_0} \frac{\partial}{\partial\theta} \log f = \mathrm{E}_{\theta_0} \frac{f'}{f} = \int \frac{f'}{f} f = \int f' = \frac{\partial}{\partial\theta} \underbrace{\int f}_{=1} = 0,$$

and

$$\mathrm{Var}_{\theta_0}(Y_i) = \mathrm{E}_{\theta_0}[Y_i^2] = I_f(\theta_0).$$

Hence, by the CLT: $g(\theta_0) \xrightarrow{d} N(0, I_f(\theta_0))$

By Slutsky's lemma, $\frac{g(\theta_0)}{-\frac{1}{\sqrt{n}}g'(\tilde{\theta})} \xrightarrow{d} N(0, \frac{1}{I_f(\theta_0)})$.

6 Confidence Regions

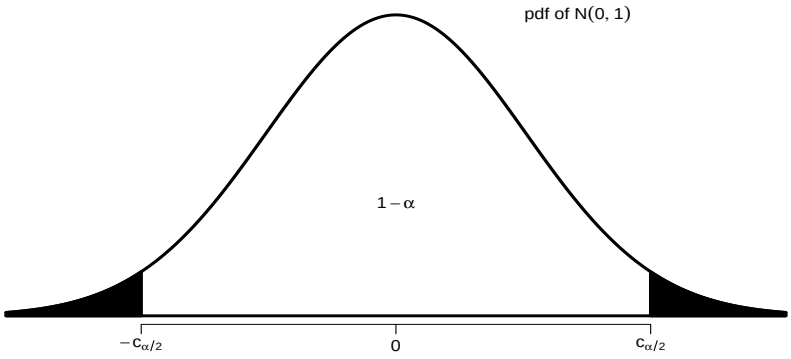
Point estimator: one number only.
Confidence interval: random interval that contains the true parameter with a certain probability.

Example 24
$Y_1, \dots, Y_n$ iid $N(\mu, \underbrace{\sigma_0^2}_{\text{known}})$ and $\mu \in \mathbb{R}$ is unknown.

Want: random interval that contains μ with probability $1-\alpha$ for some $\alpha > 0$, e.g. $\alpha = 0.05$

$\bar{Y} = \frac{1}{n}\sum Y_i \sim N(\mu, \sigma_0^2/n)$ Hence,

$$\frac{\bar{Y}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1) \quad \forall \mu \in \mathbb{R}.$$



Thus,

$$1-\alpha = \mathrm{P}(-c_{\alpha/2} < \frac{\bar{Y}-\mu}{\sigma_0/\sqrt{n}} < c_{\alpha/2}),$$

where $0 < \alpha < 1$ and $\Phi(c_{\alpha/2}) = 1-\alpha/2$. ϕ is the cdf of $N(0,1)$

Rewrite this as

$$1-\alpha = \mathrm{P}(\underbrace{\bar{Y} + c_{\alpha/2}\sigma_0/\sqrt{n}}_{\text{random}} > \underbrace{\mu}_{\text{non-random}} > \underbrace{\bar{Y} - c_{\alpha/2}\sigma_0/\sqrt{n}}_{\text{random}}]).$$

The interval $(\bar{Y} - c_{\alpha/2}\sigma_0/\sqrt{n}, \bar{Y} + c_{\alpha/2}\sigma_0/\sqrt{n})$ is a random interval which contains the true μ with probability $1 - \alpha$.

The observed value of the random interval is $(\bar{y} - c_{\alpha/2}\sigma_0/\sqrt{n}, \bar{y} + c_{\alpha/2}\sigma_0/\sqrt{n})$.

This is called a $1 - \alpha$ *confidence interval* for μ.

Remarks:

- α is usually small, often $\alpha = 0.05$. this is the usual convention
- When speaking of a confidence interval we can either mean the realisation of the random interval or the random interval itself (this should hopefully be clear from the context).
- Could use asymmetrical values, but symmetrical values ($\pm c_{\alpha/2}$) give the shortest interval in this case.
- The value $\sigma_0/\sqrt{n}$ is exactly the standard error of $\bar{Y}$.

Example 25
In an industrial process, past experience shows it gives components whose strengths are $N(40, 1.21^2)$. The process is modified but s.d.(=1.21) remains the same.

After modification, $n = 12$ components give an average of 41.125.

New strength $\sim N(\mu, 1.21^2)$.
$n = 12$, $\sigma_0 = 1.21$, $\bar{y} = 41.125$, $\alpha = 0.05$, $c_{\alpha/2} \approx 1.96$.
$\rightarrow$ a 95% CI for μ is $(40.44, 41.81)$.

This does *not* mean that we are 95% confident that the true μ lies in $(40.44, 41.81)$. It means that if we were to take an infinite number of (indep) samples then in 95% of cases the calculated CI would contain the true value.

Note that our CI does not include 40 - an indication that the modification seems to have increased strength ($\rightarrow$ hypothesis testing)

Definition 9
A $1-\alpha$ confidence interval for θ is a random interval I that contains the 'true' parameter with probability $\geq 1-\alpha$, i.e.

$$P_\theta(\theta \in I) \geq 1-\alpha \quad \forall \theta \in \Theta$$

In the above, I can be any type of interval. For example, if L and U are random variables with $L \leq U$ then I could be the open interval (L, U), the closed interval $[L, U]$, the unbounded interval $[L, \infty)$, ...

Example 26
$X \sim$ Bernoulli(θ) $\theta \in [0,1]$ unknown. Want: $1-\alpha$ CI for θ (suppose $0 < \alpha < 1/2$).

Let

$$[L, U] = \begin{cases} [0, 1-\alpha], & \text{for } X = 0 \\ [\alpha, 1], & \text{for } X = 1 \end{cases}$$

This is indeed a $1-\alpha$ CI, since

$$\mathrm{P}_\theta(\theta \in [L, U]) = \begin{cases} P_\theta(X=0) = 1-\theta \geq 1-\alpha & \text{for } \theta < \alpha, \\ 1 & \text{for } \alpha \leq \theta \leq 1-\alpha, \\ P_\theta(X=1) = \theta \geq 1-\alpha & \text{for } \theta > 1-\alpha. \end{cases}$$

Remark $L = -\infty$ and $U = \infty$ is allowed.

Example 27 (One-sided confidence interval)
Suppose $Y_1, \ldots, Y_n$ are independent measurements of a pollutant θ, where higher values indicate worse pollution. To be "on the safe side" when reporting the amoung of pollutant, we want a $1-\alpha$ CI for θ of the form $(-\infty, h(\mathbf{y})]$. For this h needs to be a function $h : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathrm{P}_\theta(\theta \leq h(\mathbf{Y})) = 1-\alpha \quad \forall \theta.$$

6.1 Construction of Confidence Intervals

Features of $\frac{\bar{Y}-\mu}{\sigma_0/\sqrt{n}}$ in the first example (where σ_0 is known):

1. it is a function of the unknown μ and the data only (σ_0 is known)
2. its distribution is *completely* known.

More generally, consider a situation, where we are interested in a (scalar) unknown parameter θ. There may be *nuisance* parameters (i.e. other unknown parameters we are not interested in).

> **Definition 10**
> A *pivotal quantity* for θ is a function $t(\mathbf{Y}, \theta)$ of the data and θ (and NOT any further nuisance parameters) s.t. the distribution of $t(\mathbf{Y}, \theta)$ is know, i.e. does NOT depend on ANY unknown parameters.

Suppose $t(\mathbf{Y}, \theta)$ is a pivotal quantity for θ. Then we can find constants a_1, a_2 s.t.

$$\mathrm{P}(a_1 \leq t(\mathbf{Y}, \theta) \leq a_2) \geq 1 - \alpha$$

because we know the distribution of $t(\mathbf{Y}, \theta)$. (there may be many pairs (a_1, a_2); $\geq$ is needed for discrete distributions)

In many cases (as above) we can rearrange terms to give

$$\mathrm{P}(h_1(\mathbf{Y}) \leq \theta \leq h_2(\mathbf{Y})) \geq 1 - \alpha$$

$[h_1(\mathbf{Y}), h_2(\mathbf{Y})]$ is a random interval. The observed interval

$$[\underbrace{h_1(\mathbf{y})}_{\text{lower confidence limit}}, \underbrace{h_2(\mathbf{y})}_{\text{upper confidence limit}}]$$

is a $1 - \alpha$ confidence interval for θ.

> **Example 28**
> $Y_1, \ldots, Y_n$ i.i.d $N(\mu, \sigma^2)$, μ, σ^2 both unknown

1. Want: confidence interval for μ.

   σ is unknown $\implies$ can't use $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ as a pivotal quantity;

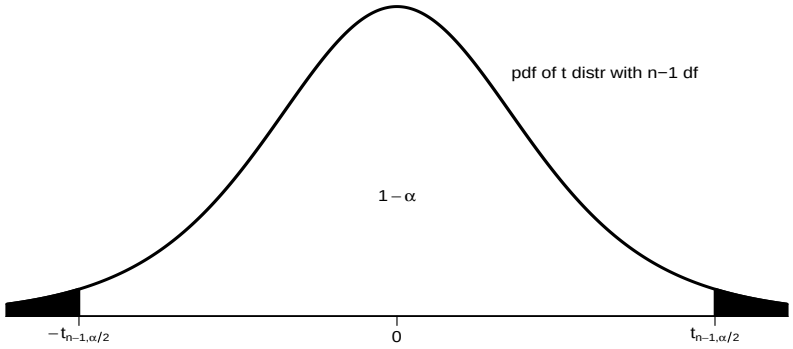
   Replace σ by S, where

$$S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2 \quad \text{(sample variance)}$$

   to give

$$T = \frac{\sqrt{n}}{S}(\bar{Y} - \mu).$$

   T follows a Student-t distribution with $n-1$ degrees of freedom. (You may have seen this previously, but we will prove more general results in the 2nd part of the course that includes this as a special case)

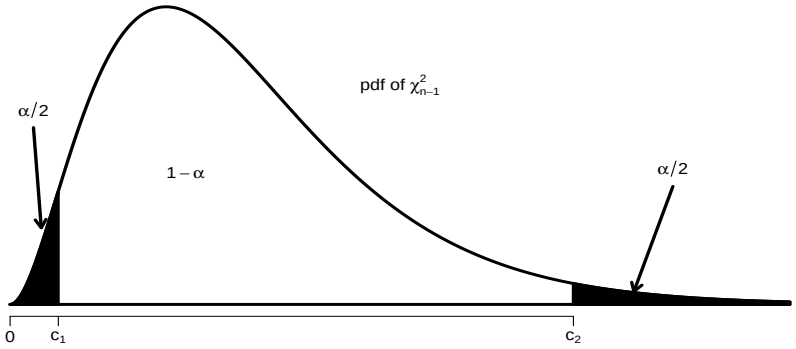


$$\begin{aligned} 1-\alpha &= \mathrm{P}(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) \\ &= \mathrm{P}(\bar{Y} - \frac{S}{\sqrt{n}}t_{n-1,\alpha/2} \le \mu \le \bar{Y} + \frac{S}{\sqrt{n}}t_{n-1,\alpha/2}) \end{aligned}$$

   $1-\alpha$ CI is $(\bar{y} - \frac{s}{\sqrt{n}}t_{n-1,\alpha/2}, \bar{y} + \frac{s}{\sqrt{n}}t_{n-1,\alpha/2})$

2. Want: confidence interval for σ (or σ^2).
   We will see that:

$$\frac{\sum(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$$



c_1 and c_2 such that

$$\mathrm{P}\left(c_1 \leq \frac{\sum(Y_i - \bar{Y})^2}{\sigma^2} \leq c_2\right) = 1 - \alpha$$

$\Longrightarrow$ a $1-\alpha$ CI for σ^2 is $\left(\frac{\sum(y_i-\bar{y})^2}{c_2}, \frac{\sum(y_i-\bar{y})^2}{c_1}\right)$ and a $1-\alpha$ CI for σ is $\left(\sqrt{\frac{\sum(y_i-\bar{y})^2}{c_2}}, \sqrt{\frac{\sum(y_i-\bar{y})^2}{c_1}}\right)$.

6.2 Asymptotic confidence intervals

Often, we only know $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta))$ (e.g. asymptotic distribution of the MLE). This implies $\sqrt{n}\frac{T_n-\theta}{\sigma(\theta)} \xrightarrow{d} N(0,1)$ and thus approximately:

$$\sqrt{n}\frac{T_n - \theta}{\sigma(\theta)} \sim N(0,1)$$

and we can use the LHS as a pivotal quantity. The resulting confidence interval is called *asymptotic confidence interval*.

Definition 11
A sequence of random intervals I_n is called an asymptotic $1-\alpha$ CI for θ if

$$\lim_{n\to\infty} \mathrm{P}_\theta(\theta \in I_n) \geq 1 - \alpha \quad \forall\theta.$$

Trying to construct an asymptotic CI directly from $\sqrt{n}\frac{T_n-\theta}{\sigma(\theta)}$ is usually not easy. This is because σ depends on θ and thus it may be difficult to solve the resulting inequalities for θ.

Simplification:
Suppose we have a consistent estimator $\hat{\sigma}_n$ for $\sigma(\theta)$. Thus, by definition, $\hat{\sigma}_n \xrightarrow{P_\theta} \sigma(\theta)$ for all θ. Hence, using the Slutsky lemma and the fact that $X \sim N(0, \sigma^2(\theta))$ implies $X/\sigma(\theta) \sim N(0,1)$,

$$\sqrt{n}\frac{T_n-\theta}{\hat{\sigma}_n} \xrightarrow{d} N(0,1).$$

Using the LHS as the pivotal quantity leads to the approximate confidence limits

$$T_n \pm c_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$$

where $\Phi(c_{\alpha/2}) = 1-\alpha/2$.

The quantity $\hat{\sigma}_n/\sqrt{n}$ in the simplification is an estimate of the standard error $\mathrm{SE}(T_n)$. Hence, we can also write the approximate confidence limits as

$$T_n \pm c_{\alpha/2}\,\mathrm{SE}(T_n).$$

Example 29
$Y \sim \text{Binomial}(n,\theta)$, $\theta \in (0,1)$ unknown. n is known.

Then $\sqrt{n}(Y/n-\theta) \xrightarrow{d} N(0,\theta(1-\theta))$ as $n\to\infty$. To see this use the CLT or the large sample properties of the MLE.
As $\sqrt{n}\frac{Y/n-\theta}{\sqrt{\theta(1-\theta)}}$ is approx. $N(0,1)$, we get

$$\mathrm{P}(-c_{\alpha/2} \le \frac{Y-n\theta}{\sqrt{n\theta(1-\theta)}} \le c_{\alpha/2}) \approx 1-\alpha$$

The conf. limits (approx) are the roots of

$$(y-n\theta)^2 = c_{\alpha/2}^2 n\theta(1-\theta)$$

Solving this gives the confidence interval

$$\left(\frac{1}{2}\,\frac{2\,yn + c^2 n + \sqrt{4\,yn^2c^2 + c^4n^2 - 4\,y^2c^2n}}{n\,(n+c^2)}, \frac{1}{2}\,\frac{2\,yn + c^2 n - \sqrt{4\,yn^2c^2 + c^4n^2 - 4\,y^2c^2n}}{n\,(n+c^2)}\right)$$

Using the above simplification idea: For $\hat{\sigma}^2 = \frac{Y}{n}(1-\frac{Y}{n})$ one can show $\hat{\sigma}^2 \xrightarrow{P} \theta(1-\theta)$ (LLN: $Y/n \xrightarrow{P} \theta$, rules for $\xrightarrow{P}$). Using the (asymptotic) pivotal quantity

$$\sqrt{n}\frac{Y/n-\theta}{\sqrt{\frac{Y}{n}(1-\frac{Y}{n})}}$$

leads to the confidence limits

$$\frac{y}{n} \pm \frac{c_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{y}{n}(1-\frac{y}{n})}.$$

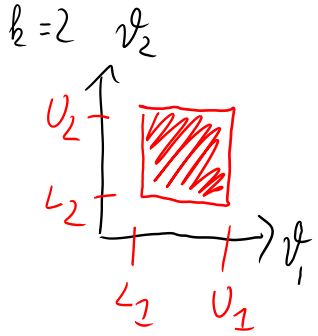
Both of the above confidence intervals are asymptotic confidence intervals with coverage probability $1-\alpha$.

6.3 Simultaneous Confidence Intervals/Confidence Regions

Extension of confidence intervals to more than one parameter / to a parameter vector:
Suppose $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T \in \Theta \subset \mathbb{R}^k$ and suppose that we have random intervals $(L_i(\mathbf{Y}), U_i(\mathbf{Y}))$ such that

$$\forall \boldsymbol{\theta}: \quad \mathrm{P}_{\boldsymbol{\theta}}(L_i(\mathbf{Y}) < \theta_i < U_i(\mathbf{Y}) \text{ for } i = 1, \dots, k) \geq 1-\alpha$$

then we call $(L_i(\mathbf{y}), U_i(\mathbf{y}))$, $i = 1, \dots, k$ a $1-\alpha$ *simultaneous confidence intervals* for $\theta_1, \dots, \theta_k$. Can we construct simultaneous confidence intervals from one-dimensional confidence intervals?



Remark (Bonferroni correction) Suppose $[L_i, U_i]$ is a $1-\alpha/k$ confidence interval for θ_i, $i = 1, \dots, k$. Then $[(L_1, \dots, L_k)^T, (U_1, \dots, U_k)^T] = (L_1, U_1) \times \cdots \times (L_k, U_k)$ is a $1-\alpha$ simultaneous confidence interval for $(\theta_1, \dots, \theta_k)^T$. Indeed,

$$\mathrm{P}(\theta_i \in [L_i, U_i], i = 1, \dots, k) = 1 - \mathrm{P}(\bigcup_{i=1}^{k}\{\theta_i \notin [L_i, U_i]\}) \geq 1 - \sum_{i=1}^{k}\underbrace{\mathrm{P}(\theta_i \notin [L_i, U_i])}_{\leq \alpha/k} \geq 1-\alpha.$$

This Bonferroni correction works without assuming anything about the joint distribution of the intervals. It also works if not all confidence intervals have the same coverage probabilities:

Example 30
Suppose $[L_1, U_1]$ is a 99% confidence interval for θ_1 and $[L_2, U_2]$ is a 97% confidence interval for θ_2. Then $[L_1, U_1] \times [L_2, U_2]$ is a 96% simultaneous confidence interval for the parameter vector (θ_1, θ_2).

The Bonferroni correction is conservative, i.e. the actual coverage probabilty is higher than the result from the correction.

Example 31
Suppose $X_1, \dots, X_n \sim N(\mu, 1)$, $Y_1, \dots, Y_n \sim N(\theta, 1)$ independent with (μ, θ) being the unknown parameter. Then we have seen previously that $I = (\bar{X} - c_{\alpha/2}/\sqrt{n}, \bar{X} + c_{\alpha/2}/\sqrt{n})$, where $\Phi(c_{\alpha/2}) = 1 - \alpha/2$, is a $1 - \alpha$ confidence interval for μ and $J = (\bar{Y} - c_{\alpha/2}/\sqrt{n}, \bar{Y} + c_{\alpha/2}/\sqrt{n})$ is a $1 - \alpha$ confidence interval for θ.

Then we know, using the Bonferroni correction, that $I \times J$ is a $1 - 2\alpha$ confidence region for (μ, θ).

In this example we know that I and J are independent, thus the actual coverage probability of $I \times J$ is

$$\mathrm{P}_{(\mu,\theta)}((\mu, \theta) \in I \times J) = \mathrm{P}_{(\mu,\theta)}(\mu \in I)\, \mathrm{P}_{(\mu,\theta)}(\theta \in J) = (1 - \alpha)^2.$$

For example, for $\alpha = 0.1$, the Bonferroni correction only guarantees a coverage probability of 80%, whereas the actual coverage probability is $0.9^2 = 0.81$.

If one uses a more complicated form than rectangles, i.e. one uses a random set $A(\mathbf{Y})$ such that for all $\boldsymbol{\theta} \in \Theta$ $\mathrm{P}_\theta(\boldsymbol{\theta} \in A(\mathbf{Y})) \geq 1 - \alpha$ one calls $A(\mathbf{y})$ a $1 - \alpha$ *confidence region* for $\boldsymbol{\theta}$. [In the second part of the course there will be an example in which the random set is an ellipse]

7 Hypothesis Tests

Beyond point estimation, statistical inference often makes use of *hypothesis testing* methods. Here, a *hypothesis* is any statement about a population parameter. We will

consider pairs of complementary hypotheses with the aim of deciding between the two.

Definition 12
The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis, denoted by H_0 and H_1, respectively.

Let θ denote the population parameter of interest taking values in a set Θ. The set Θ of all possible parameter values is usually called the *parameter space*. Formally, H_0 is a statement that $\theta \in \Theta_0$ for some $\Theta_0 \subset \Theta$. Then, H_1 corresponds to $\Theta_1 = \Theta \setminus \Theta_0$.

Example 32
(Ox weights, continued) Recall that a reasonable probability model for the guess of a participant is $X \sim N(\mu, \sigma^2)$, with $\mu = 543.4$ kg. We might therefore be interested in testing the hypothesis $H_0 : \mu = 543.4$ against the alternative $H_1 : \mu \neq 543.4$. If we reject H_0, then we are deciding that evidence suggest the average guess of a participant does not equal the true weight of the ox.

In practice, we will often use the result of a hypothesis test to assert that H_0 or H_1 is true. For example, in clinical trials the decision to be made is whether or not there is a difference in the primary outcome between the population receiving standard of care and the group receiving a novel treatment. To make this process rigorous, we must define a formal procedure for decision-making.

Definition 13

- A *hypothesis test* is a rule that specifies for which values of the sample $X_1, \ldots, X_n$ the decision is made to accept H_0 as true and for which values to reject H_0 and accept H_1 as true.
- The subset of the sample space for which H_0 will be rejected is called the *rejection region* or *critical region*.

We typically regard H_0 as the 'status quo' which we do not reject unless there is (considerable) evidence against it.

Example 33
Medical statistics:
H_0: new treatment is not better
H_1: new treatment is better.

When we use hypothesis tests, there are two types of errors we can make:

	H_0 true	H_0 false
do not reject H_0	✓	Type II error
reject H_0	Type I error	✓

A test is of level α ($0 < \alpha < 1$) if

$$\mathsf{P}_\theta(\text{reject } H_0) \leq \alpha \quad \forall \theta \in \Theta_0.$$

Usually α is small, e.g. 0.01 or 0.05.

Loosely speaking: the probability of a type I error is less than α.

There is no such bound for the probability of a type II error.

7.1 Power of a Test

Setup: Θ parameter space, $\Theta_0 \subset \Theta$, $\Theta_1 = \Theta \setminus \Theta_0$. Consider

$$H_0 : \theta \in \Theta_0 \text{ v.s. } H_1 : \theta \in \Theta_1$$

Suppose we have some test for this hypothesis.

The *power function* is defined as the mapping

$$\beta : \Theta \to [0, 1], \beta(\theta) = \mathsf{P}_\theta(\text{reject } H_0)$$

If $\theta \in \Theta_0$ then we want $\beta(\theta)$ to be small.
If $\theta \in \Theta_1$ then we want $\beta(\theta)$ to be large.

Example 34
$X \sim N(\theta, 1)$, $\theta \in \mathbb{R}$ unknown.

$$H_0 : \theta \leq 0 \quad \text{against} \quad H_1 : \theta > 0$$

Thus $\Theta = \mathbb{R}, \Theta_0 = (-\infty, 0], \Theta_1 = (0, \infty)$. Suppose we use the critical region

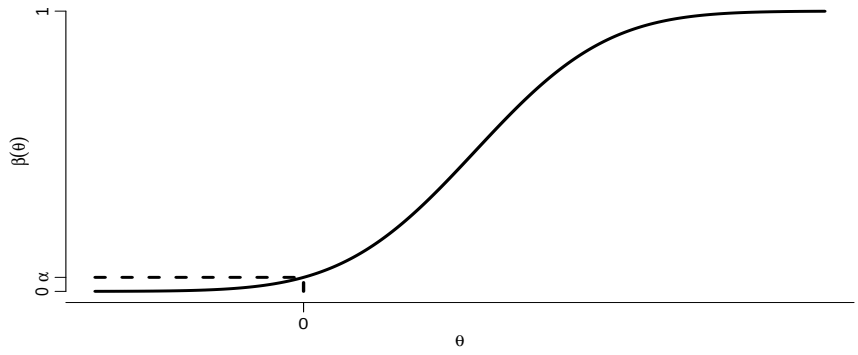
$$R = [c, \infty)$$

(i.e. we reject if $X \in R$). We will choose the critical value c s.t. the test is of level α. For $\theta \leq 0$:

$$\mathrm{P}_\theta(\text{reject } H_0) = \mathrm{P}_\theta(X \geq c) = \mathrm{P}_\theta(\underbrace{X - \theta}_{\sim N(0,1)} > c - \theta) = 1 - \Phi(c - \theta) \leq 1 - \Phi(c)$$

where Φ is the standard normal cdf. Choose c such that $\Phi(c) = 1 - \alpha$. Then $\mathrm{P}_\theta(\text{reject } H_0) \leq \alpha$ for all $\theta \in \Theta_0$.
Note: In this case it was sufficient to construct a test of level α for the boundary case θ=0.

In this situation we can work out the power function $\beta(\theta) = \mathrm{P}_\theta(\text{reject } H_0)$ [try it!]. Below is a sketch of $\beta(\theta)$:



The above is a typical $\beta(\theta)$ for a one-sided test. Examples for power functions of two-sided tests are in the next section.

7.2 p-value

Often the so-called p-value is reported (instead of a test decision):

$$p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(\text{observing something "at least as extreme" as the observation})$$

Reject H_0 iff $p \leq \alpha \rightarrow \alpha$-level test.
Advantage for computer packages: User does not have to specify the level.

If the test is based on the statistic T with rejection for large values of T then

$$p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(T \geq t),$$

where t is the observed value.

In the above example (where $X \sim N(\theta, 1)$ and $H_0 : \theta \leq 0 \quad \text{against} \quad H_1 : \theta > 0$) the p-value is:

$$p = \sup_{\theta \in \Theta_0} \mathrm{P}_\theta(X \geq x) = \mathrm{P}_0(X \geq x) = 1 - \Phi(x)$$

Example 35
Two-sided test with known variance. $X_1, \ldots, X_n \sim N(\mu, 1)$ iid, μ unknown parameter

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0$$

Under H_0: $T = \sqrt{n}(\bar{X} - \mu_0) \sim N(0, 1)$. Rejection region (based on T):

$$(-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, \infty),$$

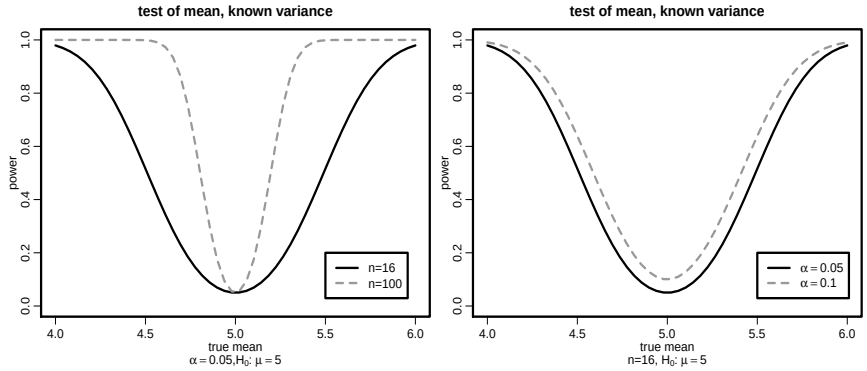
where $\Phi(c_{\alpha/2}) = 1 - \alpha/2$.

Test rejects for large values of $|T|$.
Hence, for the observation t the p-value is:

$$p = \mathrm{P}_{\mu_0}(|T| \geq |t|) = \mathrm{P}(T \leq -|t| \text{ or } T \geq |t|) = \Phi(-|t|) + 1 - \Phi(|t|) = 2 - 2\Phi(|t|)$$

Power: Note that $T \sim N(\sqrt{n}(\mu - \mu_0), 1)$.

$$\begin{aligned}
\beta(\mu) =& \mathrm{P}_\mu(|T| \geq c_{\alpha/2}) = 1 - \mathrm{P}_\mu(-c_{\alpha/2} \leq T \leq c_{\alpha/2}) \\
=& 1 - \mathrm{P}_\mu(-\sqrt{n}(\mu - \mu_0) - c_{\alpha/2} \leq T - \sqrt{n}(\mu - \mu_0) \leq -\sqrt{n}(\mu - \mu_0) + c_{\alpha/2}) \\
=& 1 - \Phi(-\sqrt{n}(\mu_0 - \mu) + c_{\alpha/2}) + \Phi(-\sqrt{n}(\mu_0 - \mu) - c_{\alpha/2})
\end{aligned}$$



Example 36 (Student's t-Test; One-Sample t-Test)
$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ iid, μ and σ unknown parameters

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0$$

Under H_0: $T = \sqrt{n}\frac{\bar{X}-\mu_0}{S} \sim t_{n-1}$.

Rejection region:

$$(-\infty, -c] \cup [c, \infty),$$

where $c = t_{n-1,\alpha/2}$. ($t_{n-1,\alpha/2}$ is chosen such that if $Y \sim t_{n-1}$ then $\mathrm{P}(Y > t_{n-1,\alpha/2}) = \alpha/2$)

(one gets similar plots for the power function)

7.3 Connection between tests and confidence intervals

Constructing a test from a confidence region: Let Y be the random observations. Suppose $A(Y)$ is a $1-\alpha$ confidence region for θ, i.e.

$$\mathrm{P}_\theta(\theta \in A(Y)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

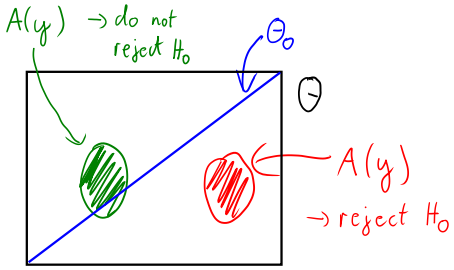
Then one can define a test for

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \notin \Theta_0$$

(where Θ_0 is some fixed subset of Θ) with level α as follows:

$$\text{Reject } H_0 \text{ if } \Theta_0 \cap A(y) = \emptyset.$$

In words: Reject H_0 if none of its elements are in the confidence region.



To see that the above test has the appropriate level: For all $\theta \in \Theta_0$,

$$\mathrm{P}_\theta(\text{type I error}) = \mathrm{P}_\theta(\text{reject}) = \mathrm{P}_\theta(\Theta_0 \cap A(Y) = \emptyset) \leq \mathrm{P}_\theta(\theta \notin A(Y)) \leq \alpha.$$

Remark If the null hypothesis is simple, i.e. $\Theta_0 = \{\theta_0\}$, then the above rule is equivalent to

$$\text{Reject } H_0 \text{ if } \theta_0 \notin A(y).$$

Constructing a confidence region from tests:

Suppose that $\forall \theta_0 \in \Theta$ we have a level α test ϕ_{θ_0} for

$$H_0^{\theta_0} : \theta = \theta_0 \quad \text{v.s.} H_1^{\theta_0} : \theta \neq \theta_0,$$

i.e. a decision rule ϕ_{θ_0} to reject or not reject $H_0^{\theta_0}$ that satisfies

$$\mathrm{P}_{\theta_0}(\phi_{\theta_0} \text{ rejects } H_0^{\theta_0}) \leq \alpha$$

Consider the random set

$$A := \{\theta_0 \in \Theta : \phi_{\theta_0} \text{ does not reject } H_0^{\theta_0}\}$$

This is a $1 - \alpha$ confidence region for θ. Indeed, $\forall \theta \in \Theta$,

$$\mathrm{P}_\theta(\theta \in A) = \mathrm{P}_\theta(\phi_\theta \text{ does not reject}) = 1 - \mathrm{P}_\theta(\phi_\theta \text{ rejects}) \geq 1 - \alpha$$

8 Likelihood Ratio Tests

Idea behind the maximum likelihood estimator: parameter with the highest likelihood is "best". Can this idea be used to create a test?

More precisely, consider the hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1 := \Theta \setminus \Theta_0$$

Main idea: compare the maximised likelihood L under H_0 ($\sup_{\theta\in\Theta_0} L(\theta)$) to the unrestricted maximum likelihood ($\sup_{\theta\in\Theta} L(\theta)$). If the latter is (much?) larger then $\sup_{\theta\in\Theta_1} L(\theta) >> \sup_{\theta\in\Theta_0} L(\theta)$, casting doubt on H_0.

Definition 14
Suppose we observe the data $\mathbf{y}$. The likelihood ratio test statistic is

$$t(\mathbf{y}) = \frac{\sup_{\theta\in\Theta} L(\theta;\mathbf{y})}{\sup_{\theta\in\Theta_0} L(\theta;\mathbf{y})} = \frac{\text{max. lik. under } H_0 + H_1}{\text{max. lik. under } H_0}$$

(Other equivalent definitions are possible) Note: $t(y) \geq 1$.

If $t(\mathbf{y})$ is "large" this will indicate support for H_1, so reject H_0 when

$$t(\mathbf{y}) \geq k,$$

where k is chosen to make

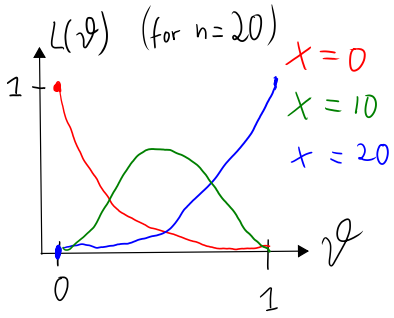
$$\sup_{\theta\in\Theta_0} \mathrm{P}_\theta(t(\mathbf{Y}) \geq k) = (\text{or } \leq)\alpha$$

(e.g. $\alpha = 0.05$).

Remark The choice of k ensures that we get a test to the level α.

Example 37
$X \sim$Binomial(n, θ), $\theta \in (0, 1) = \Theta$. n is known.

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta \neq 0.5$$



Here, $\Theta_0 = \{0.5\}$, $\Theta_1 = (0, 0.5) \cup (0.5, 1)$.
Suppose we observe the realisation x. The likelihood is

$$L : \Theta \to \mathbb{R},\ \theta \mapsto \mathrm{P}_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

We have shown previously that the MLE is $\hat{\theta} = \frac{x}{n}$. Thus

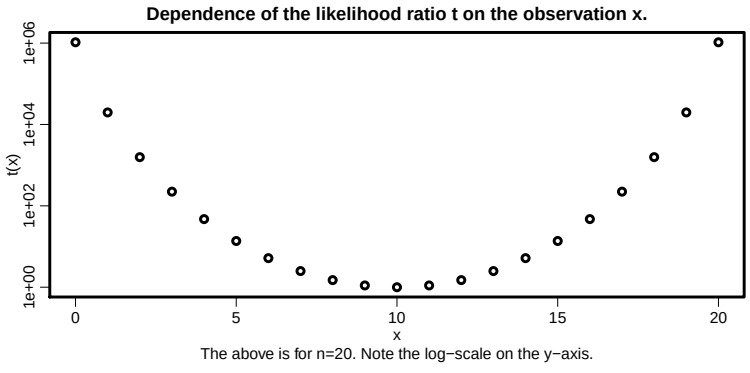
$$\sup_{\theta \in \Theta} L(\theta) = L(\hat{\theta}) = \binom{n}{x} \left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x}$$

Under H_0 :

$$\sup_{\theta \in \Theta_0} L(\theta) = L(0.5) = \binom{n}{x} 0.5^x (1-0.5)^{n-x} = \binom{n}{x} 0.5^n$$

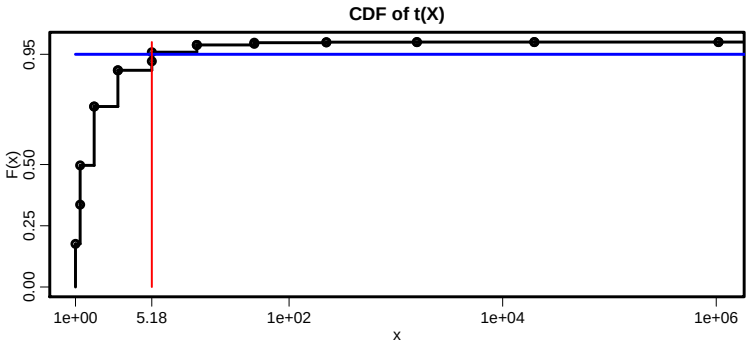
Thus, the likelihood ratio statistic is

$$t = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{\left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x}}{0.5^n}$$



To construct the test (find the threshold k or compute a p-value) we need the distribution of t under H_0.

It is possible to do this here, as the null hypothesis is simple, (i.e. consists of only one element). As X is a discrete random variable, $t(X)$ is a discrete random variable, allowing to compute the cdf of $t(X)$.



In this case rejecting if $t > 5.19$ leads to a test with level 5%.

Example 38
$X_i \sim \text{Binomial}(n, \theta_i), i = 1, 2$ indep., $\theta = (\theta_1, \theta_2) \in (0,1)^2 = \{(x, y) : x, y \in (0,1)\} = \Theta$

$$H_0 : \theta_1 = \theta_2 \text{ v.s. } H_1 : \theta_1 \neq \theta_2$$

Suppose we observe the realisation (x_1, x_2).

$$L(\theta) = \mathrm{P}_\theta(X_1 = x_1, X_2 = x_2) = \binom{n}{x_1} \theta_1^{x_1} (1-\theta_1)^{n-x_1} \binom{n}{x_2} \theta_2^{x_2} (1-\theta_2)^{n-x_2}$$

We want to derive the likelihood ratio statistic.

First we find the MLE over all of Θ. The likelihood is the product of two positive functions, the first only depending on θ_1, the second depending only on θ_2. Maximising them separately (which is just the standard maximisation of a Binomial likelihood) leads to the MLE

$$\hat{\theta}_1 = \frac{x_1}{n}, \quad \hat{\theta}_2 = \frac{x_2}{n}.$$

Under H_0, we have $\theta_1 = \theta_2$. Hence,

$$L(\theta) = \mathrm{P}_\theta(X_1 = x_1, X_2 = x_2) = \binom{n}{x_1} \binom{n}{x_2} \theta_1^{x_1 + x_2} (1-\theta_1)^{2n - (x_1 + x_2)}$$

Except from a different constant at the beginning this is the likelihood of a single Binomial$(2n, \theta_1)$ observation. Maximising this leads to the estimator

$$\hat{\theta}_1 = \hat{\theta}_2 = \frac{x_1 + x_2}{2n}$$

Thus the likelihood ratio test statistic is

$$t = \frac{L(x_1/n, x_2/n)}{L(\frac{x_1+x_2}{2n}, \frac{x_1+x_2}{2n})}$$

To construct a test, I would need the distribution of t under H_0. This is not easy to obtain as for each value of $(\theta, \theta) \in \Theta_0$ the distribution of t may be different.

Example 39
m factories producing light bulbs. Are all factories producing bulbs of the same quality? Observation: life-length of n light bulbs from each factory; Y_{ij} = life-length of bulb j from factory i.

Model: Y_{ij} indep. Exp(λ_i), i=1,. . . ,m; j=1,. . . n, $\lambda_i > 0$ unknown, $i = 1, \dots, m$.

$$H_0 : \lambda_1 = \cdots = \lambda_m \quad \text{v.s.} \quad H_1 : \text{ not } H_0$$

Interpretation of H_0 : all factories produce bulbs of equal quality

Likelihood (using $\boldsymbol{\theta}^T = (\lambda_1, \dots, \lambda_m)$):

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \prod_{j=1}^{n} \lambda_i \exp(-\lambda_i Y_{ij}) = \prod_{i=1}^{m} \lambda_i^n e^{-\lambda_i \sum_j y_{ij}}$$

As the elements of the product are all nonnegative, and only contain different components of the parameter, we can maximise them separately. Each element is the likelihood of iid Exponential(λ_i) observations, leading to the MLE

$$\hat{\lambda}_i = \frac{1}{\bar{y}_i} \text{ where } \bar{y}_i = \frac{1}{n} \sum_j y_{ij}.$$

Hence,

$$\sup_{\theta \in \Theta} L(\theta; y) = \frac{e^{-mn}}{(\prod_i \bar{y}_i)^n}$$

Under H_0 (setting $\lambda := \lambda_1 = \cdots = \lambda_m$)

$$L(\boldsymbol{\theta}; \mathbf{y}) = \lambda^{mn} e^{-\lambda \sum_{i,j} y_{ij}}$$

Again this is the likelihood for iid Exponential(λ) observations. Thus, the MLE is $\hat{\lambda} = \frac{1}{\bar{y}}$. Hence,

$$\sup_{H_0} L(\theta; y) = \frac{e^{-mn}}{\bar{y}^{mn}}$$

$\implies t(\mathbf{y}) = \frac{\bar{y}^{mn}}{(\prod_i \bar{y}_i)^n}$ To construct a test we would need to know the distr. of $t(\mathbf{Y})$ under H_0. Not easy!
Even if it were known - the distribution of $t(\mathbf{Y})$ may depend on λ and hence, choosing k according to $\sup_{\lambda>0} \mathrm{P}_\lambda(t(\mathbf{Y}) \geq k) = \alpha$ may not be easy.

Definition 15
Let $X_1, X_2, \ldots, X_n \sim N(0,1)$ independently. Then $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.

Theorem 6
Under certain regularity conditions (in particular H_0 must be "nested" in H_1, i.e. Θ_0 is a lower-dimensional subspace/subset of Θ),

$$2 \log t(\mathbf{Y}) \xrightarrow{d} \chi_r^2 \quad (n \to \infty)$$

under H_0, where $r = \#$independent restrictions on $\boldsymbol{\theta}$ needed to define H_0.

Alternative way to derive the degrees of freedom r:

$r = \#$ of independent parameters under full model$-\#$ of independent parameters under H_0

Example 40
In the above examples:

- $X \sim$Binomial(n, θ), $\theta \in (0,1) = \Theta$ with $H_0 : \theta = 0.5$ v.s. $H_1 : \theta \neq 0.5$: r=1
- $X_i \sim$ Binomial(n, θ_i), $i = 1, 2$ indep., $\theta \in (0,1)^2$ with $H_0 : \theta_1 = \theta_2$ v.s. $H_1 : \theta_1 \neq \theta_2$: r=1
- "light bulbs": $r = m - 1$

Proof (Sketch of a proof for the case $\Theta_0 = \{\theta_0\}$) Suppose $\Theta \subset \mathbb{R}^r$. Then

$$2 \log t(\mathbf{Y}) = 2(\log L(\hat{\theta}) - \log L(\theta_0)),$$

where $\hat{\theta}$ denotes the MLE of θ. Using a Taylor expansion,

$$\log L(\theta_0) \approx \log L(\hat{\theta}) + (\theta_0 - \hat{\theta})^T \underbrace{\frac{\partial \log L(\theta)}{\partial \theta}\Big|_{\hat{\theta}}}_{=0;\text{(suff. cond. for min)}} + \frac{1}{2}(\theta_0 - \hat{\theta})^T \frac{\partial\partial \log L(\theta)}{\partial\theta\partial\theta^T}\Big|_{\hat{\theta}}(\theta_0 - \hat{\theta})$$

Hence,

$$2 \log t(\mathbf{Y}) \approx (\theta_0 - \hat{\theta})^T \left(-\frac{\partial\partial \log L(\theta)}{\partial\theta\partial\theta^T}\Big|_{\hat{\theta}}\right)(\theta_0 - \hat{\theta}).$$

By a multivariate version of Theorem 5 (asymptotics of MLE),

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_f(\theta_0)^{-1}),$$

where $I_f(\theta) = \mathrm{E}_\theta[(\frac{\partial}{\partial\theta} \log L(\theta))(\frac{\partial}{\partial\theta} \log L(\theta))^T]$.
By the law of large numbers (and a few more arguments similar to the proof of the asymptotics of the MLE),

$$-\frac{1}{n}\frac{\partial\partial \log L(\theta)}{\partial\theta\partial\theta^T}\Big|_{\hat{\theta}} \xrightarrow{P_{\theta_0}} I_f(\theta_0).$$

Hence, $2 \log t(\mathbf{Y}) \approx \mathbf{Z}^T I_f(\theta_0)\mathbf{Z}$, where $\mathbf{Z} \sim N_p(0, I_f(\theta_0)^{-1})$.
Results on quadratic forms of normal random vectors (which will be derived in the second part of this course) imply $\mathbf{Z}^T I_f(\theta_0)\mathbf{Z} \sim \chi^2_r$.

9 Linear Models with Second Order Assumptions

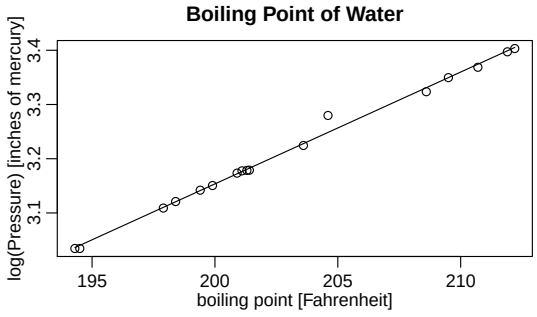
Many scientific problems focus on evaluating associations between an outcome of interest and a set of predictor variables. In the remainder of the course, we consider how linear regression can be used for this purpose.

9.1 Simple Linear Regression

Example 41
Boiling Point of Water

Can the boiling point of water be used to measure altitude? Altitude can be measured by atmospheric pressure. Early barometers were difficult to transport, so, in the 19th century, James David Forbes, a Scottish scientist, conducted a study if the boiling point of water can predict atmoshperic pressure (and thus ultimately altitude).
He collected 17 observations on boiling point (Fahrenheit) and barometric pressure (inches of mercury) at various locations in the alps.
(S. Weisberg (1980), Applied Linear Regression, Wiley.)

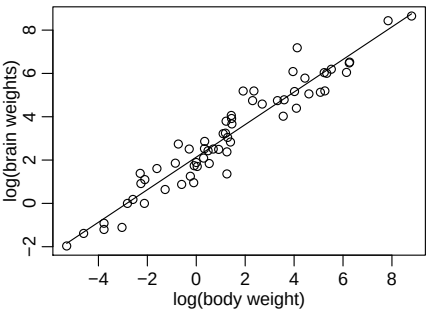
bp	pressure	bp	pressure
194.5	20.79	194.3	20.79
197.9	22.4	198.4	22.67
199.4	23.15	199.9	23.35
200.9	23.89	201.1	23.99
201.4	24.02	201.3	24.01
203.6	25.14	204.6	26.57
209.5	28.49	208.6	27.76
210.7	29.04	211.9	29.88
212.2	30.06		



Example 42
Brain and Body Weights for 62 Species of Land Mammals

	body	brain
Chimpanzee	52.160	440.0
Human	62.000	1320.0
Donkey	187.100	419.0
Horse	521.000	655.0
Cow	465.000	423.0
Grey wolf	36.330	119
Goat	27.660	115
African giant pouched rat	1.000	6
Asian elephant	2547.000	4603
African elephant	6654.000	5712
Baboon	10.550	179
Rat	0.280	1
Red fox	4.235	50
Brazilian tapir	160.000	169
Jaguar	100.000	157
Pig	192.000	180.0
Desert hedgehog	0.550	2.4
Slow loris	1.400	12.5
Golden hamster	0.120	1.0

...



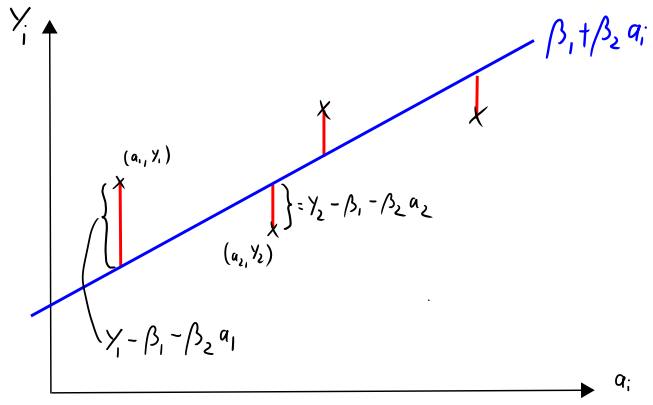
The *simple linear model* is

$$Y_i = \beta_1 + a_i\beta_2 + \epsilon_i, \quad i = 1, \ldots, n$$

- Y_i "outcome", "response"; observable random variable.
- a_i "covariate"; observable constant.
- β_1, β_2 unknown parameters.
- Error $\epsilon_1, \ldots, \epsilon_n$ iid, $\mathrm{E}\,\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. $\sigma^2 > 0$ is another unknown parameter. The errors $\epsilon_1, \ldots, \epsilon_n$ are not observable.

Forbes data: a_i =boiling point and Y_i = log(pressure)

Mammals data: a_i = log(body weight) and Y_i =log(brain weight).



The so-called *least squares estimators* $\hat{\beta}_1$, $\hat{\beta}_2$ of β_1 and β_2 are defined as the minimisers of

$$S(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - a_i\beta_2)^2.$$

Remark Note that:

- $e_i = y_i - \hat{\beta}_1 - a_i\hat{\beta}_2$, the so-called residuals, are observable. They are not iid, as dependence is introduced via $\hat{\beta}_1$, $\hat{\beta}_2$.
- The unknown parameters are β_1, β_2 and σ^2.
- In linear regression models $Y_1, \ldots, Y_n$ are generally not iid observations. Independence will still hold if the errors $\epsilon_1, \ldots, \epsilon_n$ are independent. However, the Y_i do not have the same distribution; the distribution of Y_i depends on the covariate a_i.

The general formulation of the linear model makes heavy use of matrix notation and properties of random vectors. In the next two subsections, some useful tools for this are developed.

9.2 Matrix Algebra

This section contains some useful results about matrices. A^T denotes the transpose of a matrix. I will use the terms "invertible" and "non-singular" synonymously.

Lemma 5
Let $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times n}$. Then $(AB)^T = B^T A^T$

Let $A \in \mathbb{R}^{n\times n}$ be non-singular. Then $(A^{-1})^T = (A^T)^{-1}$.

You can try the proof of the above lemma yourself.

Let $A = (A_{ij}) \in \mathbb{R}^{n\times n}$. Its trace, denoted by $\text{trace}(A)$ is the sum of its diagonal elements, i.e. $\text{trace}(A) = \sum_{i=1}^{n} A_{ii}$.

Lemma 6
$\text{trace}(AB) = \text{trace}(BA)$.

Proof Recall that $AB = (\sum_j A_{ij}B_{jk})_{i,k}$. Thus,
$\text{trace}(AB) = \sum_i \sum_j A_{ij}B_{ji} = \sum_j \sum_i B_{ji}A_{ij} = \text{trace}(BA)$

Example 43
Let $A = (1, 1)$, $B = (1, 1)^T$
Then $AB = 2 \neq BA = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, but $\text{trace}(AB) = 2 = \text{trace}(BA)$.

Lemma 7
Let X be an $n \times p$ matrix. Then $\text{rank}(X^T X) = \text{rank}(X)$.

Proof Let $\text{kern}(X) = \{\mathbf{x} : X\mathbf{x} = \mathbf{0}\}$. Then $p = \text{rank}\, X + \dim \text{kern}(X)$ (this is the "Rank-nullity theorem", which you know from previous courses). Similarly, $p = \text{rank}\, X^T X + \dim \text{kern}(X^T X)$
It suffices to show: $\text{kern}(X) = \text{kern}(X^T X)$.
If $\mathbf{x} \in \text{kern}(X)$ then $\mathbf{0} = X\mathbf{x}$ and hence $\mathbf{0} = X^T X\mathbf{x}$ which shows $\mathbf{x} \in \text{kern}(X^T X) = \{\mathbf{y} : X^T X\mathbf{y} = \mathbf{0}\}$.
If $\mathbf{x} \in \text{kern}(X^T X)$ then $\mathbf{0} = X^T X\mathbf{x}$ and thus $\mathbf{0} = \mathbf{x}^T X^T X\mathbf{x} = (X\mathbf{x})^T X\mathbf{x} = \|X\mathbf{x}\|^2$ which shows $X\mathbf{x} = \mathbf{0}$, i.e. $\mathbf{x} \in \text{kern}(X)$.

Recall that a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called positive definitive if $\forall \mathbf{x} \in \mathbb{R}^n \subset \{0\} : \mathbf{x}^T A \mathbf{x} > 0$.

Lemma 8
If $A \in \mathbb{R}^{n \times n}$ is symmetric then $\exists$ an orthogonal matrix P (i.e. $P^T P = I$) s.t. $P^T A P$ is diagonal (with diagonal entries equal to the eigenvalues of A).
If A is an $n \times n$ positive definite symmetric matrix, then $\exists$ a non-singular matrix Q s.t. $Q^T A Q = I_n$.

Proof The first result is a standard piece of linear algebra book-work.
The second result can be derived from it: Suppose A is positive definite, then its eigenvalues are all positive. Hence, $P^T A P = D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_i > 0 \, \forall i$. Let $E = D^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ and define $Q = PE^{-1}$. Then

$$Q^T A Q = (PE^{-1})^T A P E^{-1} = (E^{-1})^T P^T A P E^{-1} = (E^{-1})^T E E E^{-1} = I.$$

9.3 Review of rules for E, cov for random vectors

Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector. Then

$$\mathrm{E}(\mathbf{X}) = (\mathrm{E}\, X_1, \ldots, \mathrm{E}\, X_n)^T,$$

i.e. the expectation is defined componentwise. For random matrices the expectation is also defined componentwise.

Lemma 9
Let $\mathbf{X}$ and $\mathbf{Y}$ be n-variate random vectors. Then the following hold:

- $\mathrm{E}(\mathbf{X} + \mathbf{Y}) = \mathrm{E}\, \mathbf{X} + \mathrm{E}\, \mathbf{Y}$.
- Let $a \in \mathbb{R}$ then $\mathrm{E}(a\mathbf{X}) = a\, \mathrm{E}(\mathbf{X})$
- Let A, B be deterministic matrices of "suitable dimensions" (deterministic means that they are not random). Then $\mathrm{E}(A\mathbf{X}) = A\, \mathrm{E}(\mathbf{X})$ and $\mathrm{E}(\mathbf{X}^T B) = \mathrm{E}(\mathbf{X})^T B$.

Proof Use properties of one-dimensional random variables, for example

$$\mathrm{E}(A\mathbf{X}) = (\mathrm{E}(\sum_j A_{ij}X_j))_i = (\sum_j A_{ij}\,\mathrm{E}(X_j))_i = A\,\mathrm{E}\,\mathbf{X}$$

Definition 16
If $\mathbf{X}$, $\mathbf{Y}$ are random vectors then

$$\begin{aligned}\mathrm{cov}(\mathbf{X},\mathbf{Y}) :=& (\mathrm{cov}(X_i, Y_j))_{i,j} \\ &= \mathrm{E}[(\mathbf{X}-\mathrm{E}(\mathbf{X}))(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))^T] = \mathrm{E}[\mathbf{X}\mathbf{Y}^T] - \mathrm{E}(\mathbf{X})\,\mathrm{E}(\mathbf{Y})^T.\end{aligned}$$

Furthermore $\mathrm{cov}(\mathbf{X}) := \mathrm{cov}(\mathbf{X},\mathbf{X})$.

The above definition really contains three equivalent definitions of $\mathrm{cov}(\mathbf{X},\mathbf{Y})$. It is straightforward to show that they are equivalent.

Lemma 10
If $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are random vectors, A, B are deterministic matrices and $a, b \in \mathbb{R}$ are constants then (assuming appropriate dimensions)

$\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \mathrm{cov}(\mathbf{Y},\mathbf{X})^T$

$\mathrm{cov}(a\mathbf{X}+b\mathbf{Y},\mathbf{Z}) = a\,\mathrm{cov}(\mathbf{X},\mathbf{Z}) + b\,\mathrm{cov}(\mathbf{Y},\mathbf{Z})$

$\mathrm{cov}(A\mathbf{X}, B\mathbf{Y}) = A\,\mathrm{cov}(\mathbf{X},\mathbf{Y})B^T$

$\mathrm{cov}(A\mathbf{X}) = A\,\mathrm{cov}(\mathbf{X})A^T$

$\mathrm{cov}(\mathbf{X})$ is positive semidefinite and symmetric,
i.e. $\mathbf{c}^T\,\mathrm{cov}(\mathbf{X})\mathbf{c} \geq 0$ for all vectors $\mathbf{c}$, or, equivalently, all eigenvalues of $\mathrm{cov}(\mathbf{X})$ are nonnegative.

- If $\mathbf{X}$ and $\mathbf{Y}$ are independent then $\mathrm{cov}(\mathbf{X},\mathbf{Y}) = 0$.

Proof Play back to properties of the one-dimensional covariance or work with one of the vector definitions of the covariance. For example, to see the last property,

$$\mathbf{c}^T\,\mathrm{cov}(\mathbf{X})\mathbf{c} = \mathrm{E}(\mathbf{c}^T(\mathbf{X}-\mathrm{E}\,\mathbf{X})(\mathbf{X}-\mathrm{E}\,\mathbf{X})^T\mathbf{c}) = \mathrm{E}[(\mathbf{c}^T(\mathbf{X}-\mathrm{E}\,\mathbf{X}))^2] \geq 0$$

Example 44
Let $X \sim Binomial(17, 0.4)$. Then

$\text{cov}(X) =$

Example 45
If $Y_1, \ldots, Y_n$ are independent then
$\text{cov}(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}) =$

Example 46
Let X, Y be independent r.v. with $X \sim N(5, 2)$ and $Y \sim \text{Binomial}(10, 0.5)$. Then

$$\text{cov}\left(\begin{pmatrix} X \\ -X \end{pmatrix}\right) =$$

$$\text{cov}\left(\begin{pmatrix} X \\ X+Y \end{pmatrix}\right) =$$

$$\text{cov}\left(X, \begin{pmatrix} 2X \\ X-Y \end{pmatrix}\right) =$$

9.4 Linear Model

General formulation that allows for any number of covariates influencing an observation.

Definition 17
In a *linear model*

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$\mathbf{Y}$ is an n-dimensional random vector (observable),

$X \in \mathbb{R}^{n \times p}$ is a known matrix (often called "design matrix"),

$\boldsymbol{\beta} \in \mathbb{R}^p$ is an *unknown parameter* and

$\boldsymbol{\epsilon}$ is an n-variate random vector (not observable) with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$.

Remark Unless mentioned otherwise, we will assume $n > p$.

Remark Concerning notation: From now on, a vector will implicitly always be a column vector. Vectors will be printed in bold. The transpose of a vector $\mathbf{x}$ is denoted by $\mathbf{x}^T$. Expectations of random vectors are defined componentwise, i.e. $\mathrm{E}(\boldsymbol{\epsilon}) = (\mathrm{E}(\epsilon_1), \dots, \mathrm{E}(\epsilon_n))^T$.

Remark Shorter way of writing a linear model:

$$\mathrm{E}\,\mathbf{Y} = X\boldsymbol{\beta}$$

Example 47 (Forbes, Mammals)
In the examples of the previous section, we had $Y_i = \beta_1 + a_i\beta_2 + \epsilon_i$. This fits into the general framework with $X = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{pmatrix}$, where

Forbes data: a_i =boiling point and $Y_i = \log(\text{pressure})$

Mammals data: $a_i = \log(\text{body weight})$ and Y_i =log(brain weight).

Example 48
20 patients, 2 drugs, A and B
10 given A, 10 given B
Y_{Aj} = response of jth patient to receive A, $j = 1, \ldots, 10$
Y_{Bj} = response of jth patient to receive B, $j = 1, \ldots, 10$
The simplest model is $\mathrm{E}(Y_{Aj}) = \mu_A$, $\mathrm{E}(Y_{Bj}) = \mu_B$. In matrix form:

$$\mathrm{E}\begin{pmatrix} Y_{A1} \\ \vdots \\ Y_{A,10} \\ Y_{B1} \\ \vdots \\ Y_{B,10} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

Example 49
10 pairs of twins, 2 drugs: A and B
one twin in each pair receives A, the other one receives B
twins are alike - we need to modify our previous model:

$\mathrm{E}(Y_{Aj}) = \mu_A + \tau_j$, $\mathrm{E}(Y_{Bj}) = \mu_B + \tau_j$, where τ_j=effect of twin pair j.

$$\mathrm{E}\begin{pmatrix} Y_{A1} \\ \vdots \\ Y_{A,10} \\ Y_{B1} \\ \vdots \\ Y_{B,10} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 1 \end{pmatrix}}_{=X} \underbrace{\begin{pmatrix} \mu_A \\ \mu_B \\ \tau_1 \\ \vdots \\ \tau_{10} \end{pmatrix}}_{\beta}$$

Example 50
We might be interested in a model of the following form:

$$\mathrm{E}(Y_i) = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i3}^2,$$

where x_{ik} is the value of the kth predictor for observation i. This is a linear model, as the parameters act linearly.

Example 51

$$\mathrm{E}(Y_i) = \beta_1 + \beta_2 x_i^{\beta_3}$$

is not a linear model. The influence of the parameter β_3 is not linear.

Assumptions:

We will frequently use some of the following further assumptions about the linear model.

Second Order Assumption (SOA): $\mathrm{cov}(\boldsymbol{\epsilon}) = (\mathrm{cov}(\epsilon_i, \epsilon_j))_{\substack{i=1,\dots,n \\ j=1,\dots,n}} = \sigma^2 I_n$ for some $\sigma^2 > 0$.

(SOA) really consists of two parts: First, that the errors of two different observations, ϵ_i and ϵ_j for $i \neq j$ are uncorrelated. Second, that the variance of all errors is identical (recall: $\mathrm{Var}(\epsilon_i) = \mathrm{cov}(\epsilon_i, \epsilon_i)$).

For the remainder of this Chapter we will assume (SOA)

Normal theory assumptions (NTA): $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ for some $\sigma^2 > 0$.

N denotes the n-dimensional multivariate normal distribution. One could equivalently define the NTA as: $\epsilon_1, \dots, \epsilon_n \sim N(0, \sigma^2)$ independently.

(NTA) implies (SOA). We will use (NTA) in Chapter 3 to construct tests and confidence intervals.

Full rank (FR) The matrix X has full rank.

We say that a matrix has "full rank" if it has the highest possible rank for its dimensions, i.e. if $\text{rank}(X) = \min(n, p)$. As we are mostly working with the situation $n > p$, (FR) reduces to $\text{rank}(X) = p$.

In the following, we will denote the rank of X always by $r = \text{rank}(X)$.

9.5 Identifiability

In statistical models, one of the main aims is to determine the unknown parameter. If two parameter values lead to the same distribution for the observed data we cannot distinguish between these parameter values.

> **Definition 18**
> Suppose we have a statistical model with unknown parameter θ. We call θ *identifiable* if no two different value of $\boldsymbol{\theta}$ lead to the same distribution of the observed data.

For a linear model: the main parameter we are interested is $\boldsymbol{\beta}$ and the observation is $\mathbf{Y}$. It turns out that if $r < p$, then the parameter vector $\boldsymbol{\beta}$ is not identifiable. The following example shows this.

> **Example 52**
> Consider again the linear model for the study with twins in which $r = 11$ and $p = 12$.
>
> Let $\boldsymbol{\beta} = (\mu_A, \mu_B, \tau_1, \dots, \tau_{10})^T \in \mathbb{R}^p$ and for some $\delta > 0$ let
> $\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \mu_A - \delta \\ \mu_B - \delta \\ \tau_1 + \delta \\ \vdots \\ \tau_{10} + \delta \end{pmatrix}$. Then $X\boldsymbol{\beta} = \begin{pmatrix} \mu_a + \tau_1 \\ \vdots \\ \mu_a + \tau_{10} \\ \mu_b + \tau_1 \\ \vdots \\ \mu_b + \tau_{10} \end{pmatrix} = \begin{pmatrix} \mu_a - \delta + \tau_1 + \delta \\ \vdots \\ \mu_a - \delta + \tau_{10} + \delta \\ \mu_b - \delta + \tau_1 + \delta \\ \vdots \\ \mu_b - \delta + \tau_{10} + \delta \end{pmatrix} = X\tilde{\boldsymbol{\beta}}$. Thus $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ lead to the same expected value of $\mathbf{Y}$. Thus $\boldsymbol{\beta}$ is not identifiable.

9.6 Least Squares Estimation

We will estimate $\boldsymbol{\beta}$ by least squares.

Least squares: choose $\boldsymbol{\beta}$ to minimise

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$$

(Later we will see that under (NTA) the least squares estimator is the MLE).

Multiplying out we get

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}^T X^T\mathbf{Y} - \mathbf{Y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \\ &= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \end{aligned}$$

The derivative wrt $\boldsymbol{\beta}$ is

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_i} \right)_{i=1,\ldots,p} = -2X^T\mathbf{Y} + 2X^T X\boldsymbol{\beta}$$

[You can check this by componentwise differentiation.] The least squares estimator $\hat{\boldsymbol{\beta}}$ should satisfy $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0}$ and hence we get the

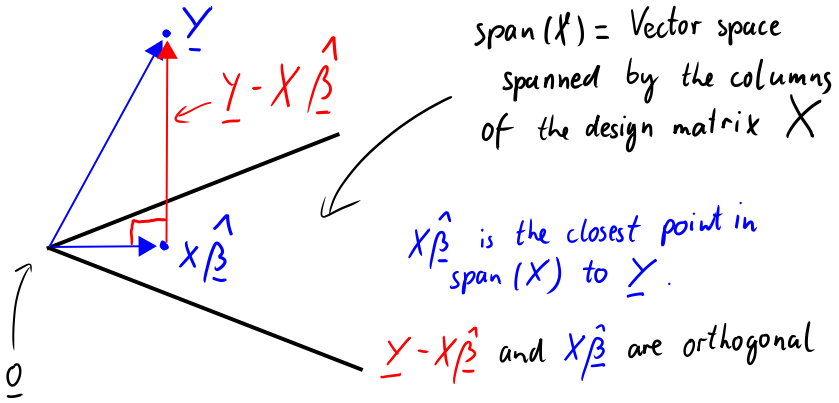
$$\text{Least Squares Equations (LSE): } X^T X\hat{\boldsymbol{\beta}} = X^T\mathbf{Y}$$

These equations have a unique solution $\iff$ $X^T X$ is invertible (i.e. has rank p).
But $\operatorname{rank} X^T X = \operatorname{rank} X$ (see Lemma 7).
Hence the LSE have a unique solution $\iff$ the linear model is of full rank.

Suppose $\hat{\boldsymbol{\beta}}$ satisfies the LSE. Then it does in fact minimise $S(\boldsymbol{\beta})$. To see this, for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{Y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) \\ &= S(\hat{\boldsymbol{\beta}}) + \underbrace{(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})}_{\geq 0} + 2\underbrace{(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}})}_{=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(X^T\mathbf{Y} - X^T X\hat{\boldsymbol{\beta}}) = 0} \\ &\geq S(\hat{\boldsymbol{\beta}}) \end{aligned}$$

Geometrical Interpretation



To see that $\mathbf{Y} - X\hat{\beta}$ is orthogonal to $X\hat{\beta}$:

$$(X\hat{\beta})^T(\mathbf{Y} - X\hat{\beta}) = \hat{\beta}^T X^T(\mathbf{Y} - X\hat{\beta}) = \hat{\beta}^T \underbrace{(X^T\mathbf{Y} - X^T X\hat{\beta})}_{=\mathbf{0} \text{ by LSE}} = 0$$

9.7 Properties of Least Squares Estimation

In this section: assume (FR) and (SOA)
Then

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

- $\hat{\beta}$ is linear in $\mathbf{Y}$.
  More precisely: the function $\mathbb{R}^n \to \mathbb{R}^p$, $\mathbf{y} \mapsto (X^T X)^{-1} X^T \mathbf{y}$ is a linear mapping.
- $\hat{\beta}$ is unbiased for β.

  Indeed, for all β:

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

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

Indeed, letting $A = (X^T X)^{-1} X^T$ we have

$$\begin{aligned} \operatorname{cov} \hat{\boldsymbol{\beta}} &= \operatorname{cov}(A\mathbf{Y}) = A \operatorname{cov}(\mathbf{Y}) A^T = \sigma^2 A A^T \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

We now compute the least squares estimator explicitly in one simple situation.

Example 53 (Simple linear regression)
Let

$$Y_i = \beta_1 + \beta_2 a_i + \epsilon_i, \quad \mathrm{E}\, \epsilon_i = 0 \qquad i = 1, \dots, n,$$

where $a_1, \dots, a_n$ are known deterministic constants. Assume that $n \geq 2$

$\mathbf{Y}^T = (Y_1, \dots, Y_n)$, $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$ and $X = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{pmatrix}$.
Assume SOA and that not all a_is are equal (to ensure FR). Then

$$\begin{aligned} X^T X &= \begin{pmatrix} n & n\bar{a} \\ n\bar{a} & \sum a_i^2 \end{pmatrix} \\ (X^T X)^{-1} &= \frac{1}{n \sum a_i^2 - n^2 \bar{a}^2} \begin{pmatrix} \sum a_i^2 & -n\bar{a} \\ -n\bar{a} & n \end{pmatrix} \\ X^T \mathbf{Y} &= \begin{pmatrix} n\overline{Y} \\ \sum a_i Y_i \end{pmatrix}. \end{aligned}$$

Now we can find $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$, hence

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \frac{1}{\sum a_i^2 - n\bar{a}^2} \begin{pmatrix} \overline{Y} \sum a_i^2 - \bar{a} \sum a_i Y_i \\ \sum a_i Y_i - n\bar{a}\overline{Y} \end{pmatrix}.$$

$$\hat{\beta}_2 = \frac{\sum (a_i - \bar{a})(Y_i - \overline{Y})}{\sum (a_i - \bar{a})^2}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \bar{a}.$$

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n \sum (a_i - \bar{a})^2} \begin{pmatrix} \sum a_i^2 & -n\bar{a} \\ -n\bar{a} & n \end{pmatrix}.$$

If $\bar{a} = 0$ everything becomes easier: the covariance matrix is diagonal and $\widehat{\beta}_1 = \bar{Y}$.

To get to this situation, we can re-parametrise the model by letting $\gamma_1 = \beta_1 + \bar{a}\beta_2$ and $\gamma_2 = \beta_2$. Then

$$\mathsf{E}(\mathbf{Y}) = X\boldsymbol{\beta} = \begin{pmatrix} \beta_1 + a_1\beta_2 \\ \vdots \\ \beta_1 + a_n\beta_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 + (a_1 - \bar{a})\gamma_2 \\ \vdots \\ \gamma_1 + (a_n - \bar{a})\gamma_2 \end{pmatrix} = \begin{pmatrix} 1 & (a_1 - \bar{a}) \\ \vdots & \vdots \\ 1 & (a_n - \bar{a}) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

Then $\sum_{i=1}^{n}(a_i - \bar{a}) = 0$, $\hat{\gamma} = \bar{Y}$, $\hat{\gamma}_2 = \hat{\beta}_2$.

If the main interest lies in β_2 then one can work with the simpler transformed model and get the same estimates via γ_2.

The following theorem justifies the use of the least squares estimator - it can be used to construct a best linear unbiased estimator (*BLUE*).

An estimator $\hat{\gamma}$ is called linear if there exists $\mathbf{L} \in \mathbb{R}^n$ such that $\hat{\gamma} = \mathbf{L}^T\mathbf{Y}$.

Theorem 7 (The Gauss-Markov Theorem for full-rank linear models)
Assume (FR), (SOA). Let $\mathbf{c} \in \mathbb{R}^p$ and let $\hat{\boldsymbol{\beta}}$ be a least squares estimator of $\boldsymbol{\beta}$ in a linear model. Then the following holds: The estimator $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators for $\mathbf{c}^T\boldsymbol{\beta}$.

Remark For $i \in \{1, \dots, n\}$, let $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ with the 1 being in the ith component. Choosing $\mathbf{c} = \mathbf{e}_i$ we have $\mathbf{c}^T\boldsymbol{\beta} = \beta_i$ and $\mathbf{c}^T\hat{\boldsymbol{\beta}} = \hat{\beta}_i$. Thus $\hat{\beta}_i$ has the smallest variance among all linear unbiased estimators of β_i.

Proof $\mathbf{c}^T\hat{\boldsymbol{\beta}}$ is linear and unbiased
Let $\hat{\gamma} = \mathbf{L}^T\mathbf{Y}$ be any other linear unbiased estimator of $\mathbf{c}^T\boldsymbol{\beta}$.

$$\begin{aligned} \mathsf{Var}(\hat{\gamma}) &= \mathsf{Var}(\mathbf{L}^T\mathbf{Y}) = \mathsf{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}} + \underbrace{(\mathbf{L}^T - \mathbf{c}^T(X^TX)^{-1}X^T)}_{=:\mathbf{D}^T}\mathbf{Y}) \\ &= \mathsf{cov}(\mathbf{c}^T\hat{\boldsymbol{\beta}} + \mathbf{D}^T\mathbf{Y}, \mathbf{c}^T\hat{\boldsymbol{\beta}} + \mathbf{D}^T\mathbf{Y}) \\ &= \mathsf{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}) + \mathsf{Var}(\mathbf{D}^T\mathbf{Y}) + 2\,\mathsf{cov}(\mathbf{c}^T\hat{\boldsymbol{\beta}}, \mathbf{D}^T\mathbf{Y}) \end{aligned}$$

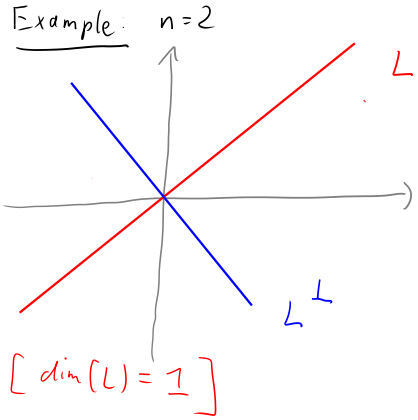
$$\operatorname{cov}(\mathbf{c}^T\hat{\beta}, \mathbf{D}^T\mathbf{Y}) = \mathbf{c}^T(X^TX)^{-1}X^T \underbrace{\operatorname{cov}(\mathbf{Y})}_{=\sigma^2 I_n} \mathbf{D} = \mathbf{c}^T(X^TX)^{-1}(\underbrace{\mathbf{D}^TX}_{\overset{(*)}{=}\mathbf{0}^T})^T\sigma^2 = 0.$$

To see $(*)$: both est unbiased $\implies 0 = \mathrm{E}(\hat{\gamma}) - \mathrm{E}(\mathbf{c}^T\hat{\beta}) = \mathrm{E}(\mathbf{D}^T\mathbf{Y}) = \mathbf{D}^TX\beta$ for all β. Hence, $\mathbf{D}^TX = \mathbf{0}^T$. Thus,

$$\operatorname{Var}(\hat{\gamma}) = \operatorname{Var}(\mathbf{c}^T\hat{\beta}) + \operatorname{Var}(\mathbf{D}^T\mathbf{Y}) \geq \operatorname{Var}(\mathbf{c}^T\hat{\beta}).$$

9.8 Projection Matrices

Let L be a linear subspace of $\mathbb{R}^n$, $\dim L = r \leq n$.



> **Definition 19**
> $P \in \mathbb{R}^{n\times n}$ is a projection matrix onto L, if
>
> 1. $P\mathbf{x} = \mathbf{x} \quad \forall \mathbf{x} \in L$
> 2. $P\mathbf{x} = \mathbf{0} \quad \forall \mathbf{x} \in L^\perp = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T\mathbf{y} = \mathbf{0}\, \forall \mathbf{y} \in L\}$

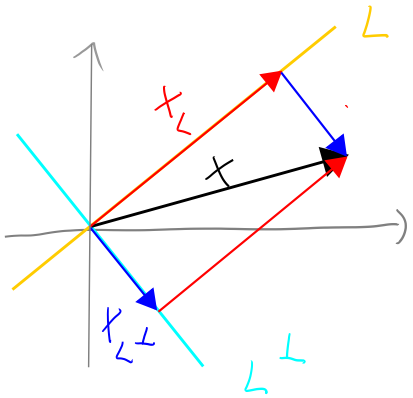
By definition, $\operatorname{rank} P = \dim L = r$.

Remark Projection matrices will be very useful when proving results for linear models. We will often use $L = \operatorname{span}(X)$, the space spanned by the columns of the design matrix X, or $L = \operatorname{span}(X)^\perp$.

> **Lemma 11**
> P is a projection matrix $\iff$ $\underbrace{P^T = P}_{P \text{ symmetric}}$ and $\underbrace{P^2 = P}_{P \text{ idempotent}}$.

Proof $\Rightarrow$:

Recall that any $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as $\mathbf{x} = \mathbf{x}_L + \mathbf{x}_{L^\perp}$, where $\mathbf{x}_L \in L$ and $\mathbf{x}_{L^\perp} \in L^\perp$.



Let $\mathbf{x} \in \mathbb{R}^n$. Then $P^2\mathbf{x} = P(P\mathbf{x}) = P\mathbf{x}_L = P\mathbf{x}$. Hence, $P^2 = P$.

For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{x}^T P^T \mathbf{y} = (\underbrace{P\mathbf{x}}_{\in L})^T \mathbf{y} = (\underbrace{P\mathbf{x}}_{\mathbf{x}_L})^T \mathbf{y}_L = \mathbf{x}_L^T P\mathbf{y} = \mathbf{x}^T P\mathbf{y}$$

Hence, $P^T = P$

$\Leftarrow$:
Let L be the space spanned by the columns of P.

- Let $\mathbf{x} \in L$. Then $\exists \mathbf{z} \in \mathbb{R}^n : \mathbf{x} = P\mathbf{z}$. Hence, $P\mathbf{x} = P^2\mathbf{z} \overset{\text{idempot}}{=} P\mathbf{z} = \mathbf{x}$.
- Let $\mathbf{x} \in L^\perp$. Then for all $y \in \mathbb{R}^n$: $(P\mathbf{x})^T\mathbf{y} = \mathbf{x}^T P^T \mathbf{y} \overset{\text{symm}}{=} \mathbf{x}^T \underbrace{P\mathbf{y}}_{\in L} = \mathbf{0}$. Hence $P\mathbf{x} = \mathbf{0}$.

The projection matrix is unique. Indeed, for each i, the vector $\mathbf{e}_i$ can be uniquely written as $\mathbf{e}_i = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in L$ and $\mathbf{y} \in L^\perp$. Then the ith column of P is $P\mathbf{e}_i = \mathbf{x}$.

If $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are a basis of L then the projection onto L is given by

$$P = X(X^T X)^{-1} X^T,$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_r)$. [prove this directly via the definition of the projection matrix or check $P^2 = P$, $P^T = P$, $\underbrace{\text{span}(P)}_{\text{space spanned by the columns of } P} = L$ or .]

If $\mathbf{x}_1, \ldots, \mathbf{x}_r$ are an *orthonormal* basis then $P = XX^T$.

$I_n - P$ is the projection matrix onto $L^\perp$

(can be checked using original definition).

Example 54
$n = 3$
If $L = \text{span}(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix})$ then $P = \begin{pmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 0\end{pmatrix}$.
If $L = \text{span}(\begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix})$ then $P = \begin{pmatrix}1/2 & 1/2 & 0\\1/2 & 1/2 & 0\\0 & 0 & 1\end{pmatrix}$.
If $L = \text{span}(\mathbf{x})$ for some $\mathbf{x} \in \mathbb{R}^n$ then $P = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}$.

Lemma 12
If A is an $n \times n$ projection matrix (i.e. $A = A^T$, $A^2 = A$) of rank r then

1. r of the eigenvalues of A are 1 and $n - r$ are 0,
2. $\text{rank}\, A = \text{trace}\, A$,

Proof Let $\mathbf{x}$ be an eigenvector of A, with eigenvalue λ. Then $\lambda\mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = \lambda A\mathbf{x} = \lambda^2\mathbf{x}$. $\stackrel{\mathbf{x}\neq\mathbf{0}}{\Longrightarrow} \lambda = \lambda^2 \implies \lambda \in \{0, 1\}$.

1. A symmetric $\implies \exists P$ (orthogonal) s.t. $P^{-1}AP = D$, where D is diagonal with 0s and 1s on the diagonal. Since P is non-singular, $\text{rank}\, A = \text{rank}\, D$. Hence D has r ones down the diagonal.
2. $\text{trace}(A) = \text{trace}(APP^{-1}) \stackrel{Le6}{=} \text{trace}(P^{-1}AP) = \text{trace}\, D = \text{rank}\, A$

9.9 Residuals, Estimation of the variance

Definition 20
$\hat{\mathbf{Y}} = X\hat{\beta}$, where $\hat{\beta}$ is a least squares estimator, is called the *vector of fitted values*.

In the full rank case, $\hat{\mathbf{Y}} = X(X^TX)^{-1}X^T\mathbf{Y}$.

Lemma 13
$\hat{\mathbf{Y}}$ is unique and

$$\hat{\mathbf{Y}} = P\mathbf{Y},$$

where P is the projection matrix onto the column space of X.

Because of this lemma, P is sometimes called the *hat matrix* (it puts the hat on $\mathbf{Y}$, i.e. $\hat{\mathbf{Y}} = P\mathbf{Y}$).

Proof Suppose $\hat{\beta}$ is a LSE of β. We already know P is unique, hence $P\mathbf{Y}$ is unique. Thus it suffices to show $\hat{\mathbf{Y}} = P\mathbf{Y}$. Since $P\mathbf{Y} \in \text{span}(X)$ there exists γ s.t. $X\gamma = P\mathbf{Y}$. Then

$$\begin{aligned} S(\hat{\beta}) =& \|\mathbf{Y} - P\mathbf{Y} + P\mathbf{Y} - X\hat{\beta}\|^2 \\ =& \underbrace{\|\mathbf{Y} - P\mathbf{Y}\|^2}_{=S(\gamma)} + \|P\mathbf{Y} - X\hat{\beta}\|^2 + 2\underbrace{\underbrace{(\mathbf{Y} - P\mathbf{Y})^T}_{=\mathbf{Y}^T(I-P)}\underbrace{(P\mathbf{Y} - X\hat{\beta})}_{\in \text{span}(X)}}_{=0} \\ \geq& S(\hat{\beta}) + \|P\mathbf{Y} - X\hat{\beta}\|^2, \end{aligned}$$

since $\hat{\beta}$ minimises S. Thus $\|P\mathbf{Y} - X\hat{\beta}\| = 0$. Therefore, $P\mathbf{Y} = X\hat{\beta}$.

Definition 21
$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ is called the *vector of residuals*.

Remark Equivalent form: Using Lemma 13,

$$\mathbf{e} = \mathbf{Y} - P\mathbf{Y} = Q\mathbf{Y},$$

where $Q = I - P$ is the projection matrix onto $\text{span}(X)^T$.

Remark

$$\mathrm{E}(\mathbf{e}) = \mathrm{E}[Q\mathbf{Y}] = Q\,\mathrm{E}\,\mathbf{Y} = \underbrace{QX}_{=0}\beta = \mathbf{0}$$

$\mathbf{e}$ can be used to see how well the model and the data agree and to see if certain observations are larger or smaller than predicted by the model.

Diagnostic plots:

Suppose the data comes from the model

$$\mathbf{Y} = X\beta + \mathbf{Z}\gamma + \boldsymbol{\epsilon}, \quad \mathsf{E}\,\boldsymbol{\epsilon} = \mathbf{0}$$

where $\mathbf{Z} \in \mathbb{R}^n \setminus \text{span}(X)$ and $\gamma \in \mathbb{R}$ are deterministic, but the analyst erroneously works with

$$\mathbf{Y} = X\beta + \boldsymbol{\epsilon}, \quad \mathsf{E}\,\boldsymbol{\epsilon} = \mathbf{0}.$$

Thus, if $\gamma \neq 0$ then the analyst uses the wrong model. Then

$$\mathsf{E}(\mathbf{e}) = \mathsf{E}(Q\mathbf{Y}) = \mathsf{E}(Q(X\beta + \mathbf{Z}\gamma + \boldsymbol{\epsilon})) = Q\mathbf{Z}\gamma$$

Thus plotting $Q\mathbf{Z}$ against the residuals should give a line (with some noise) through the origin. If the slope of this line is different from 0 then we should consider including $\mathbf{Z}$ in the model.

Example 55 (Plots of Residuals - Brain and Body Weight of Land Mammals) o_i=b<u>o</u>dy weight of the ith animal, r_i=b<u>r</u>ain weight of the ith animal.

Consider the model

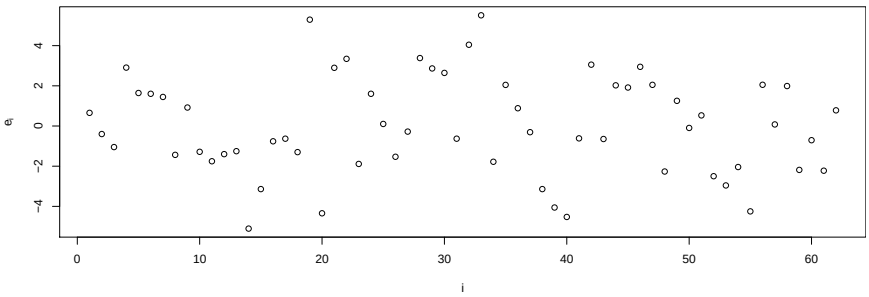
$$\log(r_i) = \beta_1 + \epsilon_i, \quad \mathsf{E}(\epsilon_i) = 0. \tag{1}$$

In matrix form this is

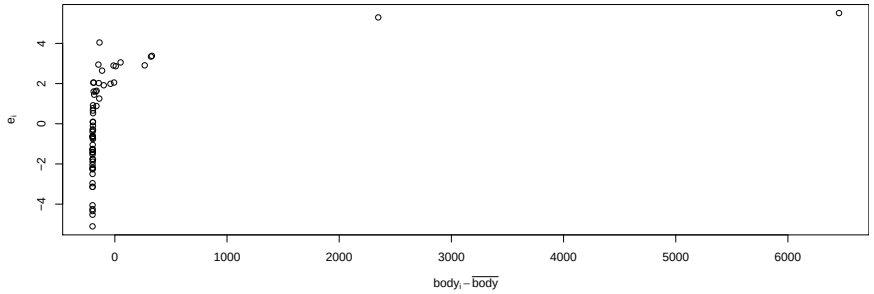
$$\mathsf{E}[\begin{pmatrix} \log(r_1) \\ \vdots \\ \log(r_n) \end{pmatrix}] = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_1.$$

Here,
$P =$

The following is a plot of the residuals in this model:

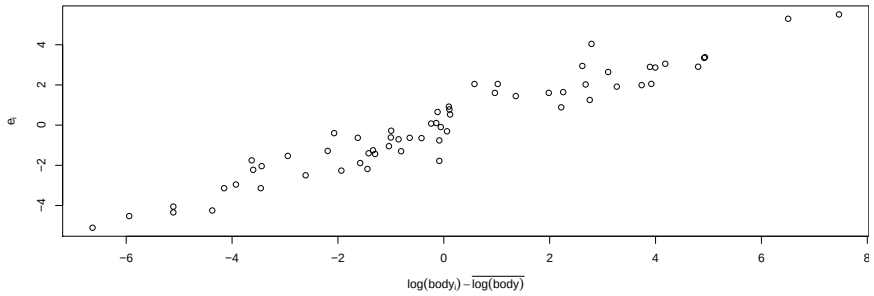


Suppose we suspect $\mathbf{Z} = \mathbf{o}$ might be important. To investigate this, we plot $(Q\mathbf{o})_i = o_i - \bar{o}$ vs e_i. If the model (1) is true then the plot below should roughly look like the previous plot.



The fit of (1) does not seem to be good; however simply including $\mathbf{o}$ in the model does not seem to be reasonable because the above plot does not look like a linear relationship.

Let $z_j = \log(o_j)$. A plot of $(Q\mathbf{z})_i$ vs e_i:



This looks like a linear relationship with slope $\neq 0 \rightarrow$ could include $z_i\beta_2$ in model (1).

Residual Sum of Squares

Definition 22
RSS $= \mathbf{e}^T\mathbf{e}$ is called the *residual sum of squares*.

RSS quantifies the departure of the data from the model. It is the minimum of $S(\beta)$.

Remark Other forms:

- $\text{RSS} = \sum_{i=1}^{n} e_i^2$
- $\text{RSS} = S(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$
- $\text{RSS} = (Q\mathbf{Y})^T Q\mathbf{Y} = \mathbf{Y}^T Q^T Q\mathbf{Y} = \mathbf{Y}^T Q\mathbf{Y}$
- $\text{RSS} = \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$.
  Indeed, $\text{RSS} = (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}^T\mathbf{Y} - 2\hat{\mathbf{Y}}^T\mathbf{Y} + \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} = \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$.
  The last equality holds because $\hat{\mathbf{Y}}^T\mathbf{Y} = (P\mathbf{Y})^T\mathbf{Y} = (PP\mathbf{Y})^T\mathbf{Y} = \mathbf{Y}^T P^T P^T\mathbf{Y} = (P\mathbf{Y})^T P\mathbf{Y} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$.

Theorem 8
$\hat{\sigma}^2 := \frac{\text{RSS}}{n-r}$ is an unbiased estimator of σ^2.

Recall: $r = \text{rank}(X)$

Proof Let $Q = I - P$. Since P is a projection matrix, Q is a projection matrix as well. Hence, $\text{RSS} = Y^T QY$.

$$\begin{aligned}
\mathrm{E}(\text{RSS}) &= \mathrm{E}\,\text{trace}\,\text{RSS} = \mathrm{E}\,\text{trace}(\mathbf{Y}^T Q\mathbf{Y}) \overset{\text{Le 6}}{=} \mathrm{E}\,\text{trace}(Q\mathbf{Y}\mathbf{Y}^T) = \text{trace}(Q\,\mathrm{E}(\mathbf{Y}\mathbf{Y}^T)) \\
&= \text{trace}(Q[\text{cov}\,\mathbf{Y} + \mathrm{E}(\mathbf{Y})\,\mathrm{E}(\mathbf{Y})^T]) = \text{trace}(Q\sigma^2) + \text{trace}(QX\beta(X\beta)^T) \\
&= \sigma^2\,\text{trace}(I - P) + 0 = \sigma^2(n - \text{trace}(P)) \\
&\overset{\text{Le 12}}{=} \sigma^2(n - \text{rank}(P)) = \sigma^2(n - r).
\end{aligned}$$

Remark This is a generalisation of the result that the sample variance s^2 is an unbiased estimator for σ^2 when $Y_1, \ldots, Y_n$ are i.i.d. with unknown mean μ and unknown variance σ^2.

Indeed, we can write this iid setup as the linear model $\mathbf{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \boldsymbol{\epsilon}$ with $\mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$ and $\text{cov}\,\boldsymbol{\epsilon} = \sigma^2 I$. Then $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ and thus $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - P\mathbf{Y} =$

$\mathbf{Y} - \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$. Hence,

$$\frac{\text{RSS}}{n-r} = \underbrace{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}_{=s^2=\text{sample variance}}$$

which we already know is unbiased for σ^2.

Coefficient of Determination (R^2)

In the simplest model with only an intercept term, i.e. in

$$\mathbf{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_1 + \boldsymbol{\epsilon}, \quad \mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$$

we have $\text{RSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Larger models, i.e. models with more columns in X will only lead to smaller RSS.

For models containing an *intercept term*, (i.e. X contains a column consisting of 1s (or any other constant)), a popular measure of the quality of a model is

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

called the *coefficient of determination* or simply R^2. A smaller RSS is "better", thus we want a large R^2. Note: $0 \leq R^2 \leq 1$ and $R^2 = 1$ for a "perfect" model.

Remark (Intuitive interpretation) RSS/n is an estimator of $\boldsymbol{\sigma}^2$. $\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ is an estimator of $\boldsymbol{\sigma}^2$ in the model with only the intercept term (let us call this the "total variance").

Thus $\frac{\text{RSS}/n}{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2} \approx \frac{\text{Variance in the model}}{\text{total variance}}$ and hence

$$R^2 \approx \frac{\text{total variance} - \text{variance in model}}{\text{total variance}}$$

Hence, $R^2 \approx$ fraction of the total variance of the data that "is explained" by the model.

Example 56
Boiling Point: $R^2 = 0.995$

Mammals: Model $\log(\text{brain}_i) = \beta_1 + \log(\text{body}_i)\beta_2 + \epsilon_i$: $R^2 = 0.92$
Model: $\text{brain}_i = \beta_1 + \text{body}_i\beta_2 + \epsilon_i$: $R^2 = 0.87$

Note: These are unusually high values! Often, R^2 can be much smaller.

Remark Adding columns to $\mathbf{X}$ will never decrease R^2. Thus one should not use R^2 directly for model comparisons; one should penalise models with a larger number of parameters. More about this in Chapter 4.

10 Linear Models with Normal Theory Assumptions

In this Chapter we will again consider a linear model $\mathbf{Y} = X\beta + \boldsymbol{\epsilon}$, $\mathbf{E}\epsilon = \mathbf{0}$. In order to construct confidence intervals or test hypotheses we need assumptions about the distribution of $\mathbf{Y}$ (or equivalently about the distribution of $\boldsymbol{\epsilon}$).

We will work with the (NTA), which are $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$.

10.1 Distributional Results

We first define the multivariate normal distribution and some distributions constructed from it. After that some useful properties are shown.

10.1.1 The Multivariate Normal Distribution

The multivariate normal distribution, denoted by $N(\boldsymbol{\mu}, \Sigma)$, is a distribution of a random vector. It has two parameters: one vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and one positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. It will turn out that $\boldsymbol{\mu}$ is its expecation and Σ is its covariance.

It can be defined in several ways. In M2S1 you have defined it via the joint pdf as follows (this definition only works if $\boldsymbol{\Sigma}$ is positive definite):
$\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$ if $\mathbf{Z}$ has a pdf of the form

$$f(\mathbf{z}) = \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z}-\boldsymbol{\mu})\right),$$

where $|\Sigma|$ denotes the determinant of Σ.

Example 57
$\mathbf{Z} \sim N(\boldsymbol{\mu}, \sigma^2 I)$ for some $\sigma^2 > 0$. Then

$$\begin{aligned} f(\mathbf{z}) =& \frac{1}{\sqrt{2\pi}^n |\sigma^{2n}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T (\sigma^{-2} I)(\mathbf{z}-\boldsymbol{\mu})) \\ =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(z_i-\mu_i)^2}{2\sigma^2}) \end{aligned}$$

Thus: $Z_1, \ldots, Z_n$ are independent, with $Z_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$.

The three definitions mentioned below (which are all equivalent) will also work if $\mathbf{\Sigma}$ is only positive semidefinite.

Definition 23

- An n-variate random vector $\mathbf{Z}$ follows a multivariate normal distribution if for all $\mathbf{a} \in \mathbb{R}^n$ the random variable $\mathbf{a}^T\mathbf{Z}$ follows a univariate normal distribution (the degenerate case $N(\cdot, 0)$ is allowed).
- Let $X_1, \dots X_r \sim N(0,1)$ be iid, let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times r}$. Then $\mathbf{Z} = A\mathbf{X} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, AA^T)$.
- $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$ if its characteristic function $\phi : \mathbb{R}^n \to \mathbb{C}$, $\phi(\mathbf{t}) = \mathrm{E}(\exp(i\mathbf{Z}^T\mathbf{t}))$ satisfies
$$\phi(\mathbf{t}) = \exp\left(i\boldsymbol{\mu}^T\mathbf{t} - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right) \quad \forall t \in \mathbb{R}^n.$$
where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix.

Remark (Concerning the definition via the characteristic function) You have previously met the moment generating function. One of the key results about moment generating functions is that the moment generating function uniquely identifies the distribution of a random variable.

Similarly, there is a moment generating function defined for n-variate random vectors $\mathbf{X}$, namely $M : \mathbb{R}^n \to \mathbb{R}, \mathbf{t} \mapsto \mathrm{E}(\exp(\mathbf{t}^T\mathbf{X}))$. Again, M identifies the distribution of a random vector.

The characteristic function (which was used in the above definition), is often used instead of the moment generating function. It has similar properties (in particular it uniquely defines a distribution). The i in it is the complex number i. Furthermore, if $Z = X + iY$ is a complex-valued random variable then $\mathrm{E}(Z) := \mathrm{E}(X) + i\,\mathrm{E}(Y)$.

Remark (Useful properties) Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$. Then

- $\mathrm{E}\,\mathbf{Z} = \boldsymbol{\mu}$,
- $\mathrm{cov}\,\mathbf{Z} = \Sigma$,
- if A is a deterministic matrix and $\mathbf{b}$ is a deterministic vector of appropriate dimension then
$$A\mathbf{Z} + \mathbf{b} \sim N(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T).$$

In general: if X and Y are random variables then $\operatorname{cov}(X, Y) = 0$ does not imply that X and Y are independent. To put it briefly: uncorrelated does not imply independence. The following lemma shows (in a general form) that *uncorrelated and jointly normal* does imply independence.

Lemma 14
For $i = 1, \dots, k$, let $A_i \in \mathbb{R}^{n_i \times n_i}$ be pos. semidef. and symmetric and let $\mathbf{Z}_i$ be an n_i-variate random vector. If $\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_k \end{pmatrix} \sim N(\boldsymbol{\mu}, \Sigma)$, for some $\boldsymbol{\mu} \in \mathbb{R}^{\sum_{i=1}^k n_i}$ and $\Sigma = \operatorname{diag}(A_1, \dots, A_k) = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_k \end{pmatrix}$ then $\mathbf{Z}_1, \dots, \mathbf{Z}_k$ are independent.

Proof In the special case that all A_i are positive definite, this can be shown by using $\Sigma^{-1} = \operatorname{diag}(A_1^{-1}, \dots, A_k^{-1})$ and $|\Sigma| = \prod_{i=1}^k |A_i|$ to factor the pdf.
The full proof works via the characteristic (or via the moment generating) functions; to show independence one needs to show that the characteristic functions can be written as product of the characteristic functions of the components, i.e. one needs to show $\mathrm{E} \exp(it^T \mathbf{Z}) = \prod_{i=1}^k \mathrm{E} \exp(i \mathbf{t}_i^T \mathbf{Z}_i)$ for all $\mathbf{t} = (\mathbf{t}_1^T, \dots, \mathbf{t}_k^T)^T \in \mathbb{R}^n$.

Example 58
Let $k = 3$, $A_1 = 2 = (2)$, $A_2 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$, $A_3 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Let

$$\Sigma = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & -0.5 & 0 & 0 \\ 0 & -0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}.$$

If $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$ for some $\mu \in \mathbb{R}^5$ then Z_1, $\begin{pmatrix} Z_2 \\ Z_3 \end{pmatrix}$, $\begin{pmatrix} Z_4 \\ Z_5 \end{pmatrix}$ are independent.

10.1.2 Distributions derived from the Multivariate Normal

In this section we define several distributions derived from the multivariate normal distribution. They will appear as distributions of pivotal quantities and test statistics. You should have met the central χ^2- and t-distribution previously.

Definition 24
Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, I_n)$, where $\boldsymbol{\mu} \in \mathbb{R}^n$.
$\mathbf{U} = \mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^n Z_i^2$ is said to have a *non-central* χ^2*-distribution* with n degrees of freedom (d.f.) and non-centrality parameter

$$\delta = \sqrt{\boldsymbol{\mu}^T\boldsymbol{\mu}}.$$

Notation: $\mathbf{U} \sim \chi_n^2(\delta)$, $\chi_n^2 = \chi_n^2(0)$.

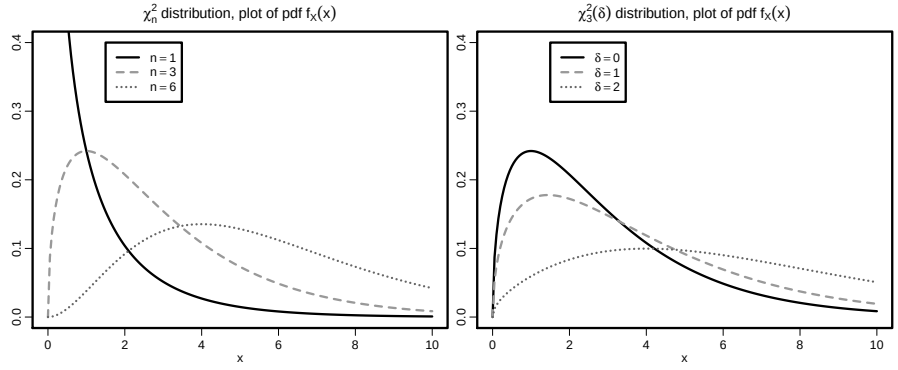
Remark In order for this to be a proper definition we need to show that the distr. of U depends on $\boldsymbol{\mu}$ only through $\boldsymbol{\mu}^T\boldsymbol{\mu}$. (One of the questions on the problem sheet asks you to prove this). One approach is to show that the moment generating function of U equals

$$M_U(t) = \frac{1}{(1-2t)^{n/2}} \exp\left(\frac{t\delta^2}{1-2t}\right)$$

The following lemma contains some properties of the χ^2-distribution.

Lemma 15
Let $U \sim \chi_n^2(\delta)$. Then $\mathrm{E}(U) = n + \delta^2$ and $\mathrm{Var}(U) = 2n + 4\delta^2$.
If $U_i \sim \chi_{n_i}^2(\delta_i)$, $i = 1, \ldots, k$ and U_i's are indep. then $\sum_{i=1}^k U_i \sim \chi^2_{\sum n_i, \sqrt{\sum \delta_i^2}}$.

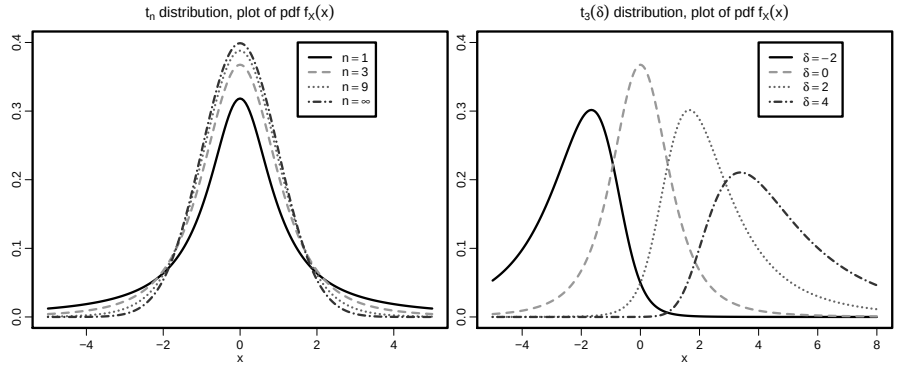
Proof: $\rightarrow$ Problem Sheet.



Definition 25
Let X and U be independent random variables with $X \sim N(\delta, 1)$ and $U \sim \chi^2_n$. The distribution of

$$Y = \frac{X}{\sqrt{U/n}}$$

is called *non-central t-distribution* with n degrees of freedom and non-centrality parameter δ.
Notation: $Y \sim t_n(\delta)$, $t_n = t_n(0)$.



Remark Convergence to the normal distribution: Suppose $Y_n \sim t_n$ for all $n \in \mathbb{N}$. Then

$$Y_n \xrightarrow{d} N(0,1) \quad (n \to \infty)$$

To see this: Let $X \sim N(0,1)$ and $U_n = \sum_{i=1}^n Z_i^2$ for $Z_1, Z_2, \cdots \sim N(0,1)$ indep. Then, by the weak law of large numbers: $U_n/n \xrightarrow{P} E(Z_1^2) = 1$. As $x \mapsto \sqrt{x}$ is continuous,

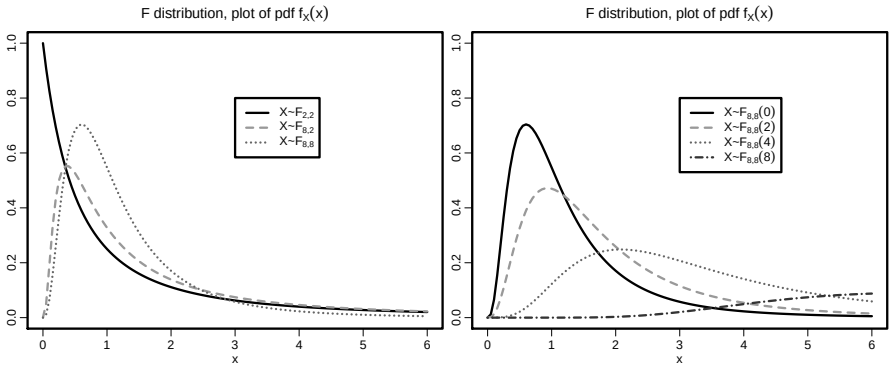
this implies $\sqrt{U_n/n} \xrightarrow{P} \sqrt{1} = 1$. Thus, by Slutsky's Lemma:

$$Y_n = \frac{X}{\sqrt{U_n/n}} \xrightarrow{d} \frac{X}{1} = X \sim N(0,1).$$

Definition 26
If $W_1 \sim \chi^2_{n_1}(\delta)$, $W_2 \sim \chi^2_{n_2}$ independently then

$$F = \frac{W_1/n_1}{W_2/n_2}$$

is said to have a *non-central F* distribution with (n_1, n_2) d.f. and n.c.p.$=\delta$.
Notation: $F \sim F_{n_1,n_2}(\delta)$, $F_{n_1,n_2} = F_{n_1,n_2}(0)$.



Remark If the n.c.p. $\delta = 0$ then the above are called *central* χ^2, *t, F distribution*.

10.1.3 Some Independence Results

In order to show that a random variable follows a *t*-distribution via the definition we need to show independence between a normal and a χ^2-distributed random variable. Similarly, to show that a random variable is *F*-distributed, we need to show independence of two χ^2 distributed random variables. This section provides results that help in doing this.

Lemma 16
Let $A \in \mathbb{R}^{n\times n}$ be a pos. semidefinite symmetric matrix with rank r. Then there exists $L \in \mathbb{R}^{n\times r}$ such that $\operatorname{rank} L = r$, $A = LL^T$ and $L^TL = \operatorname{diag}(\text{nonzero eigenvalues of } A)$.

Proof (Lemma 8) $\Longrightarrow$ $\exists$ an orthogonal matrix P st

$$P^TAP = D = \operatorname{diag}(\text{eigenvalues of } A)$$

Precisely r elements of D are positive and the others are zero. Hence, $A = PDP^T = PD^{\frac{1}{2}}D^{\frac{1}{2}}P^T = PD^{\frac{1}{2}}(PD^{\frac{1}{2}})^T$ Let L consist of the nonzero columns of $PD^{\frac{1}{2}}$. Then $A = LL^T$.

Because P is orthogonal (i.e. $P^TP = I$) we get $L^TL = \operatorname{diag}(\text{nonzero eigenvalues of } A)$.

Lemma 17
Let $\mathbf{X} \sim N(\boldsymbol{\mu}, I)$, $A \in \mathbb{R}^{n,n}$ pos. semidefinite symmetric and let B be a matrix such that $BA = 0$.
Then $\mathbf{X}^TA\mathbf{X}$ and $B\mathbf{X}$ are independent.

Proof Let $r = \operatorname{rank}(A)$. By Lemma 16, $\exists\ L \in \mathbb{R}^{n\times r}$ such that $\operatorname{rank} L = r$ and $A = LL^T$.

$$\operatorname{cov}(B\mathbf{X}, L^T\mathbf{X}) = B\operatorname{cov}(\mathbf{X})L = BL = BLL^TL(L^TL)^{-1} = BAL(L^TL)^{-1} = 0$$

Thus $B\mathbf{X}$ and $L^T\mathbf{X}$ are independent (because they are jointly normally distributed).

Hence, $B\mathbf{X}$ and $\mathbf{X}^TLL^T\mathbf{X} = \mathbf{X}^TA\mathbf{X}$ are independent.

Lemma 18
If $\mathbf{Z} \sim N(\boldsymbol{\mu}, I_n)$ and A is an $n \times n$ projection matrix of rank r, then

$$\mathbf{Z}^TA\mathbf{Z} \sim \chi^2_r(\delta) \quad \text{with } \delta^2 = \boldsymbol{\mu}^TA\boldsymbol{\mu}$$

Proof All nonzero eigenvalues of A are equal to 1.

By Lemma 16, $\exists\ L \in \mathbb{R}^{n\times r}$ such that $A = LL^T$ and $L^TL = I_r$. Let $\mathbf{V} = L^T\mathbf{Z}$. Then $\mathbf{V} \sim N(L^T\boldsymbol{\mu}, \underbrace{I}_{=L^TL})$ and

$$\mathbf{Z}^TA\mathbf{Z} = \mathbf{Z}^TLL^T\mathbf{Z} = \mathbf{V}^T\mathbf{V} \sim \chi^2_r(\delta),$$

where $\delta^2 = \left(L^T\boldsymbol{\mu}\right)^T L^T\boldsymbol{\mu} = \boldsymbol{\mu}^T\underbrace{LL^T}_{=A}\boldsymbol{\mu} = \boldsymbol{\mu}^TA\boldsymbol{\mu}$.

> **Lemma 19**
> If $\mathbf{Z} \sim N(\boldsymbol{\mu}, I_n)$ and $A_1,\ A_2 \in \mathbb{R}^{n\times n}$ are projection matrices and $A_1A_2 = 0$ then $\mathbf{Z}^TA_1\mathbf{Z}$ and $\mathbf{Z}^TA_2\mathbf{Z}$ are independent.

Proof $\operatorname{cov}(A_1\mathbf{Z}, A_2\mathbf{Z}) = A_1\underbrace{\operatorname{cov}(\mathbf{Z},\mathbf{Z})}_{=I}A_2^T = A_1A_2^T = 0$

Because $A_1\mathbf{Z}$ and $A_2\mathbf{Z}$ are jointly normally distributed this shows that they are independent.

As $\mathbf{Z}^TA_i\mathbf{Z} = (A_i\mathbf{Z})^T(A_i\mathbf{Z})$ for $i = 1, 2$ (symm+ idempotent) this implies that $\mathbf{Z}^TA_1\mathbf{Z}$ and $\mathbf{Z}^TA_2\mathbf{Z}$ are independent.

This result extends to $\mathbf{Z}^TA_1\mathbf{Z}, \dots, \mathbf{Z}^TA_k\mathbf{Z}$, where $A_iA_j = 0$ $(i \neq j)$.

> **Lemma 20**
> If $A_1, \dots, A_k$ are symmetric $n \times n$ matrices such that $\sum A_i = I_n$ and if rank $A_i = r_i$ then the following are equivalent:
>
> 1. $\sum r_i = n$
> 2. $A_iA_j = 0$ for all $i \neq j$
> 3. A_i is idempotent for all $i = 1, \dots, k$.

Proof (2)→(3) $\forall j$: $A_1 + \cdots + A_k = I_n \implies A_1A_j + \cdots + A_kA_j = A_j \implies A_j^2 = A_j$.

(3)→(1) $n = \text{trace}\, I_n = \sum \text{trace}\, A_i \overset{12.}{=} \sum \text{rank}\, A_i = \sum r_i$

(1)→(2) Let $V_i = \{A_i\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\} = \text{span}(A_i)$. Then $\dim V_i = r_i$. Let B_i be a basis for V_i and let $B = \bigcup_i B_i$. Since $\mathbf{x} = I\mathbf{x} = \sum A_i\mathbf{x} \forall \mathbf{x} \in \mathbb{R}^n$, B spans $\mathbb{R}^n$ and since B has at most $\sum r_i = n$ elements, B must form a basis of $\mathbb{R}^n$. Hence, any $\mathbf{x} \in \mathbb{R}^n$ can be written uniquely as $\sum \mathbf{u}_i$ where $\mathbf{u}_i \in V_i$. Let $\mathbf{x}$ be a column of A_j. Then $\underbrace{\mathbf{x}}_{\in V_j} + \sum_{i \neq j} \mathbf{0} = \sum A_i\mathbf{x}$. By uniqueness, $A_i\mathbf{x} = \mathbf{0}$ for all $i \neq j$.

Theorem 9 (The Fisher-Cochran Theorem)
If $A_1, \ldots, A_k$ are $n \times n$ projection matrices such that $\sum_{i=1}^n A_i = I_n$, and if $\mathbf{Z} \sim N(\boldsymbol{\mu}, I_n)$ then $\mathbf{Z}^T A_1 \mathbf{Z}, \ldots, \mathbf{Z}^T A_k \mathbf{Z}$ are independent and

$$\mathbf{Z}^T A_i \mathbf{Z} \sim \chi^2_{r_i}(\delta_i), \quad \text{where } r_i = \text{rank}\, A_i \text{ and } \delta_i^2 = \boldsymbol{\mu}^T A_i \boldsymbol{\mu}.$$

Proof By Lemma 20, $A_i A_j = 0$ for all $i \neq j$. Hence, $\mathbf{Z}^T A_1 \mathbf{Z}, \ldots, \mathbf{Z}^T A_k \mathbf{Z}$ are independent.

The rest of the theorem is a consequence of Lemma 18.

10.2 The Linear Model with Normal Theory Assumptions

In this section we will consider the linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ with (NTA).

Recall that the (NTA) are $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$. In particular, this implies $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$. The joint probability density function of $\mathbf{Y}$ is thus

$$f(\mathbf{y}) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right)$$

Estimation using the maximum likelihood approach:

- The log-likelihood of the data is

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

- maximising L with respect to β (for fixed σ^2) is equivalent to minimising $S(\beta) = (\mathbf{Y} - X\beta)^T(\mathbf{Y} - X\beta)$, i.e. maximum likelihood is equivalent to least squares for estimating β.
- The maximum likelihood estimator for σ^2 is RSS $/n$ (look at first/second derivative of $L(\hat{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}$ RSS wrt σ^2).

10.3 Confidence Intervals, Tests for one-dimensional quantities

The following gives a pivotal quantity for σ^2 which can be used to construct CIs or tests.

Lemma 21 (Distribution of RSS)
Assume (NTA). Then

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-r}$$

where $r = \operatorname{rank} X$.

Proof Let P denote the projection matrix onto span(X). Then

$$\text{RSS} = \mathbf{e}^T\mathbf{e} = (\underbrace{(I-P)}_{=:Q}\mathbf{Y})^T(I-P)\mathbf{Y} = \mathbf{Y}^T \underbrace{Q^TQ}_{=QQ=Q}\mathbf{Y} = \mathbf{Y}^TQ\mathbf{Y}$$

and Q is the projection onto the space orthogonal to the columns of X. Hence,

$$\frac{\text{RSS}}{\sigma^2} = \frac{\mathbf{Y}}{\sigma}^T Q\frac{\mathbf{Y}}{\sigma} = \mathbf{Z}^TQ\mathbf{Z}$$

where $\mathbf{Z}\mathbf{Y}/\sigma \sim N(X\beta/\sigma, I)$ and Q is a projection matrix. Furthermore, $Q + P = I$ and Q and P are projection matrices. Thus, by Lemma 20, $\operatorname{rank} Q + \underbrace{\operatorname{rank} P}_{=r} = n$, implying $\operatorname{rank} Q = n - r$.

Thus, by Lemma 18,

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-r}$$

since $(X\beta/\sigma)^T \underbrace{QX}_{=0}\beta/\sigma = 0$.

Often, we will be interested in parts of the parameter vector β, e.g. one of its components β_i. We want to construct tests, or construct confidence intervals.

The following lemma provides a flexible way of doing this; it gives a pivotal quantity for $\mathbf{c}^T\beta$ for some (known) $\mathbf{c} \in \mathbb{R}^p$. We have worked with $\mathbf{c}^T\beta$ already in the Gauss-Markov Theorem.

Lemma 22
Assume (FR), (NTA) in a linear model. Let $\mathbf{c} \in \mathbb{R}^p$. Then

$$\frac{\mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta}{\sqrt{\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\frac{\text{RSS}}{n-p}}} \sim t_{n-p}$$

Proof Since $\mathbf{c}^T\hat{\beta} = \mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Y}$ and $\mathbf{Y} \sim N(X\beta, \sigma^2 I)$,

$$\mathrm{E}\,\mathbf{c}^T\hat{\beta} = \mathbf{c}^T\beta$$

$$\begin{aligned}\mathrm{Var}(\mathbf{c}^T\hat{\beta}) =& \mathrm{Var}(\mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Y}) = \mathbf{c}^T(X^TX)^{-1}X^T\,\mathrm{cov}(\mathbf{Y})X(X^TX)^{-1}\mathbf{c}\\ =&\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2\end{aligned}$$

and thus $\mathbf{c}^T\hat{\beta} \sim N(\mathbf{c}^T\beta, \mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2)$. Hence,

$$\frac{\mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta}{\sqrt{\mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2}} \sim N(0,1).$$

We already know $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.

Thus the lemma is a consequence of the definition of the t-distribution once we have shown that $\hat{\beta} = (X^TX)^{-1}X^T\mathbf{Y}$ and $\text{RSS} = \mathbf{Y}^TQ\mathbf{Y}$ are independent. The latter is a consequence of Lemma 17 (using $\mathbf{Z} = \mathbf{Y}/\sigma$), since $(X^TX)^{-1}X^TQ = (X^TX)^{-1}\underbrace{(QX)}_{=0}^T = 0$.

Example 59
Data Set: Tooth Growth (see File in Additional Material)

Remark Suppose we construct a tests with the above pivotal quantity for $\mathbf{c}^T\boldsymbol{\beta}$. It turns out that the test statistic has a non-central t-distribution under the alternative.

10.4 The F-Test

In the previous section, we derived pivotal quantities for one-dimensional parameters (σ^2 or linear combinations $\mathbf{c}^T\boldsymbol{\beta}$ of the components of $\boldsymbol{\beta}$ such as, for some i, $\mathbf{e}_i^T\boldsymbol{\beta} = \beta_i$). If we are interested in how more than one component of the parameter behaves, e.g. if the null-hypotheses $\beta_2 = \beta_3 = 0$ is of interest then we would have to do more than one test (and this would result in similar problems as the "joint confidence intervals" mentioned earlier and a correction such as the Bonferroni correction would be necessary). This section presents a method to test more complicated hypotheses about $\boldsymbol{\beta}$.

Example 60
Suppose we have a linear model with $p = 3$ and design matrix

$$X = \begin{pmatrix} 1 & a_1 & b_1 \\ \vdots & \vdots & \vdots \\ 1 & a_n & b_n \end{pmatrix}$$

Suppose we are interested in testing the hypotheses

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

Under H_0, we can write the linear model as

$$\mathrm{E}\, Y = X_0\beta_1, \qquad \text{where } X_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Thus we can rewrite the hypotheses as

$$H_0 : \mathrm{E}\,\mathbf{Y} \in \operatorname{span}(X_0) \quad \text{against} \quad H_1 : \mathrm{E}\,\mathbf{Y} \notin \operatorname{span}(X_0)$$

In general, suppose we want to test whether a sub-model of a linear model $\mathrm{E}\,\mathbf{Y} = X\beta$ is true, i.e. we want to test

$$H_0 \ : \ \mathrm{E}\,\mathbf{Y} \in \mathrm{span}(X_0) \text{ against } H_1 \ : \ \mathrm{E}\,\mathbf{Y} \notin \mathrm{span}(X_0)$$

for some matrix X_0 with $\mathrm{span}(X_0) \subset \mathrm{span}(X)$. In other words, the null hypothesis says that the sub-model

$$\mathrm{E}(\mathbf{Y}) = X_0\beta_0$$

is true.

Example 61
Continuing the previous example, one may also be interested in $X_0 = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{pmatrix}$ [equivalent to $\beta_3 = 0$] or $X_0 = \begin{pmatrix} 1 & a_1 - b_1 \\ \vdots & \vdots \\ 1 & a_n - b_n \end{pmatrix}$ [equivalent to $\beta_3 = -\beta_2$].

Let

- RSS= the residual sum of squares in the full model $\mathrm{E}\,\mathbf{Y} = X\beta$
- RSS= the residual sum of squares in the sub-model $\mathrm{E}\,\mathbf{Y} = X_0\tilde{\beta}$

Lemma 23
Under $H_0 \ : \ \mathrm{E}\,\mathbf{Y} \in \mathrm{span}(X_0)$,

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}} \cdot \frac{n-r}{r-s} \sim F_{r-s,n-r}$$

where $r = \mathrm{rank}\, X$, $s = \mathrm{rank}\, X_0$.

Let P be the projection matrix onto $\mathrm{span}\, X$, and let $Q = I - P$ the projection matrix onto $(\mathrm{span}\, X)^{\perp}$.
Likewise, let P_0 be the projection matrix onto $\mathrm{span}\, X_0$, and let $Q_0 = I - P_0$ the projection matrix onto $(\mathrm{span}\, X_0)^{\perp}$. Then, as we derived in the previous chapter,

$$\mathrm{RSS} = \mathbf{Y}^T Q\mathbf{Y}, \quad \mathrm{RSS}_0 = \mathbf{Y}^T Q_0\mathbf{Y}.$$

Using this gives

$$\begin{aligned} F =& \frac{\mathbf{Y}^T Q_0 \mathbf{Y} - \mathbf{Y}^T Q \mathbf{Y}}{\mathbf{Y}^T Q \mathbf{Y}} \cdot \frac{n-r}{r-s} \\ =& \frac{\mathbf{Y}^T (P - P_0)\mathbf{Y}/\sigma^2}{\mathbf{Y}^T (I - P)\mathbf{Y}/\sigma^2} \cdot \frac{n-r}{r-s} \end{aligned}$$

To show: numerator$\sim \chi^2_{r-s}$, denominator$\sim \chi^2_{n-r}$, numerator and denominator are independent.

Proof We will use the Fisher-Cochran theorem. Let $\mathbf{Z} = \mathbf{Y}/\sigma$, $A_1 = I - P$, $A_2 = P - P_0$, $A_3 = P_0$.

Clearly, $A_1 + A_2 + A_3 = I$. We already know that A_1 and A_3 are projection matrices.

To show: $A_2 = P - P_0$ is a projection matrix. $P - P_0$ is symmetric as P_0 and P are both symmetric. Furthermore,

$$(P - P_0)^2 = P^2 + P_0^2 - PP_0 - P_0 P$$

Every column $\mathbf{y}$ of P_0 is an element of span(X_0) and thus an element of span(X). Thus, $P\mathbf{y} = \mathbf{y}$.
Hence,

$$PP_0 = P_0.$$

This also implies $P_0 P = (P^T P_0^T)^T = (PP_0)^T = P_0^T = P_0$.

Thus,

$$(P - P_0)^2 = P + P_0 - P_0 - P_0 = P - P_0$$

The Fisher-Cochran theorem now implies

- $\mathbf{Z}^T (P - P_0)\mathbf{Z}$ and $\mathbf{Z}^T (I - P)\mathbf{Z}$ are independent,
- $\mathbf{Z}^T (P - P_0)\mathbf{Z} \sim \chi^2_{\text{rank}(P-P_0)}(\mathsf{E}\,\mathbf{Z}^T (P - P_0)\,\mathsf{E}\,\mathbf{Z})$,
- $\mathbf{Z}^T (I - P)\mathbf{Z} \sim \chi^2_{\text{rank}(I-P)}(\mathsf{E}\,\mathbf{Z}^T (I - P)\,\mathsf{E}\,\mathbf{Z})$.

Next, we show that the non-centrality parameters are 0.

Under H_0, we know $\mathsf{E}\,\mathbf{Z} = \frac{1}{\sigma}\mathsf{E}\,\mathbf{Y} \in \text{span}(X_0) \subset \text{span}(X)$ Thus,

$$(P - P_0)\,\mathsf{E}\,\mathbf{Z} = \underbrace{P\,\mathsf{E}\,\mathbf{Z}}_{=\mathsf{E}\,\mathbf{Z}} - \underbrace{P_0\,\mathsf{E}\,\mathbf{Z}}_{=\mathsf{E}\,\mathbf{Z}} = \mathbf{0}.$$

Hence, $\mathrm{E}\,\mathbf{Z}^T(P - P_0)\,\mathrm{E}\,\mathbf{Z} = 0$. Furthermore,

$$E\mathbf{Z}^T(I - P)\,\mathrm{E}\,\mathbf{Z} = E\mathbf{Z}^T(\mathrm{E}\,\mathbf{Z} - \underbrace{P\,\mathrm{E}\,\mathbf{Z}}_{=\mathrm{E}\,\mathbf{Z}}) = 0.$$

Concerning the degrees of freedom:

- By Lemma 20,

  $$n = \mathrm{rank}(P) + \mathrm{rank}(I - P) = \mathrm{rank}\,X + \mathrm{rank}(I - P) = r + \mathrm{rank}(I - P)$$

  Thus, $\mathrm{rank}(I - P) = n - r$.

- Using Lemma 20 again,

  $$\underbrace{\mathrm{rank}(P_0)}_{=\mathrm{rank}(X_0)=s} + \mathrm{rank}(P - P_0) + \underbrace{\mathrm{rank}(I - P)}_{n-r} = n$$

  Thus, $\mathrm{rank}(P - P_0) = r - s$.

To summarise, we have shown

$$\mathbf{Z}^T(P - P_0)\mathbf{Z} \sim \chi^2_{r-s},\ \mathbf{Z}^T(I - P)\mathbf{Z} \sim \chi^2_{n-r} \text{ independently.}$$

Thus, by definition, $F \sim F_{r-s,n-r}$.

If H_0 is not true then the proof is still valid, except for the non-centrality parameter of $\mathbf{Z}^T(P - P_0)\mathbf{Z}$. Now,

$$\mathrm{E}\,\mathbf{Z}^T(P - P_0)\,\mathrm{E}\,\mathbf{Z} = \frac{1}{\sigma^2}\boldsymbol{\beta}^T X^T(P - P_0)X\boldsymbol{\beta}.$$

Thus, without assuming H_0, we get

$$F \sim F_{r-s,n-r}(\delta),\ \text{where } \delta^2 = \frac{1}{\sigma^2}(X\boldsymbol{\beta})^T(P - P_0)X\boldsymbol{\beta}.$$

This implies that F will take on larger values if H_0 is not true.

Thus it is advisable to reject for large values of F. In particular, if we want a test to the level $\alpha > 0$, we reject if

$$F > c,$$

where c is such that $\mathrm{P}(X \geq c) = \alpha$ for $X \sim F_{r-s,n-r}$.

10.5 Confidence Regions

Suppose $\mathrm{E}\,\mathbf{Y} = X\beta$ is a linear model satisfying (FR), (NTA). We are interested in finding a confidence region for β, i.e. we want a random set D s.t. $\mathrm{P}(\beta \in D) \geq 1-\alpha$ for all β, σ^2.

Let

$$A = \frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{\mathrm{RSS}} \cdot \frac{n-p}{p}$$

If we can work out the distribution of A, we can use A as a pivotal quantity for β.

The numerator of the first fraction can be rewritten as

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

where P is the projection onto the space spanned by the columns of X. Indeed,

$$(\mathbf{Y}-X\beta)^T P(\mathbf{Y}-X\beta) = (\mathbf{Y}-X\beta)^T PP(\mathbf{Y}-X\beta) = (P(\mathbf{Y}-X\beta))^T P(\mathbf{Y}-X\beta)$$

Using $P = X(X^TX)^{-1}X^T$ this is equal to

$$(X(\hat{\beta}-\beta))^T X(\hat{\beta}-\beta)$$

Furthermore, RSS can be written as

$$\mathrm{RSS} = \mathbf{Y}^T Q\mathbf{Y} = (\mathbf{Y}-X\beta)^T Q(\mathbf{Y}-X\beta)$$

where $Q = I - P$. Thus, letting $\mathbf{Z} = \frac{1}{\sigma}(\mathbf{Y}-X\beta)$,

$$A = \frac{\mathbf{Z}^T P\mathbf{Z}}{\mathbf{Z}^T Q\mathbf{Z}} \cdot \frac{n-p}{p}$$

with $Z \sim N(\mathbf{0}, I)$, $P+Q=I$, $\operatorname{rank} P = p$, P and Q projection matrices.

Thus the Fisher-Cochran theorem shows that $A \sim F_{p,n-p}$.

Hence, a $1-\alpha$ confidence region R for β is defined by all $\gamma \in \mathbb{R}^p$ such that

$$\frac{(\hat{\beta}-\gamma)^T X^T X(\hat{\beta}-\gamma)}{\mathrm{RSS}} \cdot \frac{n-p}{p} \leq F_{p,n-p,\alpha},$$

where $\mathrm{P}(Z \geq F_{p,n-p,\alpha}) = \alpha$ for $Z \sim F_{p,n-p}$. The region R is an ellipsoid centred at $\hat{\beta}$ (use diagonalisation).

Remark General definition of an ellispoid: $\{\mathbf{z} \in \mathbb{R}^p : (\mathbf{z}-\mathbf{z}_0)^T A^{-1}(\mathbf{z}-\mathbf{z}^0) \leq 1\}$ where A is pos. def. and $\mathbf{z}_0 \in \mathbb{R}^p$.

Example 62
$p = 2$, X has full rank.

Let $\mathbf{a} = \hat{\beta} - \beta$, $B = X^T X$ and $c = p\frac{\text{RSS}}{n-p} F_{p,n-p,\alpha}$. Hence we want to find for which $\mathbf{a}$,

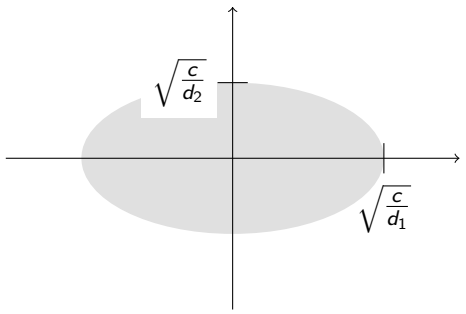
$$\mathbf{a}^T B \mathbf{a} \leq c \tag{2}$$

B is pos. def. Hence, $\exists$ an orthogonal matrix R and a diagonal matrix $D = \text{diag}(d_1, d_2)$ s.t. $B = R^T D R$ and d_1, d_2 are positive. [D consists of the eigenvalues and R consists of the corresponding normalised eigenvectors.]

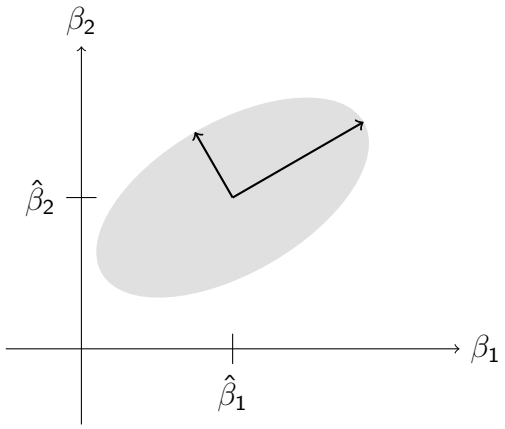
Let $\tilde{\mathbf{a}} = R\mathbf{a}$ (this rotates the coordinate axes). Then (2) is equivalent to

$$\tilde{a}_1^2 \frac{d_1}{c} + \tilde{a}_2^2 \frac{d_2}{c} \leq 1.$$

This describes an ellipse with half-axes of lenght $\sqrt{\frac{c}{d_1}}$ and $\sqrt{\frac{c}{d_2}}$.



Transforming everything back via $\beta = \hat{\beta} - \mathbf{a} = \hat{\beta} - R^T \tilde{\mathbf{a}}$ gives a rotated and translated ellipse, centered at $\hat{\beta}$.



Remark We could construct individual CIs for β_1 and β_2 via Lemma 22 and combine them via the Bonferroni correction. The advantage of the above construction is that the resulting ellipsoid has a smaller area.

11 Diagnostics, Model Selection, Extensions

11.1 Outliers

An *outlier* is an observation that does not conform to the general pattern of the rest of the data. Potential causes:

- Error in the data recording mechanism (example - iron content of spinach).
- Data set may be "contaminated" - i.e. it may be the mixture of two or more populations.
- Indication that the model/underlying theory needs to be improved. Further investigations needed. Outliers may be the "most interesting" observations.

Practical method for spotting outliers: Look for residuals that are "too large". When is a residual too large?

Recall: $\mathbf{e} = (I - P)\mathbf{Y}$, where P is the projection onto span(X). If X is full rank then $P = X(X^T X)^{-1} X^T$. Note that

$$\operatorname{cov} \mathbf{e} = (I - P) \operatorname{cov} \mathbf{Y} (I - P)^T = \sigma^2 (I - P)$$

and $\mathrm{E}\, \mathbf{e} = \mathbf{0}$. Thus, unter NTA, $e_i \sim N(0, \sigma^2(1 - P_{ii}))$, where P_{ii} is the ith diagonal element of P. Hence,

$$\frac{e_i}{\sqrt{(1 - P_{ii})\sigma^2}} \sim N(0, 1).$$

We do not know $\sigma^2 \rightarrow$ plug in the unbiased estimate

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p}.$$

This gives the *standardised* residuals

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 (1 - P_{ii})}}$$

This of course means that r_i are not (necessarily) N(0,1) distributed.

Nevertheless, the standardized residuals should be roughly independent, and their distribution should be close to a N(0,1)-distribution.

Remark Let X N(0,1). Then the probabilities for larges values of X are very rapidly decreasing as the following table shows. [The normal distribution has *light tails*.]

x	3	4	5	6	7	8
$P(X > x)$	1.3e-03	3.2e-05	2.9e-07	9.9e-10	1.3e-12	6.2e-16

Thus if (NTA) holds then the standardized residuals should be relatively small.

Remark r_i is also *not* student t; the usual proof (that we used for t-tests) does not work because $\hat{\sigma}^2$ and e_i are not independent.

Remark An entire branch of statistics, called "robust statistics", is concerned with the development of methods/statistics that give meaningful results even in the presence of outliers.

Suppose we are interested in the "centre" of a distribution and have observations $x_1, \ldots x_n$. The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is heavily affected by outliers - in fact, just one outlier can make $\bar{x}$ take any value. Other statistics are far more *robust* to outliers. For example the median cannot be changed arbitrarily by one outlier [you would have to move half of the observations.]

11.2 Leverage

What is the potential impact of individual observations on the model fit?

$$\text{cov}(\mathbf{e}) = \sigma^2 (I_n - P)$$

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

> **Definition 27**
> The *leverage* of the ith observation in a linear model is P_{ii}, the ith diagonal matrix of the hat matrix P.

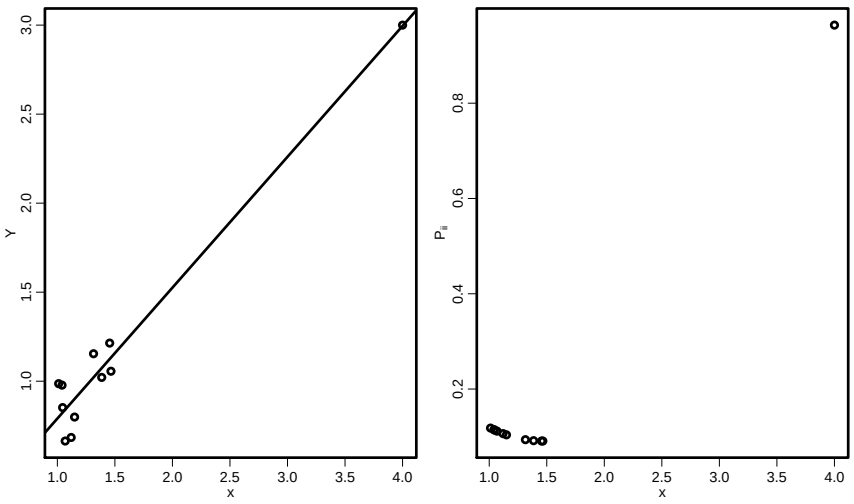
(Recall: P is the projection matrix onto span(X), where X is the design matrix).

If $P_{ii} \approx 1$ then the variance of the ith residual is very low. This is totally determined by X, i.e. the design matrix is forcing the model fit to be good at the covariates of the ith observation. In this case the ith observation is said to have **high leverage**.

$\sum_{i=1}^{n} P_{ii} = \text{trace}(P) = \text{rank}(X) =: r$ (see Lemma 12), so the "average" is r/n and a rule of thumb is to take notice when

$$p_{ii} > \frac{2r}{n}.$$

Example 63 (Linear regression)
E $Y_i = \beta_1 + \beta_2 x_i$



11.3 Cook's Distance

To measure how much the ith observation changes the estimator $\hat{\beta}$ one can consider the following measure, called Cook's distance:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p\,\text{RSS}/(n-p)},$$

where $\hat{\beta}_{(i)}$ is the least squares estimator with the ith observation removed. Alternatively,

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p\,\text{RSS}/(n-p)},$$

where $\hat{\mathbf{Y}}_{(i)} = X\hat{\beta}_{(i)}$. Rule of thumb: take notice if D_i gets close to 1.

Algebraically equivalent expression:

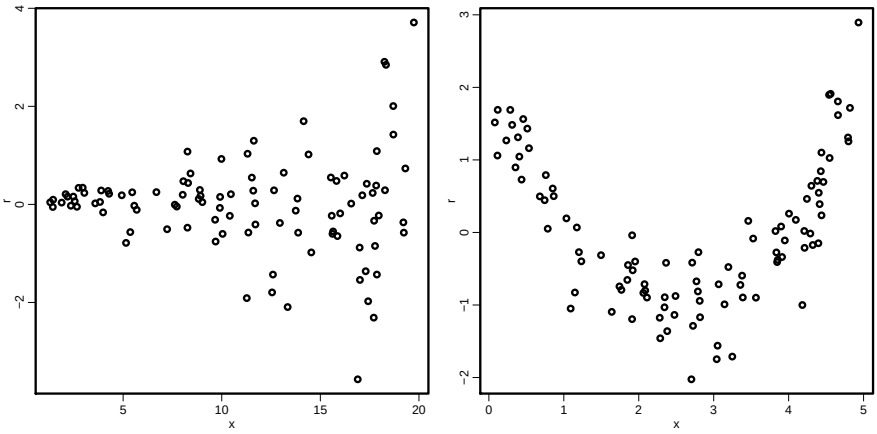
$$D_i = r_i^2 \frac{P_{ii}}{(1 - P_{ii})r},$$

where r_i is the standardised residual and $r = \text{rank}(X)$. Cook's distance *combines leverage and residual*.

11.4 Residual Plots

Goal: To detect problems with a model; in other words: to detect a lack of fit of a model:
Approach: Plot standardised residuals against some other variable (e.g. a column of X, potentially interesting additional covariates, $\hat{Y}$, ...)

If the model is correct then the resulting plots should just show "noise", with no distinct patterns.



The left plot suggests a non-constant variance (a heteroscedastic error).

The right hand plot indicates that the covariate may have a nonlinear influence.

11.5 Distributional Checks

Is the normal theory assumption justified? So-called Quantile-Quantile plots (qq-plots) can plots can reveal if it is not. [note: it can never prove that it is correct]

This is because essentially we try to test the following hypotheses:

$$H_0: \text{ model correct} \quad \text{against} \quad H_1 \text{model not correct}$$

Not rejecting H_0 is no evidence that H_0 is correct.

Basic idea of the Quantile-Quantile plots:

Consider two cdfs F and G. Consider the curve

$$(0,1) \to \mathbb{R}^2, t \mapsto \begin{pmatrix} F^{-1}(t) \\ G^{-1}(t) \end{pmatrix}$$

If $F = G$ this gives a line through the origin.

Check if data comes from a specific distribution If one wants to check if observed iid data $Y_1, \dots, Y_n$ come from a fixed distribution G one replaces F by the *empirical cdf*

$$F_n : \mathbb{R} \to [0,1], \quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{I}(Y_i \leq x)$$

where $\mathsf{I}(Y_i \leq x) = \begin{cases} 1 & Y_i \leq x \\ 0 & \text{otherwise} \end{cases}$. Then, if $Y_1, \dots, Y_n \sim F$, by the strong law of larger numbers

$$F_n(x) \to F(x) \quad (n \to \infty) \text{a.s.}$$

and so we can use F_n as consistent estimator for F.

Check if data is normally distributed

Often, one does not want to check if data comes from a fixed distribution but whether it comes from a given class, e.g. a normal distribution.

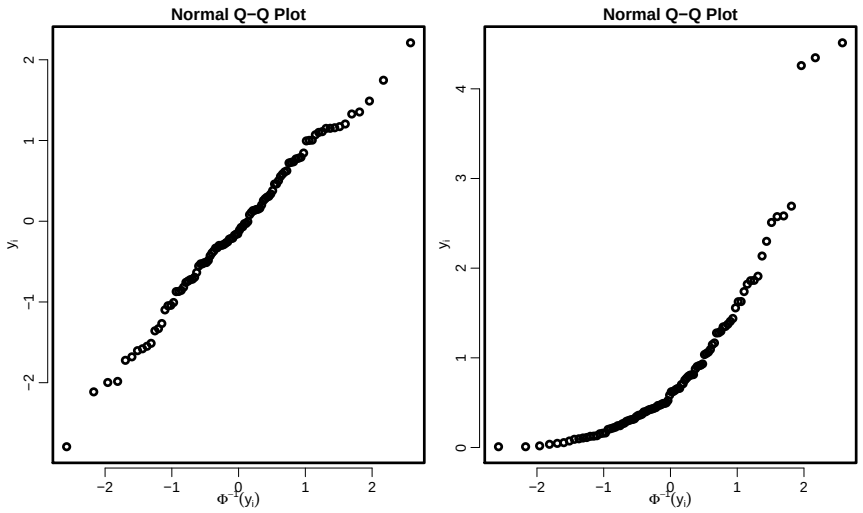
suppose $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ and $G = \Phi$ where Φ is the cdf of N(0,1). $F^{-1}(t) = \sigma\Phi^{-1}(t) + \mu$ and $G^{-1}(t) = \Phi^{-1}(t)$. Hence, the qq-plot is

$$(0,1) \to \mathbb{R}^2, t \mapsto \begin{pmatrix} F^{-1}(t) \\ G^{-1}(t) \end{pmatrix} = \begin{pmatrix} \sigma\Phi^{-1}(t) + \mu \\ \Phi^{-1}(t) \end{pmatrix}$$

which is a line.

To apply to real data, F is replaced by the empirical cdf F_n and then, if the observations are coming from a normal distribution, the qq plot should show a straight line.

Example 64



n=100; LHS: $Y_i \sim N(0,1)$; RHS: $Y_i \sim Exp(1)$.

Use standardised residuals to check assumption of normality in the linear model.

(Similar plots can also be used for other (specific) distributions)

More formal approaches: Goodness-of-fit tests.

$$H_0 : Y \sim F, \quad \text{vs} \quad H_1 : Y \not\sim F$$

Observations $Y_1, \ldots, Y_n \to$ Empirical cdf F_n
Test statistic (Kolmogorov-Smirnov test):

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Reject H_0 for large values.

11.6 Weighted Least Squares

So far we have assumed $\text{cov}(\mathbf{Y}) = \sigma^2 I_n$. Now suppose $\text{cov}(\mathbf{Y}) = \sigma^2 V$, where V is known, symmetric and positive definite.

Example 65
Var $Y_i \propto b_i^2$, Y_i's uncorrelated. Then $V = \begin{pmatrix} b_1^2 & & 0 \\ & \ddots & \\ 0 & & b_n^2 \end{pmatrix}$.

How to estimate β? What is a BLUE in this situation? Main idea: transform the model to a situation in which (SOA) hold true, i.e. in which $\text{cov}(\epsilon) = \sigma^2 I$.

V is symmetric and positive definite. There $\exists$ a nonsingular matrix T such that $T^T V T = I_n$ and $TT^T = V^{-1}$. Indeed, by Lemma 8 , $\exists$ an orthogonal matrix P and a diagonal matrix D with the eigenvalues of V on the diagonal s.t.

$$P^T V P = D$$

Let $T = PD^{-1/2}P^T$. Since P is orthogonal, $V = PDP^T$ and thus

$$T^T V T = PD^{-1/2}P^T PDP^T PD^{-1/2}P^T = I.$$

Furthermore, $TT^T = PD^{-1}P^T = (P^T)^{-1}D^{-1}P^{-1} = (PDP^T)^{-1} = V^{-1}$.

Let $\mathbf{Z} = T^T\mathbf{Y}$. Then

$$\text{E}(\mathbf{Z}) = \underbrace{T^T X}_{=:\tilde{X}}\beta \quad \text{and} \quad \text{cov}(\mathbf{Z}) = T^T V T \sigma^2 = \sigma^2 I_n$$

Thus the linear model $\text{E}\,\mathbf{Z} = \tilde{X}\beta$ satisfies (SOA). Assuming (FR), we get the following least squares estimator.

$$\begin{aligned} \hat{\beta} &= [\tilde{X}^T\tilde{X}]^{-1}\tilde{X}^T\mathbf{Z} \\ &= [X^T(TT^T)X]^{-1}X^T(TT^T)\mathbf{Y} \\ &= (X^T V^{-1} X)^{-1}X^T V^{-1}\mathbf{Y}. \end{aligned}$$

Note: $\hat{\beta}$ is an optimal estimator in the sense of the Gauss-Markov theorem.

11.7 Under/overfitting

Underfitting=necessary predictors left out

Let

$$\mathbf{Y} = X\beta + \mathbf{Z}\gamma + \boldsymbol{\epsilon}$$

be the model the observations have come from and let

$$\mathbf{Y} = X\beta + \boldsymbol{\epsilon}$$

be the fitted model.

Suppose we are interested in estimating $\mathbf{c}^T\beta$. $\hat{\beta}$ is biased:

$$\mathrm{E}\,\hat{\beta} = (X^TX)^{-1}X^T(X\beta + \mathbf{Z}\gamma) = \beta + (X^TX)^{-1}X^T\mathbf{Z}\gamma$$

Hence,

$$MSE(\mathbf{c}^T\hat{\beta}) = \mathrm{Var}(\mathbf{c}^T\hat{\beta}) + (E(\mathbf{c}^T\hat{\beta} - \mathbf{c}^T\beta))^2 = \mathbf{c}^T(X^TX)^{-1}\mathbf{c}\sigma^2 + (\mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Z})^2\gamma^2$$

Let $(\hat{\beta}^F, \hat{\gamma})^T$ be the estimator in the full model. Then

$$\mathrm{cov}(\begin{pmatrix}\hat{\beta}^F \\ \hat{\gamma}\end{pmatrix}) = \sigma^2 \begin{pmatrix}X^TX & X^T\mathbf{Z} \\ \mathbf{Z}^TX & \mathbf{Z}^T\mathbf{Z}\end{pmatrix}^{-1}$$

Formulas for the inverse of 2×2 block matrices are known in the literature (see e.g. the Matrix cookbook at http://matrixcookbook.com/). Using these we get

$$\mathrm{cov}(\hat{\beta}^F) = \sigma^2(X^TX)^{-1} + \sigma^2(X^TX)^{-1}X^T\mathbf{Z}(\mathbf{Z}^TQ\mathbf{Z})^{-1}\mathbf{Z}^TX(X^TX)^{-1},$$

where Q is the projection matrix onto $(\mathrm{span}\, X)^{\perp}$. Hence,

$$\mathrm{Var}(\mathbf{c}^T\hat{\beta}^F) = \sigma^2\mathbf{c}^T(X^TX)^{-1}\mathbf{c} + \frac{\sigma^2}{\mathbf{Z}^TQ\mathbf{Z}}(\mathbf{c}^T(X^TX)^{-1}X^T\mathbf{Z})^2$$

Hence, if $\frac{\sigma^2}{\mathbf{Z}^TQ\mathbf{Z}} > \gamma^2$ then the estimator from the reduced model has the smaller MSE.

Hence, the mean squared error can be improved by omitting covariates...

Sometimes it pays to use a simpler model!

Overfitting = unnecessary predictors included.

This means that some of the components of β are 0. Estimator $\hat{\beta}$ is unbiased; however the variance will be larger than in a model where these predictors are left out.

11.8 Model Selection

What covariates/predictors to include in a model?

Example 66 (Quine data)
Aitkin(1978) discussed an observational study of S. Quine. Town in Australia. 146 childrens in the study.
Response: Number of days children where absent from school in a year.

4 factors:

- Age (primary, first, second or third form)
- Ethnicity (aboriginal or non-aboriginal)
- slow or average learner
- male or female

$4 \cdot 2 \cdot 2 \cdot 2 = 32$ combinations of factors

Want to use a model with fewer parameters (reduce MSE of parameter estimates one is interested in, better understanding of the model).

How many models? Essentially 2^{32} models possible.

Why not just minimize RSS? Including more predictors will always decrease the RSS but not the MSE (see previous section). Bias-Variance tradeoff (more parameters - higher variance and lower bias).

Criterion developed by Akaike (he called it "an information criterion"):

$$\text{AIC} = -2\log(L) + 2p,$$

where p is the number of parameters in the model, and L is the maximum of the likelihood function. In the linear model,

$$\text{AIC} = n\log(RSS/n) + 2p + \text{constant},$$

Want: model with small AIC.
There are many other criteria: BIC, Mallow's C_p, LASSO, ...

Search Strategies:

- Best subset (search all possible subsets)
  usually not possible: #of subsets $= (2^{\#\text{of predictors}})$
- Stepwise (forward, backward, both): Start with an initial model and change it using small steps (adding/deleting a predictor) always aiming to decrease the the criterion.

Current field of research: What to do if there is an extremely large number of (potential) covariates (often $p >> n$)?
Example: How does the DNA influence the occurence of diseases/survival/obesity (microarray data)?

11.9 Generalized Linear Models (GLMs)

Classical book: McCullagh and Nelder (1989).

$$g(\mathrm{E}\, Y_i) = \sum_j X_{ij}\beta_j, \quad i = 1, \ldots, n,$$

where g is the so-called link-function. One assumes that the distribution of Y_i is contained in a so-called exponential familiy of distributions. Examples of exponenial families: Normal, Exponential, Gamma, Poisson, Binomial distributions.

The models are usually fitted by the maximum likelihood method.

Example 67
Suppose $Y_1, \ldots, Y_n$ are Bernoulli outcomes. We could formulate the linear model

$$\mathrm{E}\, Y_i = \sum_j X_{ij}\beta_j.$$

One of the problems is that there is nothing in the model that enforces $\sum_j X_{ij}\beta_j \in [0, 1]$.

A popular choice for thee link function for Bernoulli outcomes is the so-call logit function given by $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$. This is called a *logit*-model.

An alternative link function is the link function $l = \Phi$ i.e. the cdf of a standard normal random variable, - such a model is called a probit model.