2S(3) Statistics – Part II – Spring 2026

Lecture Notes by Biman Chakraborty

January 16, 2026

Contents

Chapter 1

Point Estimation

1.1 Introduction

In this chapter we discuss fitting the parameters, or the ***estimation*** of the parameters, for various distributions.

We are going to discuss two modes of estimation here. The first is ***point estimation***, which involves the use of sample data to arrive at a single value (known as a statistic) which is to serve as a "best guess" for an unknown (fixed or random) population parameter. Point estimation can be contrasted with ***interval estimation***, which is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter.

As mentioned before, ***point estimation*** utilises the sample data to arrive at a single value which is to serve as a "best guess" for an unknown parameter.

Basic setup: Consider a sample $X_1, X_2, \ldots, X_n$ whose (joint) distribution depends on the (unknown) parameter θ, scalar or vector. (Recall that by a sample we mean independent and identically distributed (i.i.d.) observations from some distribution.) Assume that the X_i's are i.i.d. with the distribution $F(x|\theta)$.

Definitions:

- An ***estimator*** of θ is a function of $X_1, X_2, \ldots, X_n$ i.e. a random variable.
- Numerical value of a realisation of an estimator is an ***estimate***.
- The probability distribution of an estimator is called its ***sampling distribution***.
- The variability of an estimator is the estimate of the standard deviation of its sampling distribution, or ***standard error (SE)***.
- When the expectation of the estimator is the true value of θ, it is called an ***unbiased estimator of*** θ.
- When the estimate converges in some sense to the true value of θ, we say that it is a ***consistent estimator*** of θ in that sense.

There are many ways to obtain a point estimate. Some of the most well-known estimates are

a) **Method of Moments Estimator (MME)**

b) **Maximum Likelihood Estimator (MLE)**

c) **Minimum Variance Unbiased Estimator (MVUE)**

d) **Best Linear Unbiased Estimator (BLUE)**

e) **Minimum Mean Squared Error Estimator (MMSE)**

We shall discuss the first two in detail here, and briefly the third.

1.2 Method of Moments

Recall that the k-th ***raw moment*** of a probability distribution F is defined as $\mu_k = E\left(X^k\right)$, where the distribution of X is F.

Definition 1.1. *If $X_1, X_2, \ldots, X_n$ are i.i.d. with distribution F, then the k-th* **sample moment** *is defined as $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.*

Therefore, $\hat{\mu}_k$ **can be used as an estimate of** μ_k.

Parameters are estimated by the method of moments by finding expressions in terms of moments and then substituting corresponding sample moments. By the law of large numbers (LLN), such estimates are *consistent*, i.e. they converge to the true parameter as the sample size gets larger.

Example 1.1 (Method of moment estimation of the Poisson parameter): Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson with unknown parameter λ. We now describe how a method of moment estimate of λ can be obtained using $X_1, X_2, \ldots, X_n$.

Solution: The first moment of $X \sim \text{Poisson}(\lambda)$ is $E(X) = \lambda$. The first sample moment is $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$. Hence λ can be estimated by $\hat{\lambda} = \bar{X}$. We say that $\bar{X}$ **is a method of moment estimator (MME) of** λ. We shall see later that the method of moment estimators do not need to be unique.

Functions of λ can be estimated by the corresponding functions of the MME. For example, $P(X = 0) = e^{-\lambda}$ can be estimated by $e^{-\bar{X}}$ here.

Since the X_i's are i.i.d. Poisson(λ), $n\hat{\lambda} = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\lambda)$. Hence $E\left(\hat{\lambda}\right) = \frac{E\left(\sum_{i=1}^{n} X_i\right)}{n} = \frac{n\lambda}{n} = \lambda$, implying that $\bar{X}$ is unbiased for λ. Also see that $\bar{X} \to \lambda$ by the LLN. Hence it is also a consistent estimator of λ.

Further, note that $V\left(\hat{\lambda}\right) = \frac{V\left(\sum_{i=1}^{n} X_i\right)}{n^2} = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$ which may be estimated by $\frac{\hat{\lambda}}{n} = \frac{\bar{X}}{n}$.

Therefore an estimate of the standard error (SE) of $\bar{X}$ is $\sqrt{barX/n}$.

Example 1.2 (Method of moment estimation of the parameters of a normal distribution): Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$ with unknown parameters μ and σ^2.

The first two moments of $X \sim N\left(\mu, \sigma^2\right)$ are $E(X) = \mu$; $E\left(X^2\right) = \mu^2 + \sigma^2$.

Therefore, $\hat{\mu} = \bar{X}$; $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

One may show that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and

$$\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Finally, note that $\bar{X}$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ are **independent.** We skip the proof of this interesting result.

Example 1.3 (**Method of moment estimation of the parameters of a gamma distribution**) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. gamma with parameters α and λ. The first two moments are $\mu_1 = \frac{\alpha}{\lambda}$ and $\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$, which could be rewritten as $\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$; $\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$.

Since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$, $\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2}$; $\hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$. Obtaining the distributions of the estimates is a bit tricky in this case, and we omit that.

Example 1.4 We now show the example of a non-standard distribution. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. discrete random variables with the p.m.f.

$$p(x) = \begin{cases} p & \text{if } x = 0 \\ 2p & \text{if } x = 1 \\ 1 - 3p & \text{if } x = 2 \end{cases}$$

Since $\mu_1 = E(X) = 2 - 4p$, we have $\hat{p} = \frac{1}{4}\left(2 - \hat{\mu}_1\right) = \frac{1}{4}\left(2 - \bar{X}\right)$.

1.3 Maximum Likelihood Estimation

We are now going to discuss another method of estimation based on the 'likelihood' of the obtained sample. For that we need to introduce the idea of likelihood first. Towards that, let us denote the p.m.f. at x by $p(x|\theta)$ if the model is discrete and the p.d.f. by $f(x|\theta)$ (if the model is continuous).

Now suppose that we have an i.i.d. sample $X_1, X_2, \ldots, X_n$.

Definition 1.2. *The* **likelihood** *of an i.i.d. sample $X_1, X_2, \ldots, X_n$ is defined as,*

$$\ell(\theta|X_1, X_2, \ldots, X_n) = \begin{cases} \prod_{i=1}^{n} p(X_i|\theta) & \textit{discrete case} \\ \prod_{i=1}^{n} f(X_i|\theta) & \textit{continuous case} \end{cases}$$

Notation: To simplify notations, we drop the $X_1, X_2, \ldots, X_n$ part from the notation for likelihood and **write it simply as** $\updownarrow(\theta)$ when there is no scope of confusion.

This likelihood tells us about the model given the data. For example, if we had two possible models we might prefer the one with the ***higher likelihood***. This leads to the concept of ***maximum likelihood estimation*** of any parameter e.g. probability p in a binomial model.

Definition 1.3. *The* **maximum likelihood estimate (MLE)** *of θ is the value of θ, say $\hat{\theta}$, that maximises the likelihood, i.e. makes the data "most probable" or "most likely".*

The MLE is consistent for the true parameter when certain conditions are satisfied; (which will usually be satisfied in the cases of our interest.)

Rather than maximising the likelihood itself, it is often easier to maximise its logarithm (which is an equivalent problem because logarithm is a strictly increasing function). The log-likelihood is

$L(\theta) = \log_e \ell(\theta) = \sum_{i=1}^n \log_e f(X_i|\theta)$.

If $L(\theta)$ is twice differentiable, the MLE $\hat{\theta}$ satisfies:

$$\left.\frac{dL}{d\theta}\right|_{\theta=\hat{\theta}} = 0 \quad and \quad \left.\frac{d^2L}{d\theta^2}\right|_{\theta=\hat{\theta}} < 0$$

Example 1.5 (Maximum likelihood estimation of the Bernoulli parameter p

Sample: $X_1, X_2, \ldots, X_n$ *i.i.d. Bernoulli(p).*

Bernoulli probability function: $p(x|\theta) = \theta^x (1-\theta)^{1-x}$ for $x = 0,\ 1$.

Likelihood function: $\ell(\theta) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}$

Log-likelihood function: $L(\theta) = \log_e \ell(\theta) = \log_e(\theta)\sum_{i=1}^n X_i + \log_e(1-\theta)(n - \sum_{i=1}^n X_i)$

Now, $\frac{dL(p)}{\mathrm{dp}} = \frac{\sum_{i=1}^n X_i}{p} - \frac{n-\sum_{i=1}^n X_i}{1-p}$, so $\frac{dL(p)}{\mathrm{dp}} = 0$ implies $\hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$, the proportion of 1's in the sample $X_1, X_2, \ldots, X_n$.

Since $\frac{d^2L(p)}{dp^2} = -\frac{\sum_{i=1}^n X_i}{p^2} - \frac{n-\sum_{i=1}^n X_i}{(1-p)^2} < 0$ for any p, $\left.\frac{d^2L(p)}{dp^2}\right|_{p=\bar{X}} < 0$ and hence $\bar{X}$ is the MLE.

Example 1.6 (Maximum likelihood estimation of the Poisson parameter) Sample: $X_1, X_2, \ldots, X_n$ *i.i.d.* Poisson(λ). Assume that at least one of them is positive.

Probability mass function: $p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots,$

Likelihood function: $\ell(\lambda) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{X_i}}{X_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!}$

Log-likelihood function:

$$L(\lambda) = \log_e \ell(\lambda) = -n\lambda + (\log_e \lambda)\sum_{i=1}^n X_i - \log_e\left(\prod_{i=1}^n X_i!\right)$$

Now, $\frac{dL(\lambda)}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^n X_i - n$, which is equal to 0 for $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$.

Since $\left.\frac{d^2L(\lambda)}{d\lambda^2}\right|_{\lambda=\bar{X}} = -\ \frac{1}{\lambda^2}\sum_{i=1}^n X_i\big|_{\lambda=\bar{X}} = -\frac{n\bar{X}}{\bar{X}^2} = -\frac{n}{\bar{X}} < 0$, maximises the likelihood and hence is the MLE for λ.

Note 1: $-\frac{n}{\bar{X}} < 0$as long as $\bar{X} > 0$, i.e. at least one of is $X_1, X_2, \ldots, X_n$ positive.)

Note 2: The MLE and the MME (method of moments estimator) are the same for Poisson and hence have the same sampling distribution.

Example 1.7 (Maximum likelihood estimation of the exponential parameter): Sample: $X_1, X_2, \ldots, X_n$ *i.i.d.* Exponential(λ). Assume that at least one of them is positive. Here the distribution is continuous with p.d.f. $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x > 0$

Likelihood function:

$$\ell(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n \exp\left(-\lambda\sum_{i=1}^n X_i\right)$$

Log-likelihood function:

$$L(\lambda) = \log_e \ell(\theta) = n \log_e \lambda - \lambda \sum_{i=1}^{n} X_i$$

Now, $\frac{dL(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i$, which is equal to zero for $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean.

Also, $\frac{d^2L(\lambda)}{d\lambda^2} = -\frac{n}{\lambda^2}$, and hence $\left.\frac{d^2L(\lambda)}{d\lambda^2}\right|_{\lambda=1/\bar{X}} = -n\bar{X}^2 < 0$. Therefore $\hat{\lambda} = \frac{1}{\bar{X}}$ is the MLE.

Example 1.8 (Example 1.4 continued): To compute the likelihood, we start by writing the p.m.f. in a compact form:

$$p(x) = p^{I(x=0)} (2p)^{I(x=1)} (1-3p)^{I(x=2)}$$

$$\ell(p) = \prod_{i=1}^{n} p^{I(X_i=0)} (2p)^{I(X_i=1)} (1-3p)^{I(X_i=2)} = p^{n_1} (2p)^{n_2} (1-3p)^{n_3}$$

where $n_1 = \#(X_i = 0) = \sum_{i=1}^{n} I(X_i = 0)$, $n_2 = \#(X_i = 1)$, $n_3 = \#(X_i = 2)$, and $n_1 + n_2 + n_3 = n$.

The log-likelihood would therefore be

$$L(p) = \log(p)\, n_1 + \log(2p)\, n_2 + \log(1-3p)\, n_3 = n_2 \log 2 + (n_1 + n_2)\log(p) + n_3 \log(1-3p)$$

So $\frac{dL(p)}{\mathrm{dp}} = \frac{n_1+n_2}{p} - \frac{3n_3}{1-3p}$.

Hence, putting $\frac{dL(p)}{\mathrm{dp}} = 0$ gives $\frac{n_1+n_2}{p} = \frac{3n_3}{1-3p}$, or $\frac{n_1+n_2}{p} = \frac{3n_3}{1-3p} = \frac{3(n_1+n_2)+3n_3}{3p+1-3p} = 3n$, (as $n_1 + n_2 + n_3 = n$,) implying $\hat{p} = \frac{n_1+n_2}{3n}$. $\frac{d^2L(p)}{dp^2} = -\frac{n_1+n_2}{p^2} - \frac{9n_3}{(1-3p)^2} < 0$ for any $p \neq 0, \frac{1}{3} \Rightarrow \left.\frac{d^2L(p)}{dp^2}\right|_{p=\hat{p}} < 0 \Rightarrow \hat{p}$ is the MLE if $\neq 0, \frac{1}{3}$.

Note: See that in Example 1.1.4 we have seen that the MME of p is $\frac{(2-\bar{X})}{4}$.

Now, $\bar{X} = \frac{(0\times n_1 + 1\times n_2 + 2\times n_3)}{n} = (n_2 + 2n_3)/n$, so here $MME = \frac{1}{4}\left(2 - \bar{X}\right) = \frac{1}{4n}(2n - (n_2 + 2n_3)) = \frac{1}{4n}(2n_1 + n_2)$, is different from the MLE.

1.4 Maximum Likelihood Estimation of Two Parameters

Let us now consider a probability function represented by $f(x; \theta_1, \theta_2)$ with θ_1 and θ_2 two unknown parameters. Then, given i.i.d. sample $X_1, X_2, \ldots, X_n$, the likelihood is

$$\ell(\theta_1, \theta_2 | X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i; \theta_1, \theta_2)$$

and hence the log-likelihood is

$$L(\theta_1, \theta_2) = \log_e \ell(\theta_1, \theta_2) = \sum_{i=1}^{n} \log_e f(X_i; \theta_1, \theta_2)$$

The maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of the parameters θ_1 and θ_2 then satisfy:

$$\left.\frac{\partial L\left(\theta_1,\theta_2\right)}{\partial\theta_1}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}=0 \text{ and } \left.\frac{\partial L\left(\theta_1,\theta_2\right)}{\partial\theta_2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}=0$$

There are also some conditions on the second derivatives to ensure that a maximum occurs at $(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)$

The matrix

$$\begin{bmatrix} -\left.\frac{\partial^2 L}{\partial\theta_1^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)} & -\left.\frac{\partial^2 L}{\partial\theta_1\partial\theta_2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)} \\ -\left.\frac{\partial^2 L}{\partial\theta_1\partial\theta_2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)} & -\left.\frac{\partial^2 L}{\partial\theta_2^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)} \end{bmatrix}$$

must be positive definite,

i.e. $\left.\frac{\partial^2 L}{\partial\theta_1^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}<0$; $\left.\frac{\partial^2 L}{\partial\theta_2^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}<0$ and

$$\left.\frac{\partial^2 L}{\partial\theta_1^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}\times\left.\frac{\partial^2 L}{\partial\theta_2^2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}-\left[\left.\frac{\partial^2 L}{\partial\theta_1\partial\theta_2}\right|_{(\theta_1,\theta_2)=\left(\hat{\theta}_1,\hat{\theta}_2\right)}\right]^2>0.$$

Example 1.9 (MLE of the parameters of a normal distribution): In this example, we obtain the mle of the two normal parameters, μ and σ^2.

If $X_1,X_2,\ldots,X_n$ are i.i.d. $N\left(\mu,\sigma^2\right)$ then the likelihood function is

$$\ell\left(\mu,\sigma|X_1,X_2,\ldots,X_n\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(X_i-\mu\right)^2\right)$$

and hence the log likelihood is

$$L\left(\mu,\sigma\right)=-n\log\sigma-\frac{n}{2}\log\left(2\pi\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

The partial derivatives of the log likelihood with respect to μ and σ are respectively

$$\frac{\partial L}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right) \text{ and } \frac{\partial L}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}\left(X_i-\mu\right)^2$$

Setting the partial derivatives equal to zero, we get following estimates: $\hat{\mu}=\bar{X}$ and $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$. That these are the MLEs can be verified by calculating the matrix of second derivatives and verifying that the matrix is negative definite. The details are left for interested students.

1.5 Departure from the Identical Distribution Scenario

The random variables do not need to be i.i.d. to compute maximum likelihood estimates.

Example 1.10: Suppose that males in a society die independently within age 60 with probability p_1, and the females die within age 60 with probability p_2, where $p_2 = p_1 + d$.

We observe the records of n males and m females and try to estimate p_1 and d from that. We can assume that the people die independently.

Let $X_i = 1$ if the i-th male in our records has died within age 60, and 0 otherwise. Then X_i is a Bernoulli random variable with probability p_1, $i = 1, 2, \ldots, n$.

Further, let $Y_j = 1$ if the j-th female in our records has died within age 60, and 0 otherwise. Then Y_j is also a Bernoulli random variable, but with probability p_2, $j = 1, 2, \ldots, m$

Likelihood function in this case is the product of all probability functions:

$$\ell\left(p_1, p_2 | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right) = p_1^{\sum_{i=1}^n X_i} \left(1 - p_1\right)^{n - \sum_{i=1}^n X_i} p_2^{\sum_{j=1}^m Y_j} \left(1 - p_2\right)^{m - \sum_{j=1}^m Y_j}$$

Since $p_2 = p_1 + d$, we can re-write the likelihood as

$$\ell\left(p_1, d | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right) = p_1^{\sum_{i=1}^n X_i} \left(1 - p_1\right)^{n - \sum_{i=1}^n X_i} \left(p_1 + d\right)^{\sum_{j=1}^m Y_j} \left(1 - p_1 - d\right)^{\ m - \sum_{j=1}^m Y_j}$$

The log likelihood:

$$L\left(p_1, d\right)$$

$$= \log p_1 \sum_{i=1}^n X_i + \log\left(1 - p_1\right)\left(n - \sum_{i=1}^n X_i\right) + \log\left(p_1 + d\right) \sum_{j=1}^m Y_j + \log\left(1 - p_1 - d\right)\left(m - \sum_{j=1}^d Y_j\right)$$

Now,

$$\frac{\partial L}{\partial p_1} = \frac{\sum_{i=1}^n X_i}{p_1} - \frac{n - \sum_{i=1}^n X_i}{1 - p_1} + \frac{\sum_{j=1}^m Y_j}{p_1 + d} - \frac{m - \sum_{j=1}^d Y_j}{1 - p_1 - d} \frac{\partial L}{\partial d} = \frac{\sum_{j=1}^m Y_j}{p_1 + d} - \frac{m - \sum_{j=1}^d Y_j}{1 - p_1 - d}$$

Setting the partial derivatives to 0, we get $\hat{p}_1 = \bar{X}$; $\hat{d} = \bar{Y} - \bar{X}$. That these are the MLEs can be verified by looking at the matrix of the second derivatives. Again, the details are omitted from here.

1.6 Optimality of a Point Estimate

The point estimates could be anything, for example the number zero is always a valid point estimate for any parameter; so is the number 1 million! Hence, one needs to use some criteria by which an estimate could be considered better than another. (The MLE is unique in most situations but it is not obvious why it is any good.) One popular criterion to consider an estimate to be good is unbiasedness, another is consistency.

In general, unbiased estimates are considered as good estimates. But they are not unique, either. For example, both sample mean and variance are unbiased estimates for the Poisson parameter; (prove it!) Hence we need something more. One commonly used optimality criterion is to consider the unbiased estimate with the minimum variance, i.e. the minimum variance unbiased estimator (MVUE), which is unique. To prove the uniqueness we need to use some basic property of covariance; the proof is outside the scope of this class.

The following theorem gives a lower bound on the variance of unbiased estimators, which in turn helps us to find the MVUE.

Theorem 1.1 (Crammer-Rao Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. with probability function $f(x;\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x;\theta)$,*

$$Var(T) \geq \frac{1}{nI(\theta)};\ \textit{where } I(\theta) = E\left(\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right)^2.$$

Under smoothness conditions on f, $I(\theta)$ is also equal to $-E\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta)\right)$, and hence

$$-E\left(\frac{\partial^2}{\partial\theta^2}L(\theta)\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log f(X_i;\theta)\right) = -nE\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta)\right) = nI(\theta).$$

So we can simplify the above theorem as

$$\text{Var}(T) \geq \frac{1}{-E\left(\frac{\partial^2 L(\theta)}{\partial\theta^2}\right)}.$$

The bound $\frac{1}{-E\left(\frac{\partial^2 L(\theta)}{\partial\theta^2}\right)}$ is known as the **Crammer-Rao lower bound (CRLB).**

Note: If an unbiased estimate attains this lower bound, then it is the (unique) minimum variance unbiased estimator (MVUE).

The question now is how to obtain the MVUE. Towards that, we can use the following theorem:

Theorem 1.2. *Under regularity conditions, when the true value of θ is θ_0, the MLE $\hat{\theta}$ is approximately normal with mean θ_0 and variance = CRLB at $\theta_0 = -\left.\frac{1}{E\left(\frac{\partial^2 L(\theta)}{\partial\theta^2}\right)}\right|_{\theta=\theta_0}$ for large n.*

The above result asserts that the MLE is asymptotically the MVUE when certain regularity conditions are satisfied. We now provide an example where it is the MVUE for any sample size.

Example 1.11 (Poisson revisited): The MLE for the Poisson distribution, as seen earlier, is $\hat{\lambda} = \bar{X}$and it is unbiased for λ. Now,

$$-E\left(\frac{\partial^2}{\partial\lambda^2}L(\lambda)\right) = -E\left(-\frac{\sum_{i=1}^{n}X_i}{\lambda^2}\right) = \frac{n}{\lambda}.$$

Hence the CRLB is $\frac{\lambda}{n}$, which is also the variance of $\bar{X}$. Hence, here the MLE is also the MVUE.

Chapter 2

Interval Estimation

2.1 Introduction

Interval estimation uses the sample data to calculate an interval of possible values of an unknown population parameter with certain degree of reliability.

The most prevalent forms of interval estimation are confidence intervals. In this chapter, we shall discuss estimation of confidence intervals. Before that, however, we need to introduce some distributions that we are going to use frequently.

2.2 Distributions of Functions of Normal Random Variable: χ^2, t and F

2.2.1 The χ^2 Distribution

If $X_1, \ldots, X_n \sim N(0,1)$ are independent, then $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$. In particular, the square of a standard normal random variable is χ_1^2. Note that χ_n^2 is a gamma random variable, see Example 1.2.7.

2.2.2 The t Distribution

If $Z \sim N(0,1)$ and $U \sim \chi_n^2$ are independent, then the distribution of $Z/\sqrt{U/n}$ is called the t distribution with n degrees of freedom.

This variable has the probability density function

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

which may be derived by the methods discussed in the previous section.

Note that the density of the t distribution is symmetric about zero, and approaches the standard normal density as the number of degrees of freedom converges to ∞.

2.2.3 The F Distribution

If U and V are two independent χ^2 random variables with degrees of freedom m and n respectively, then the distribution of $W = \frac{U/m}{V/n}$ is called the F distribution with m and n

degrees of freedom and is denoted by $F_{m,n}$. The probability density function of an $F_{m,n}$ random variable is

$$f(w) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} w^{\frac{m}{2}-1}\left(1+\frac{m}{n}w\right)^{-\frac{m+n}{2}},$$

where $w \geq 0$.

It should also be obvious that the square of a t_n random variable has an $F_{1,n}$ distribution, and that if $W \sim F_{m,n}$, then $\frac{1}{W} \sim F_{n,m}$.

Example 2.0 (The Mean and Variance of a Normal Sample): Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (i.i.d.) $N\left(\mu, \sigma^2\right)$ random variables; they are sometimes referred to as a sample from $N\left(\mu, \sigma^2\right)$. We shall now find the joint distribution of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$, which are called the sample mean and the sample variance, respectively.

Since $\bar{Y}$ is a linear combination of i.i.d. normal random variables, it is also normal with

$$E\left(\bar{Y}\right) = \frac{n\mu}{n} = \mu; \quad V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{V\left(\sum_{i=1}^{n} Y_i\right)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

It may be shown that $\bar{Y}$and S^2 are independent, and that$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Since $\bar{Y}$ and S^2 are independent, $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$,

$$\frac{\left(\bar{Y}-\mu\right)/\sqrt{\frac{\sigma^2}{n}}}{\sqrt{\left(n-1\right)S^2}/\sqrt{\left(n-1\right)\sigma^2}} \sim t_{n-1};$$

as the numerator is a $N(0,1)$ random variable independent of $(n-1)S^2/\sigma^2$, which has a χ^2_{n-1} distribution. Simplifying,

$$\frac{\sqrt{n}\left(\bar{Y}-\mu\right)}{S} \sim t_{n-1}.$$

This result will play a major role when we discuss confidence intervals and testing.

2.3 Confidence Intervals

Confidence intervals are a way to express the degree of uncertainty around an estimate of a parameter. A confidence interval for a parameter θ is a random interval that contains θ with some specified probability. When estimated repeatedly, a $100(1-\alpha)\%$ **confidence interval** (C.I.) will contain the true parameter about $100\left(1-\alpha\right)\%$ times.

Most often people are interested in 95% ($\alpha = 0.05$) and 99% ($\alpha = 0.01$) confidence intervals. The confidence intervals could be one-sided, i.e. of the form $(-\infty, c)$ or (c, ∞) where c is a suitable constant, or it could be both sided, i.e. of the form (c_1, c_2). The both sided confidence intervals are usually preferred because they are finite, but the one-sided confidence intervals are also frequently used.

Confidence intervals could be estimated using various methods. We discuss two: **exact methods** and **approximations based on large sample techniques using asymptotic properties**.

Let us now discuss an example where the confidence interval is exact.

Example 2.1: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$, where both parameters are unknown. Let us see how to construct a confidence interval for μ.

Recall that $\hat{\mu} = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $n\hat{\sigma}^2/\sigma^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}$ independently, so that $\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{s} \sim t_{n-1}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.

Let $t_{n-1;\alpha/2}$ denote the point beyond which the corresponding t-distribution has probability $\alpha/2$.

By definition,

$$P\left(-t_{n-1;\frac{\alpha}{2}} \leq \frac{\sqrt{n}\left(\bar{X}-\mu\right)}{s} \leq t_{n-1;\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(\bar{X} - \frac{s}{\sqrt{n}}t_{n-1;\alpha/2} \leq \mu \leq \bar{X} + \frac{s}{\sqrt{n}}t_{n-1;\alpha/2}\right) = 1-\alpha$$

So, a $100\left(1-\alpha\right)\%$ C.I. for μ is $\bar{X} \pm \frac{s}{\sqrt{n}}t_{n-1;\alpha/2}$. We shall discuss these results later in more detail.

Note: we keep saying "a" $100\left(1-\alpha\right)\%$ rather than "the" $100(1-\alpha)\%$. That is because the confidence intervals are not unique, for the same parameter and the same confidence coefficient; different confidence intervals can be obtained. For example, in **Example 2.1**, a different $100(1-\alpha)\%$ C.I. for μ is $\left(\bar{X} - \frac{s}{\sqrt{n}}t_{n-1;3\alpha/4},\ \bar{X} + \frac{s}{\sqrt{n}}t_{n-1;\alpha/4}\right)$

Example 2.1 (Continued): Computing confidence intervals numerically: Suppose we have the following sample of size 10 from a normal distribution:

$$5.93, 4.97, 5.65, 5.34, 4.10, 3.39, 5.16, 5.76, 3.66, 6.37$$

We would now construct 90% confidence intervals for the mean μ, assuming that the variance is unknown.

Case I: both sided C.I.

Here $\alpha = 0.1$, $n = 10$, $\bar{X} = 5.033$, and $s = 1.004$.

A 90% confidence interval is therefore $\bar{X} \pm \frac{s}{\sqrt{n}}t_{n-1;\alpha/2} = 5.033 \pm \frac{1.004}{\sqrt{10}}t_{9;0.05}$

Now, $t_{9;0.05} = 1.833$ from the tables. So, the confidence interval would be $(4.451,\ 5.615)$.

Case II: one sided C.I. with an upper confidence coefficient

Here the interval would look like $(-\infty, c)$. Now, $\frac{\sqrt{n}(\bar{X}-\mu)}{s} \sim t_{n-1}$ with $n = 10$.

So the value of c will be given by $c = \bar{X} + \frac{s}{\sqrt{n}}t_{n-1;\alpha} = 5.033 + \frac{1.004}{\sqrt{10}}t_{9;0.1}$.

Now, $t_{9;0.1} = 1.383$ from the tables, so $c = 5.472$: the 90% confidence interval is $(-\infty,\ 5.472)$.

In general, exact confidence intervals are not easy to obtain. We illustrate this point by the example below.

Example 2.2: Let X_1, X_2 and X_3 be three i.i.d. Poisson random variables with parameter $\lambda > 0$. Suppose we want to find a one sided 95% confidence interval of the form $(0, c)$ for λ

based on $\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$. Now, for independent Poisson, the sum is also Poisson with the parameter the sum of parameters, so $(X_1 + X_2 + X_3)$ is Poisson with parameter 3λ. So, letting $Y = (X_1 + X_2 + X_3)$, we need a constant c such that $P\left(\bar{X} < c\right) = P(Y < 3c) = 0.95$.

So, c is such that $\sum_{x=0}^{3c-1} \frac{e^{-3\lambda}(3\lambda)^x}{x!} = 0.95$. Problem is that such a c with exact sum 0.95 might not be found, and even if we decide to use a c for which the sum is approximately 0.95, its value still depends on λ, and hence without knowing λ we can not find c!

To counter this problem, we usually use approximate confidence intervals based on the central limit theorem whenever we can.

Example 2.2 (continued): $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately $N(\lambda,\ \lambda/n)$ for "reasonably large" n by applying the central limit theorem. So, $\frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}} \sim N(0,1)$ approximately.

Hence, approximately, $P\left(-z_{0.05} < \frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}} < \infty\right) = 0.95$.

Therefore, $P\left(-z_{0.05}\sqrt{\lambda/n} < \bar{X} - \lambda < \infty\right) = 0.95$ or $P\left(-\infty < \lambda < \bar{X} + z_{0.05}\sqrt{\lambda/n}\right) = 0.95$ where $z_{0.05} = 1.645$ is the 95th percentile of the standard normal distribution. From there we can find an approximate confidence interval $\left(0, \bar{X} + z_{0.05}\sqrt{\bar{X}/n}\right)$. Since $\lambda > 0$. Here $n = 3$, so the confidence interval is $\left(0,\ \bar{X} + 1.645\sqrt{\bar{X}/3}\right)$.

However, note that to apply the central limit theorem here we need to assume that 3 is a "large" number, which is not really very reasonable. But that is the best that we could do here.

Example 2.3 Confidence intervals for the mean of an i.i.d. normal sample: Consider a sample $X_1, X_2, \ldots, X_n$ of i.i.d. $N(\mu, \sigma^2)$. We now construct two-sided 95% confidence interval for μ for two cases: (a) σ^2 is known, (b) σ^2 is unknown.

(a) When σ^2 is known:

Here $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \sigma^2\right)$. Hence, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

So, using the symmetry of the normal distribution, we can write $P\left(\left|\bar{X} - \mu\right| < \frac{z_{0.025}\sigma}{\sqrt{n}}\right) = 0.95$, or $P\left(\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right) = 0.95$, where $z_{0.025} = 1.96$ is the 97.5th percentile of the standard normal distribution.

This gives us that for an i.i.d. normal sample with mean μ and variance σ^2, $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$, is an exact 95% both sided confidence interval for μ when σ is known.

Remark: We have constructed a symmetric 95% confidence interval for μ here. However, we can construct other 95% confidence intervals, for example $\left(\bar{X} - \frac{z_{0.01}\sigma}{\sqrt{n}},\ \bar{X} + \frac{z_{0.04}\sigma}{\sqrt{n}}\right) = \left(\bar{X} - 2.326\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.75\frac{\sigma}{\sqrt{n}}\right)$ is another 95% confidence interval for μ, as

$$P\left(\bar{X} - z_{0.01}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.04}\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

(b) When σ^2 is unknown:

As we have seen in (a), we can still write $P\left(\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right) = 0.95$, but the problem is that σ is unknown. However, as we have discussed earlier, we know that $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, so using the symmetry of the t distribution, we can write $P\left(\left|\bar{X} - \mu\right| < t_{n-1,0.025}\frac{s}{\sqrt{n}}\right) = 0.95$, or

$$P\left(\bar{X} - t_{n-1;0.025}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1;0.025}\frac{s}{\sqrt{n}}\right) = 0.95,$$

where $t_{n-1;0.025}$ is the 97.5th percentile of the t_{n-1} distribution.

Hence for an i.i.d. normal sample with mean μ and variance σ^2, $\bar{X} \pm t_{n-1,0.025}\frac{s}{\sqrt{n}}$, is an exact 95% both sided confidence interval for μ when σ is unknown.

Note:

1. All the above confidence intervals are exact.
2. For large n, a t_{n-1} distribution can be approximated by a standard normal distribution, and hence the 95% confidence interval will be approximately $\bar{X} \pm z_{0.025}\frac{s}{\sqrt{n}} = \bar{X} \pm 1.96\frac{s}{\sqrt{n}}$.

Example 2.1 (continued): Suppose we know somehow that $\sigma^2 = 1$. Then a 95% confidence interval for the mean μ is $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \bar{X} \pm 1.96\frac{1}{\sqrt{n}} = 5.033 \pm \frac{1.96}{\sqrt{10}} = (4.413, 5.653)$.

When σ^2 is unknown, a 95% confidence interval for the mean μ is $\bar{X} \pm t_{n-1,0.025}\frac{s}{\sqrt{n}} = 5.033 \pm t_{9,0.025}\frac{1.004}{\sqrt{10}} = 5.033 \pm 2.262 \times \frac{1.004}{\sqrt{10}} = (4.318,\ 5.748)$.

Example 2.4 (Confidence intervals for the mean of a large i.i.d. sample): Consider a sample $X_1, X_2, \ldots, X_n$ of i.i.d. observations with mean μ and variance σ^2, but not necessarily normal. Suppose we want to construct a 95% confidence interval for the mean μ for large n.

(a) σ^2 **is known.** Then, using the fact that by the ***Central Limit Theorem***, for large n, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ approximately, we can just proceed as part (a) in Example 2.2.3.

That tells us that for large n, $P\left(\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right) = 0.95$ approximately, and hence ***for a large i.i.d. sample with mean μ and variance σ^2, $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is an approximate 95% both sided confidence interval for μ when σ is known.***

(b) σ^2 **is unknown.** We can estimate σ^2 by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$; and obtain an approximate 95% confidence interval for μ for large n by writing

$$P\left(\bar{X} - z_{0.025}\frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{s}{\sqrt{n}}\right) \approx 0.95$$

and hence ***for a large i.i.d. sample with mean μ and variance σ^2, $\bar{X} \pm 1.96\frac{s}{\sqrt{n}}$ is an approximate 95% both sided confidence interval for μ when σ is unknown.***

We now summarise Examples 2.3 and 2.4 in the following result in a slightly general form:

Theorem 2.1. *Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a distribution with mean μ and variance σ^2,*

1. *If the X_i's are normal random variables, σ^2 known, then an exact $100(1-\alpha)\%$ two-sided confidence interval for μ will be given by $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.*

2. *When X_i's are normal random variables, σ^2 is unknown, then an exact $100(1-\alpha)\%$ confidence interval for μ will be given by $\bar{X} \pm t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$. For largen, it could be approximated by $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$.*

3. *When X_i's are not necessarily normal, σ^2 known, for large n an approximate $100(1-\alpha)\%$ confidence interval for μ will be given by $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.*

4. *When X_i's are not necessarily normal, σ^2 is unknown, then for large n an approximate $100(1-\alpha)\%$ confidence interval for μ will be given by $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$.*

Note: The above theorem only discusses both sided confidence intervals. Similar to what we have done for Example 2.2.1, we can also construct one sided confidence intervals. Some formulae are given at the end of the chapter.

The following result tells us how to construct the confidence intervals in more general situations.

Theorem 2.2. *(Confidence Intervals for Parameters related to μ) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a distribution with mean μ and variance σ^2. Suppose the parameter θ is such that $\theta = f(\mu)$, where f is a differentiable function. Suppose further that $f'(\mu)$ can not be 0.*

1. *If the X_i's are normal random variables, σ^2 known, then an approximate $100(1-\alpha)\%$ two-sided confidence interval for θ will be given by $f\left(\bar{X}\right) \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \left|f'\left(\bar{X}\right)\right|$.*

2. *When X_i's are normal random variables, σ^2 is unknown, then an approximate $100(1-\alpha)\%$ confidence interval for θ will be given by $f\left(\bar{X}\right) \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \left|f'\left(\bar{X}\right)\right|$.*

3. *When X_i's are not necessarily normal, σ^2known, for largen an approximate $100(1-\alpha)\%$ confidence interval for θ will be given by $f\left(\bar{X}\right) \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \left|f'\left(\bar{X}\right)\right|$.*

4. *When X_i's are not necessarily normal random variables, σ^2 is unknown, then for large n an approximate $100(1-\alpha)\%$ confidence interval for θ will be given by $f\left(\bar{X}\right) \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \left|f'\left(\bar{X}\right)\right|$.*

Proof of this result follows from the result that for large n, $f\left(\bar{X}\right) - f(\mu) \sim N\left(0, \sigma^2 \times \left(f'(\mu)\right)^2\right)$ approximately.

Let us see how we could use the above result.

Example 2.5: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with the following distribution:

$$P(X=x) = \begin{cases} \frac{\theta^2}{4} & \text{if } x=-1 \\ 1-\theta^2 & \text{if } x=0 \\ \frac{3\theta^2}{4} & \text{if } x=1 \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$.

Suppose we want to obtain a 90% confidence interval for θ. Notice that $\mu = E(X) = -1 \times \frac{\theta^2}{4} + 1 \times \frac{3\theta^2}{4} = \frac{\theta^2}{2}$. So, $\theta = \sqrt{2\mu}$.

Further, $\sigma^2 = V(X) = E\left(X^2\right) - \mu^2 = 1 \times \frac{\theta^2}{4} + 1 \times \frac{3\theta^2}{4} - \mu^2 = \theta^2 - \left(\frac{\theta^2}{2}\right)^2 = \theta^2 - \frac{\theta^4}{4}$, which is a function of θ and therefore unknown.

Therefore, we may apply part (b) of Theorem 2.2 to obtain the approximate confidence interval $f\left(\bar{X}\right) \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \left|f'\left(\bar{X}\right)\right|$ for θ for large n, where $\alpha = 0.1$, $z_{\frac{\alpha}{2}} = 1.645$; $f(x) = \sqrt{2x}$.

Hence, $f'(x) = \sqrt{2} \times \frac{1}{2} x^{-\frac{1}{2}} = \frac{1}{\sqrt{2x}}$ giving the 90% confidence interval for θ to be

$$\sqrt{2\bar{X}} \pm 1.645 \times \frac{s}{\sqrt{2n\bar{X}}}.$$

Note that a 95% confidence interval for θ is

$$\sqrt{2\bar{X}} \pm 1.96 \times \frac{s}{\sqrt{2n\bar{X}}}.$$

Here, clearly, $f'\left(\bar{X}\right)$ should not be 0.

We are now ready for another result.

Theorem 2.3. *(Confidence Interval based on the MLE) Let $X_1, X_2, \ldots, X_n$ be i.i.d. with probability function $f(x;\theta)$. If $f''(x;\theta)$ exists and is continuous, then for large n, an approximate $100(1-\alpha)\%$ confidence interval for θwould be given by $\hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{C\hat{R}LB}$ where $\hat{\theta}$ is an MLE for θ and $C\hat{R}LB$ is an appropriate estimate of the CRLB $1/E\left(-\frac{\partial^2 L(\theta)}{\partial \theta^2}\right)$.*

Example 2.6: Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli(p). Then $\hat{p} = \bar{X}$. The loglikelihood is

$$L(p) = \log(p) \sum_{i=1}^{n} X_i + \log(1-p)\left(n - \sum_{i=1}^{n} X_i\right).$$

So, $\frac{d^2 L(p)}{dp^2} = -\frac{\sum_{i=1}^n X_i}{p^2} - \frac{n - \sum_{i=1}^n X_i}{(1-p)^2}$ giving

$$E\left(-\frac{d^2 L(p)}{dp^2}\right) = E\left(\frac{\sum_{i=1}^n X_i}{p^2} + \frac{n - \sum_{i=1}^n X_i}{(1-p)^2}\right) = \frac{E\left(\sum_{i=1}^n X_i\right)}{p^2} + \frac{n - E\left(\sum_{i=1}^n X_i\right)}{(1-p)^2}$$

$$= \frac{\mathrm{np}}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.$$

The above tells us that $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ approximately for large n.

Hence, an approximate $100(1-\alpha)\%$ confidence interval for p would be given by

$$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n\hat{I}(p)}},$$

where for $n\hat{I}(p)$ we may substitute $\frac{n}{\hat{p}(1-\hat{p})} = \frac{n}{\bar{X}(1-\bar{X})}$. This means the above approximate $100(1-\alpha)\%$ confidence interval can be re-written as

$$\bar{X} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

(We could have also deduced the above result by the central limit theorem as $\hat{p} = \bar{X}$, but now we have used the asymptotic distribution of the MLE.)

Example 2.7: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. exponential random variables with parameter $\lambda > 0$. We have seen in Example 1.1.7 that $\hat{\lambda} = n/\sum_{i=1}^{n} X_i = 1/\bar{X}$ is the MLE, where $\bar{X}$ is the sample mean. Now, $\frac{d^2L}{d\lambda^2} = -\frac{n}{\lambda^2}$.

Since $E\left(-\frac{d^2L}{d\lambda^2}\right) = E\left(\frac{n}{\lambda^2}\right) = \frac{n}{\lambda^2}$, the CRLB is $\frac{\lambda^2}{n}$, and hence for large n the distribution of $\hat{\lambda}$ is approximately $N\left(\lambda, \frac{\lambda^2}{n}\right)$.

So, an approximate $100(1-\alpha)\%$ confidence interval for λ would be given by

$$\frac{1}{\bar{X}} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\lambda}^2}{n}},$$

where $\hat{\lambda} = \frac{1}{\bar{X}}$.

Hence, the above confidence interval can be re-written as

$$\frac{1}{\bar{X}} \pm z_{\frac{\alpha}{2}} \frac{1}{\bar{X}\sqrt{n}}.$$

Chapter 3

Testing of Hypotheses

3.1 Terminology

Definition 3.1. *A statistical hypothesis test is a method of making statistical decisions using experimental data.*

3.1.1 Null Hypothesis, Alternative Hypothesis and the Test Statistic

In a hypothesis testing procedure the main task is to choose between two competing hypotheses about the distribution from which the experimental data are generated.

In testing of hypotheses:

(a) Usually people put more faith on the hypothesis that they think is more likely (or sometimes, easier to describe). It is called the *null hypothesis*, usually denoted by H_0. The alternative is called the *alternative hypothesis*, and is usually denoted by H_A or H_1.

(b) The testing of hypothesis procedure is performed given an i.i.d. sample $X_1, \ldots, X_n$.

(c) The decision of the testing of hypothesis is either to accept or to reject the null hypothesis based on the value of *a statistic*, say $T(X_1, \ldots, X_n)$ or more simply $T(\ X)$, computed on the basis of the sample, which is called the *test statistic*.

(d) Whenever the *null hypothesis* is rejected, the alternative hypothesis is accepted.

Points to note: The plausibility of the alternative hypothesis is never checked. It is accepted whenever it is decided that the null hypothesis is unacceptable. Hence the alternative hypothesis should be chosen with care.

Example 3.1.1: A test is performed to check whether a given coin is fair or biased towards heads based on 10 tosses of the coin. Suppose the observed results of the tosses are:

$$HHTHHHHTHH.$$

Here, we choose the null hypothesis to be that the *coin is fair*: $H_0 : p = \frac{1}{2}$; where $p = P(\text{Head})$.

As we are interested only in knowing whether the coin is fair or biased towards heads, we choose the alternative hypothesis $H_A : p > \frac{1}{2}$.

The *test statistic* in this case would be the *number of heads obtained in the 10 throws.* We can see that intuitively, but we would also discuss a mathematical justification later.

Let us define X_j as the indicator of the j-th toss producing a head, (i.e. $X_j = 1$ for a head in the j-th toss, and 0 if it is a tail,) $1 \leq j \leq 10$. Then, our test statistic, the number of heads, is $X = \sum_{i=1}^{10} X_i$.

3.1.2 Simple and Composite Hypothesis

Definition 3.2. *When a hypothesis considers one completely specified distribution, it is a simple hypothesis; it is a composite hypothesis otherwise.*

Example 3.1.1 (continued): The null hypothesis $H_0 : p = \frac{1}{2}$ specifies only one possibility, under which each X_j is Bernoulli$\left(\frac{1}{2}\right)$, which is a completely specified distribution. Hence, the null hypothesis in this example is a simple hypothesis.

On the other hand, under the alternative X_j is Bernoulli(p), where p can be any real number between $\frac{1}{2}$ and 1; so the alternative hypothesis is a composite hypothesis.

3.1.3 Acceptance and Rejection Regions

The null hypothesis is rejected when the value of $T(X_1, \ldots, X_n) \in R$, where R is such a set that $T(X_1, \ldots, X_n)$ is not very likely to take values is when H_0 is correct. The set R is therefore known as the *rejection region* or the *critical region.* More formally,

Definition 3.3. *The rejection region or the critical region, R, denotes the set of values of $T(X_1, \ldots, X_n)$ for which the null hypothesis H_0 is rejected.*

The complement of the set R is A, the *acceptance region:* the set of values of $T(X_1, \ldots, X_n)$ for which the null hypothesis is **not rejected**.

The acceptance and rejection regions depend on the alternative hypothesis.

Example 3.1.1 (continued): In this example, clearly the alternative is more likely when there are more heads. Hence, the *rejection region* would be of the form $X > c$, for a suitable number c. The *acceptance region* is then $X \leq c$.

3.1.4 Type I and Type II Errors, Level of Significance and Power of a Test

Two types of errors can occur when a hypothesis is tested. The following table summarises the two cases:

	H_0 is true	H_0 Is False
Decision is taken to reject H_0	Type I error	
Decision is taken not to reject H_0		Type II error

Definition 3.4. *The probability of the type I error for a test is called the level of the test or the significance level of the test.*

Notations: The level of the test is usually denoted by α. When H_0 is composite, α is the supremum of the probabilities of type I error under various possible distributions under H_0.

The probability of the type II error is denoted by β.

Definition 3.5. *The power of a test is defined as $1-\beta$, the probability of rejecting the null when it is false.*

Note that when the null is false and the alternative hypothesis is composite, then the value of β and the power of the test would depend on which particular distribution under the alternative hypothesis is true.

Example 3.1.1 (continued): Note that the X_i's are independent Bernoulli(p) random variables, so that $X \sim \text{Binomial}(10, p)$. Now, under the null hypothesis H_0, (i.e. when H_0 is true,) $p = \frac{1}{2}$, so $X \sim \text{Binomial}(10, \frac{1}{2})$.

Hence, $\alpha = P(X > c)$ when $X \sim \text{Binomial}\left(10, \frac{1}{2}\right)$, i.e.

$$\alpha = \sum_{j=c+1}^{10} \binom{10}{j} (0.5)^j (0.5)^{10-j} = (0.5)^{10} \sum_{j=c+1}^{10} \binom{10}{j}.$$

Now suppose H_0 is false, and $p = 0.7$. Then, $X \sim \text{Binomial}(10, 0.7)$, implying $\beta = P(X \leq c) = \sum_{j=0}^{c} \binom{10}{j} (0.7)^j (0.3)^{10-j}$, and so the power is

$$1 - \beta = \sum_{j=c+1}^{10} \binom{10}{j} (0.7)^j (0.3)^{10-j}$$

3.1.5 The p-value of a Test

Definition 3.6. *The p-value is the probability of obtaining a result at least as extreme as the one that was actually observed when the null hypothesis H_0 is true.*

Note: When asked to test at a significance level α_0, we reject the null hypothesis if and only if the p-value of the test statistic is less than α_0. Commonly tests are performed at significance levels 0.1, 0.05, 0.01 or 0.001.

Example 3.1.1 (continued): Here $x = 8$. Hence the p-value is $(0.5)^{10} \sum_{j=8}^{10} \binom{10}{j} = 0.055$.

General benchmarking of p-values

- < 0.1 (or 10%) indicates *marginal* evidence against H_0
- < 0.05 (or 5%) indicates *reasonable* evidence against H_0
- < 0.01 (or 1%) indicates *strong* evidence against H_0
- < 0.001 (or 0.1%) indicates *very strong* evidence against H_0

3.2 Most Powerful and Uniformly Most Powerful tests

Ideally a test should have $\alpha = \beta = 0$, i.e. there should not be any error. But that is usually not possible. Usually one of them increases when the other decreases.

It is a common practice to set an upper bound for α and then try to maximise power.

Definition 3.7. *The test with maximum power for a fixed level α is called the most powerful (MP) test at level α.*

Definition 3.8. *When H_A is composite, if a test has maximum power against every alternative, it is called the uniformly most powerful (UMP) test at level α.*

UMP tests do not exist always; we will see some examples later where they do.

Example 3.2.1: Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, 1)$ with unknown mean μ. Suppose one knows that μ is either 0 or 1, but the exact value is unknown. Suppose one person believes that 0 is the more likely value and sets $H_0 : \mu = 0$, $H_A : \mu = 1$. Both are clearly simple hypotheses.

A good estimate of μ should be used as the test statistic, and clearly $\bar{X}$ is one candidate. Since μ is higher under H_A, one should reject the null when $\bar{X}$ is high, i.e. the rejection region is of the form $\bar{X} > c$. It will be shown later that this test is the *most powerful (MP)* test.

Type I error would be committed if μ is actually 0 and we reject H_0. The level of the test α is then the probability of rejecting H_0 when $\mu = 0$.

Now, if $\mu = 0$, then $X_i \sim N(0, 1)$ for all j, and hence $\bar{X} \sim N\left(0, \frac{1}{n}\right)$.

Hence, as $\bar{X} \sim N\left(0, \frac{1}{n}\right)$ it implies that $\sqrt{n}\bar{X} \sim N(0, 1)$,

$$\alpha = P\left(\bar{X} > c\right) = P\left(\sqrt{n}\bar{X} > \sqrt{n}c\right) = 1 - \Phi\left(\sqrt{n}c\right),$$

where Φ is the distribution function of the standard normal distribution.

So, $\sqrt{n}c = z_\alpha$, giving the rejection region to be $\bar{X} > \frac{z_\alpha}{\sqrt{n}}$.

Note: For simple null and alternative, the level is also called the *size of the test*, as is the case here.

The type II error is committed if one do not reject $\mu = 0$ when actually it is not true. The probability of that is denoted by β, and power of the test is $1 - \beta$. **The most powerful test for a fixed α would be the one to maximize $1 - \beta$, or to minimize β.**

Now, in the present set up, if $\mu \neq 0$ then $\mu = 1$, and hence $\bar{X} \sim N\left(1, \frac{1}{n}\right)$.

Hence, $\beta = P\left(\bar{X} \leq c\right)$ and as $\bar{X} \sim N\left(1, \frac{1}{n}\right)$ implies that $\sqrt{n}\left(\bar{X} - 1\right) \sim N(0, 1)$, the *power of the test* is

$$1 - \beta = P\left(\bar{X} > c\right) = P\left(\bar{X} - 1 < c - 1\right) = P\left(\sqrt{n}\left(\bar{X} - 1\right) > \sqrt{n}\left(c - 1\right)\right)$$

$$= 1 - \Phi\left(\sqrt{n}\left(c - 1\right)\right) = 1 - \Phi\left(\sqrt{n}\left(\frac{z_\alpha}{\sqrt{n}} - 1\right)\right)$$

p-value: Suppose that for a particular sample we observe that the value of the sample mean is c_0. The p-value of the above test for this value of the sample mean is given by

p-value = Probability of getting a sample mean at least as extreme when H_0 is true $= P\left(\bar{X} > c_0\right) = P\left(\sqrt{n}\bar{X} > \sqrt{n}c_0\right) = 1 - \Phi\left(\sqrt{n}c_0\right)$ as under H_0, $\sqrt{n}\bar{X} \sim N(0,1)$.

Example 3.2.2 Consider $X_1, X_2, \ldots, X_n$: i.i.d. $N(\mu, 1)$ with unknown mean μ. Suppose one knows that $\mu \geq 0$, but the exact value is unknown. Let us set $H_0 : \mu = 0$, $H_A : \mu > 1$. The alternative is now composite.

The type I error would be committed if μ is actually 0 and we reject H_0. The *level* α is now the supremum of the sizes of all the tests $H_0 : \mu = 0$ versus $H_A : \mu = \mu_1$, where μ_1 is a fixed positive number.

The type II error is committed if one does not reject $\mu = 0$ when actually $\mu = \mu_1$ for some $\mu_1 > 0$. The *most powerful test* for a fixed α and a fixed μ_1 would be the one to maximize corresponding P_{μ_1} (Rejecting H_0).

If there is a level α test which would maximize power uniformly for all $\mu_1 > 0$, then it would be a UMP level α test. We will see later that the test described in previous example would be a UMP level α test.

3.3 Likelihood Ratio Tests

The likelihood ratio tests are performed on the basis of the ratio of the likelihood under the distributions of the data under the null hypothesis and the alternative hypothesis.

More specifically,

Definition 3.9. *The likelihood ratio is defined as*

$$\Lambda(X_1, \ldots, X_n) = \frac{\ell(H_0|X_1, X_2, \ldots, X_n)}{\ell(H_A|X_1, X_2, \ldots, X_n)}$$

where the likelihood function under the null hypothesis is

$$\ell(H_0|X_1, X_2, \ldots, X_n),$$

and

$$\ell(H_A|X_1, X_2, \ldots, X_n)$$

is the likelihood under the alternative hypothesis.

The basic idea is that the likelihood would be higher for the true hypothesis, and hence the likelihood ratio would be large if the null hypothesis H_0 is true, and small if H_A is false.

Example 3.3.1: Consider the model of Example 3.1.1:

Data: Results of 10 tosses of a coin: *HHTHHHHTHH.*

Hypotheses: The **null hypothesis** is $H_0 : p = 0.5$; let the **alternative hypothesis** be $H_A : p = 0.7$.

The likelihood of a specific sequence with x heads in general is given by $p^x (1-p)^{10-x}$.
So, $\ell(H_0|x) = 0.5^x (1-0.5)^{10-x} = 0.5^{10}$ and $\ell(H_A|x) = 0.7^x (1-0.7)^{10-x} = 0.7^x 0.3^{10-x}$.
Likelihood ratio: $\Lambda(x) = \frac{\ell(H_0|x)}{\ell(H_A|x)} = \frac{0.5^{10}}{0.7^x 0.3^{10-x}} = \left(\frac{0.5}{0.3}\right)^{10} \left(\frac{0.3}{0.7}\right)^x = 1.67^{10} 0.43^x$.
Rejection Criteria: Clearly, the likelihood ratio is small if x is high, and hence the *rejection region* is $x > c$ for some suitable constant c. This justifies our choice of this test in Example 3.1.1.

Hence, if a test rejects H_0 when $\frac{\ell(H_0|X_1,\ldots,X_n)}{\ell(H_A|X_1,\ldots,X_n)} < c$ for some $c > 0$, the test is then called *a likelihood ratio test.*

The likelihood ratio tests are optimal in the following sense:

Theorem 3.1. *Neyman-Pearson Lemma If a likelihood ratio test has significance level α, then any other test which has significance level $\leq \alpha$ has power less or equal to that test.*

Remark: The likelihood ratio test with level α is the most powerful (MP) test at level α for H_0 versus H_A.

Example 3.3.2 (Continuing with 3.2.1)
Data: $X_1, X_2, \ldots, X_n$: i.i.d. $N(\mu, 1)$ with unknown mean μ.
Hypotheses: $H_0 : \mu = 0$, $H_A : \mu = 1$.
By the *Neyman-Pearson lemma*, among all tests with significance level α the test that rejects for the small values of the likelihood ratio has the highest power.
Now, the likelihood ratio is

$$\Lambda(X_1,\ldots,X_n) = \frac{\ell(H_0|X_1,\ldots,X_n)}{\ell(H_A|X_1,\ldots,X_n)} = \frac{f(X_1,\ldots,X_n|\mu=0)}{f(X_1,\ldots,X_n|\mu=1)} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}X_i^2\right)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i-1)^2\right)}$$

$$= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i-1)^2\right)}$$

$$= \exp\left[\frac{1}{2}\left(\sum_{i=1}^n (X_i-1)^2 - \sum_{i=1}^n X_i^2\right)\right] = \exp\left(\frac{n}{2} - \sum_{i=1}^n X_i\right)$$

The likelihood ratio is small if $\sum_{i=1}^n X_i - \frac{n}{2}$ is large, or equivalently if $\bar{X}$ is large. So the *rejection region* is of the type $\bar{X} > c$.
We need to choose c so as the prescribed level is attained, i.e., $P\left(\bar{X} > c\right) = \alpha$ when H_0 is true, i.e., the true mean is 0.
As we have seen in example 3.2.1, this gives $\sqrt{n}c = z_\alpha$, or $c = z_\alpha/\sqrt{n}$, giving
Rejection Criteria: Reject H_0 at level α if $\bar{X} > z_\alpha/\sqrt{n}$.

Note: As the test with the rejection region $\bar{X} > c$ is the likelihood ratio test, it is the MP test.

Example 3.3.3 (Continuing with 3.2.2): Now consider testing $H_0 : \mu = 0$ against $H_A : \mu = \mu_A$; $\mu_A > 0$.

In this case, the likelihood ratio is

$$\Lambda(X_1,\ldots,X_n) = \frac{\ell(H_0|X_1,\ldots,X_n)}{\ell(H_A|X_1,\ldots,X_n)} = \frac{f(X_1,\ldots,X_n|\mu=0)}{f(X_1,\ldots,X_n|\mu=\mu_A)}$$

$$= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}X_i^2\right)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i-\mu_A)^2\right)}$$

$$= \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i-\mu_A)^2\right)}$$

$$= \exp\left[\frac{1}{2}\left(\sum_{i=1}^n (X_i-\mu_A)^2 - \sum_{i=1}^n X_i^2\right)\right] = \exp\left(\frac{n\mu_A^2}{2} - \mu_A\sum_{i=1}^n X_i\right)$$

The likelihood ratio is small if $\mu_A\sum_{i=1}^n X_i - \frac{n\mu_A^2}{2}$ is large, or equivalently if $\bar{X}$ is large.

Hence, the *rejection region* is again of the type $\bar{X} > c$. We can now proceed exactly in the same manner as the previous part and again get $c = z_\alpha/\sqrt{n}$.

Example 3.3.4: Testing H_0: $\mu = 0$ against $H_\mathbf{A}$: $\mu > 0$ in the Previous Example

In Example 3.3.3, we saw that for any fixed $\mu_A > 0$, the most powerful test rejects H_0 for $\bar{X} > c$, where c depends on n but not on μ_A. Hence this test is most powerful at level α for every $\mu_A > 0$, and hence is *UMP at level* α. Therefore, the UMP test at level α of $H_0 : \mu = 0$ against $H_A : \mu > 0$has rejection region given by $\bar{X} > z_\alpha/\sqrt{n}$.

Notes:

1. Proceeding as we did in Examples 3.3.3 and 3.3.4, we can show that the UMP level α test for testing $H_0 : \mu = \mu_0$ against $H_A : \mu > \mu_0$ is also of the form $\bar{X} > c$. The value of c would depend on μ_0. See the next two examples.

2. An alternative hypothesis as H_A above is referred to as a one sided alternative. An obvious example of a two-sided alternative is $H_A^{'} : \mu \neq \mu_0$.

Example 3.3.5: Testing H_0: $\mu = \mu_0$against H_A: $\mu > \mu_0$ where variance is 1.

Data: $X_1, X_2, \ldots, X_n$ i.i.d. $N(\mu, 1)$ with unknown mean μ.

Hypotheses: $H_0 : \mu = \mu_0$; $H_A : \mu > \mu_0$.

Rejection Criteria: The rejection region of the UMP test at level α is given by $\bar{X} > \mu_0 + z_\alpha/\sqrt{n}$. This result is left as an exercise.

Example 3.3.6: Testing H_0: $\mu = \mu_0$ against H_A: $\mu > \mu_0$ where variance is σ^2, known

Data: $X_1, X_2, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ with unknown mean μ.

Hypotheses: $H_0 : \mu = \mu_0$; $H_A : \mu > \mu_0$.

Rejection Criteria: The rejection region of the UMP test at level α is given by $\bar{X} > \mu_0 + \sigma z_\alpha/\sqrt{n}$. This result will also be left as an exercise.

3.4 Likelihood Ratio Tests for Composite Null Hypotheses

The likelihood ratio tests, as we have seen before, are optimal for simple null versus simple or one-sided alternatives. We now develop a generalisation for situations where the hypotheses are more complex. Such tests generally do not satisfy any optimality criteria, but they usually perform reasonably well.

Assume that the null hypothesis is now composite, and looks like $H_0 : \theta \in \Omega_0$, here Ω_0 is possibly a multi-element subset of Ω, the space of all possible values of the parameter θ. (In most of the examples we look at, $\Omega = \mathbb{R}$.)

Let us consider $H_A : \theta \in \Omega_0^c$, where Ω_0^c is the complement of the set Ω_0. To simplify notations, let us write $\Omega_1 = \Omega_0^c$.

For composite hypotheses, the likelihoods are evaluated for θ that maximise the likelihood under the corresponding hypothesis. In this case we define the *likelihood ratio* as

$$\Lambda^* (X_1, \ldots, X_n) = \frac{\max_{\theta \in \Omega_0} \ell (\theta | X_1, \ldots, X_n)}{\max_{\theta \in \Omega_1} \ell (\theta | X_1, \ldots, X_n)}$$

The corresponding test rejects H_0 for the small values of Λ^*.

For certain technical reasons, it is easier to deal with

$$\Lambda (X_1, \ldots, X_n) = \frac{\max_{\theta \in \Omega_0} \ell (\theta | X_1, \ldots, X_n)}{\max_{\theta \in \Omega} \ell (\theta | X_1, \ldots, X_n)}.$$

Note that $\Lambda = \min (\Lambda^*, 1)$, so small values of Λ correspond to small values of Λ^*. Hence, we reject H_0 for small values of Λ. Also note that the maximums are replaced by the supremum when the maximum for a set does not exist.

Example 3.4.1: Consider again $X_1, X_2 \ldots, X_n$, i.i.d. $N(\mu, 1)$ with unknown μ. Let $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$, where μ_0 is a known number. Here the parameter is μ, $\Omega_0 = \{\mu_0\}$, a single element set, and $\Omega = \mathbb{R}$.

So $\Lambda = \frac{\ell(\mu_0)}{\max_{\mu \in \mathbb{R}} \ell(\mu)}$; whose numerator is the likelihood of the sample under null:

$$\frac{1}{\left(\sqrt{2\pi}\right)^n} \exp \left(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu_0)^2 \right)$$

For the denominator, we need the maximum possible value of the likelihood. By definition of the maximum likelihood estimator (MLE), the likelihood is maximised when μ is equal to the MLE $\bar{X}$. The denominator is therefore

$$\frac{1}{\left(\sqrt{2\pi}\right)^n} \exp \left(-\frac{1}{2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \right)$$

Hence,

$$\Lambda = \exp \left[-\frac{1}{2} \left(\sum_{i=1}^{n} (X_i - \mu_0)^2 - \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \right) \right],$$

and we reject the null for small values of Λ.

This is equivalent to rejecting for large values of

$$-2\log(\Lambda) = \sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = n\left(\bar{X}-\mu_0\right)^2$$

Now, we know that $\sqrt{n}\left(\bar{X}-\mu_0\right) \sim N(0,1)$ when $H_0 : \mu = \mu_0$ is true, but then $n\left(\bar{X}-\mu_0\right)^2 \sim \chi^2_1$. Hence for any significance level α, the test rejects when $n\left(\bar{X}-\mu_0\right)^2 > \chi^2_{1,\alpha}$.

We could alternatively use the fact that $n\left(\bar{X}-\mu_0\right)^2 > c^2$ for any nonnegative constant c if and only if $\left|\bar{X}-\mu_0\right| > c/\sqrt{n}$. As $\sqrt{n}\left(\bar{X}-\mu_0\right) \sim N(0,1)$ under H_0, choosing $c = z_{\frac{\alpha}{2}}$ gives us an alternative expression for the rejection region of the level α test: $\left|\bar{X}-\mu_0\right| > z_{\frac{\alpha}{2}}/\sqrt{n}$.

We now state a theorem of which above derivation is a particular case.

Theorem 3.2. *Under smoothness conditions on the probability functions involved, the null distribution of* $-2\log(\Lambda)$ *is approximately* χ^2 *with degrees of freedom equal to dimension*(Ω) $-$ *dimension*(Ω_0).

The "dimension" is usually the number of unknown (or "free") parameters. For Ω in Example 3.4.1, the only unknown parameter was μ, whereas Ω_0 did not have any. Hence, $-2\log(\Lambda) \sim \chi^2_1$.

Note: The above theorem only gives an approximate result; the example gave an exact result.

Example 3.4.2: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$, where σ^2 is a known positive constant, and let $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu\_0$. Proceeding exactly as Example 3.4.1, we can set $\Lambda = \frac{\ell(\mu_0)}{\max_{\mu\in\mathbb{R}}\ell(\mu)}$; whose numerator would now be

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2,$$

and the denominator

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$$

Hence,

$$\Lambda = \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right)\right],$$

and we reject H_0 for small values of Λ.

This is equivalent to rejecting for large values of

$$-2\sigma^2\log(\Lambda) = \sum_{i=1}^{n}(X_i-\mu_0)^2 - \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = n\left(\bar{X}-\mu_0\right)^2$$

Now, we know that $\sqrt{n}\left(\bar{X}-\mu_0\right)/\sigma \sim N(0,1)$ when $H_0 : \mu = \mu_0$ is true, but then $n\left(\bar{X}-\mu_0\right)^2/\sigma^2 \sim \chi^2_1$.

Hence for any significance level α, the test rejects when $\frac{n\left(\bar{X}-\mu_0\right)^2}{\sigma^2} > \chi^2_{1,\alpha}$.

We could alternatively use the fact that $n\left(\bar{X}-\mu_0\right)^2/\sigma^2 > c^2$ for any nonnegative constant c if and only if $\left|\bar{X}-\mu_0\right| > c\sigma/\sqrt{n}$. As $\sqrt{n}\left(\bar{X}-\mu_0\right) \sim N(0,\sigma^2)$ under H_0, choosing $c = z_{\frac{\alpha}{2}}$ gives us an alternative expression for the rejection region of the level α test: $\left|\bar{X}-\mu_0\right| > \sigma z_{\frac{\alpha}{2}}/\sqrt{n}$.

3.5 Duality of Confidence Intervals and Hypothesis Tests

Tests and confidence intervals are interlinked with each other, as we see in the next example.

Example 3.5.1: Consider $X_1, X_2, \ldots, X_n$, i.i.d. $N(\mu,1)$, and let $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$. Consider the test at a specific level α that rejects for $\left|\bar{X}-\mu_0\right| > z_{\frac{\alpha}{2}}/\sqrt{n}$, as in Example 3.4.1. Therefore H_0 is accepted when $\left|\bar{X}-\mu_0\right| \leq z_{\frac{\alpha}{2}}/\sqrt{n}$, or $-z_{\frac{\alpha}{2}} \leq \sqrt{n}\left(\bar{X}-\mu_0\right) \leq z_{\frac{\alpha}{2}}$, or $\bar{X} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$.

Now, $\left(\bar{X} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\right)$ is a both sided $100(1-\alpha)\%$ confidence interval for μ. We should have taken $\left[\bar{X} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\right]$, which is the more accurate form. However, we can ignore the boundary points as they have probability zero. Therefore the test accepts the null if and only if μ_0 is in the above confidence interval. This tells us that the $100(1-\alpha)\%$ confidence interval for μ and the level α acceptance region for the above test are the same.

The following theorem summarise this duality:

Theorem 3.3. *If the acceptance region of a test with null hypothesis $H_0 : \theta = \theta_0$ at level α is the set $A(\theta_0) = \{(X_1, \ldots, X_n) : \theta_0 \in C(X_1, \ldots, X_n)\}$, then the set $C(X_1, \ldots, X_n) = \{\theta : (X_1, \ldots, X_n) \in A(\theta_0)\}$ is a $100(1-\alpha)\%$ confidence region for θ, and vice-versa.*

Example 3.5.2: We can see in Example 3.5.1 that the acceptance region for $H_0 : \mu = \mu_0$ at level α against $H_A : \mu \neq \mu_0$ is

$$A(\mu_0) = \left\{(X_1, \ldots, X_n) : -z_{\frac{\alpha}{2}} < \sqrt{n}\left(\bar{X}-\mu_0\right) < z_{\frac{\alpha}{2}}\right\}.$$

So here the set

$$C(X_1, \ldots, X_n) = \left\{\mu : -z_{\frac{\alpha}{2}} < \sqrt{n}\left(\bar{X}-\mu\right) < z_{\frac{\alpha}{2}}\right\} = \bar{X} \pm \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$$

will be a $100(1-\alpha)\%$ confidence interval for μ by Theorem 3.5.1. On the other hand, if we were given the confidence interval as above, then we can deduce the acceptance region.

Example 3.5.3: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu,1)$. Then if a $100(1-\alpha)\%$ confidence interval is chosen as $C(X_1, \ldots, X_n) = \left(\bar{X} - \frac{z_\alpha}{\sqrt{n}}, \infty\right)$, then the corresponding acceptance region at level α for the null hypothesis $H_0 : \mu = \mu_0$ will be given by

$$A(\mu_0) = \left\{(X_1, \ldots, X_n) : \mu_0 \in \left(\bar{X} - \frac{z_\alpha}{\sqrt{n}}, \infty\right)\right\}$$

$$= \left\{ (X_1, \ldots, X_n) : \mu_0 > \bar{X} - \frac{z_\alpha}{\sqrt{n}} \right\} = \left\{ (X_1, \ldots, X_n) : \bar{X} < \mu_0 + \frac{z_\alpha}{\sqrt{n}} \right\}.$$

So, this test rejects the null hypothesis for $\bar{X} > \mu_0 + \frac{z_\alpha}{\sqrt{n}}$.

Clearly, this test is similar to what we have obtained earlier in Examples 3.3.2, 3.3.3 and 3.3.4 with 0 replaced by μ_0; here the alternative has to be either of the type $H_A : \mu = \mu_A$; for some $\mu_A > \mu_0$, or more generally, $H_A : \mu > \mu_0$.

Note:

1. When we are testing the null hypothesis $H_0 : \mu = \mu_0$, and we have a confidence interval $C(X_1, \ldots, X_n)$ for μ, then we accept the null hypothesis if and only if $\mu_0 \in C(X_1, \ldots, X_n)$. For example, in the above example, $C(X_1, \ldots, X_n) = \left(\bar{X} - \frac{z_\alpha}{\sqrt{n}}, \infty\right)$, and hence

$$A(\mu_0) = \{(X_1, \ldots, X_n) : \ \mu_0 \in C(X_1, \ldots, X_n)\}$$

$$= \left\{ (X_1, \ldots, X_n) : \mu_0 > \bar{X} - \frac{z_\alpha}{\sqrt{n}} \right\} = \left\{ (X_1, \ldots, X_n) : \bar{X} < \mu_0 + \frac{z_\alpha}{\sqrt{n}} \right\}.$$

2. The form of the confidence interval that we convert the acceptance region to also provides us with the form of the rejection region.

Example 3.5.4: Consider a sample $X_1, X_2, \ldots, X_n$, i.i.d. $N(\mu, \sigma^2)$ where σ^2 is known, and let $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$. Consider the test at a specific level α that rejects for $|\bar{X} - \mu_0| > \sigma z_{\frac{\alpha}{2}}/\sqrt{n}$, as in Example 3.4.2. There H_0 is accepted when $|\bar{X} - \mu_0| \leq \sigma z_{\frac{\alpha}{2}}/\sqrt{n}$, or $-\sigma z_{\frac{\alpha}{2}} \leq \sqrt{n}(\bar{X} - \mu_0) \leq \sigma z_{\frac{\alpha}{2}}$, or $\bar{X} - \frac{\sigma z_{\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \frac{\sigma z_{\frac{\alpha}{2}}}{\sqrt{n}}$. Hence a both sided $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - \frac{\sigma z_{\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X} + \frac{\sigma z_{\frac{\alpha}{2}}}{\sqrt{n}} \right).$$

3.6 Normal Samples, Unknown Variance: t-test for Mean

Suppose we have $X_1, X_2, \ldots, X_n$, i.i.d. $N(\mu, \sigma^2)$, and we want to test $H_0 : \mu = \mu_0$ against one of the following alternatives:

$$H_{A_1} : \mu \neq \mu_0$$

$$H_{A_2} : \mu > \mu_0$$

$$H_{A_3} : \mu < \mu_0$$

where μ_0 is some known constant, σ^2 unknown. The tests we discussed so far were based on the z-statistics (i.e. the standard normal distribution), and involved the value of σ (see Example 3.3.6.)

However, when σ^2 is unknown, we need to estimate it, and we usually use

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

to estimate σ^2. We illustrate the procedure through the following example:

Example 3.6.1: Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$ with unknown mean μ. Let $H_0 : \mu = 0$, $H_A : \mu = 1$.

By the Neyman-Pearson lemma among all tests with significance level α, the test that rejects for the small values of the likelihood ratio has the highest power. Now, the likelihood ratio is

$$\Lambda\left(X_1, \ldots, X_n\right) = \frac{\ell\left(H_0 | X_1, \ldots, X_n\right)}{\ell(H_A | X_1, \ldots, X_n)}$$

$$= \frac{f\left(X_1, \ldots, X_n | \mu = 0\right)}{f\left(X_1, \ldots, X_n | \mu = 1\right)} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} X_i^2\right)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(X_i - 1\right)^2\right)}$$

$$= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(X_i - 1\right)^2\right)} = \exp\left[\frac{1}{2\sigma^2}\left(\sum_{i=1}^n \left(X_i - 1\right)^2 - \sum_{i=1}^n X_i^2\right)\right]$$

$$= \exp\left(\frac{n}{2\sigma^2} - \frac{1}{\sigma^2}\sum_{i=1}^n X_i\right)$$

Hence, the likelihood ratio is small if $\frac{1}{\sigma^2}\left(\sum_{i=1}^n X_i - \frac{n}{2}\right) = \frac{n}{\sigma^2}\left(\bar{X} - \frac{1}{2}\right)$ is large, or equivalently if $\bar{X}$ is large. So the rejection region of the MP test still is of the type $\bar{X} > c$. We need to choose c so as the prescribed level is attained, i.e., $P\left(\bar{X} > c\right) = \alpha$ when H_0 is true, i.e., the true mean is 0.

Now, under H_0, $\bar{X} \sim N\left(0, \frac{\sigma^2}{n}\right)$, and the value of σ^2 is unknown. So, we need to estimate it by s^2, and use the fact that under H_0, $\frac{\sqrt{n}\bar{X}}{s} \sim t_{n-1}$. Therefore, the rejection region of the level α MP test is now $\bar{X} > \frac{s}{\sqrt{n}} t_{n-1;\alpha}$ as $P\left(\bar{X} > \frac{s}{\sqrt{n}} t_{n-1;\alpha}\right) = P\left(\frac{\sqrt{n}\bar{X}}{s} > t_{n-1;\alpha}\right) = \alpha$.

Notes:

1. Replacing 1 by μ_A in the above example, where $\mu_A > 0$, similar calculations as Example 3.3.3 shows that the same rejection region, $\bar{X} > \frac{s}{\sqrt{n}} t_{n-1;\alpha}$ works for $H_0 : \mu = 0$ vs $H_A : \mu = \mu_A$ for the level α MP test.

2. Similar to Example 3.3.4, as we see that the rejection region is free of μ_A, the above test is also a level α UMP test for $H_0 : \mu = 0$ vs $H_A : \mu > 0$.

3. Replacing 0 by $\mu_0 (H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0)$ changes the rejection region for the level α UMP test to $\bar{X} > \mu_0 + \frac{s}{\sqrt{n}} t_{n-1;\alpha}$ (Left as an exercise.)

4. The above results could also be obtained by the duality of confidence interval argument.

5. Either by a duality of the confidence interval argument, or just by imitating Example 3.4.1, we get that for $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, the rejection region for a level α test is

$$\left|\bar{X} - \mu_0\right| > \frac{s}{\sqrt{n}} t_{n-1;\frac{\alpha}{2}}.$$

Note that this test is not UMP.

3.7 Test for Goodness of Fit

Consider i.i.d. observations $X_1, X_2, \ldots, X_n$ from an unknown distribution F. Suppose one requires to test whether F is a certain suspected distribution F_0. Set $H_0 : F = F_0$, and $H_A : F \neq F_0$.

One way to perform this test is to group the observations into a number of classes, each of which is expected to contain at least a few (usually 5 or more) observations.

Setup: Let the number of observations in the i-th group be O_i. Let the expected number of observations in that group under F_0 be E_i.

The statistic

$$\chi_0^2 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

, called Pearson's χ^2 statistic, is approximately chi-squared, χ^2, with degrees of freedom $d =$ number of cells (k) - number of parameters fitted under $H_0(p)$ -1 when H_0 is true.

We can use χ_0^2 as a test statistic to test H_0 against H_A rejecting the null if χ_0^2 is large. Hence, the critical region of the test is $\chi_0^2 > \chi_{d;\alpha}^2$ where d is the number of degrees of freedom determined as above and α is the level of the test. Note that this test is not UMP.

Example 3.7.1 Checking for the Fairness of a Die

Suppose a die is cast independently 300 times, and the numbers 1 through 6 appear with the frequencies as recorded below.

1	2	3	4	5	6
55	42	52	39	60	52

Question: is the die fair?

Hypotheses: The null hypothesis is

$$H_0 : \quad \text{The die is fair, i.e. } P(1) = P(2) = \cdots = P(6) = 1/6.$$

The alternative hypothesis is

$$H_A : \quad \text{The die is not fair, i.e. some of the probabilities do not equal } 1/6.$$

Expected Frequencies: Here $n = 300$, so under H_0 the expected frequency of each face is $300/6 = 50$. Hence, we get the following:

Face value (i)	1	2	3	4	5	6
Observed Value (O_i)	55	42	52	39	60	52
Expected Value (E_i)	50	50	50	50	50	50

Test statistic: As all the classes have expected frequency > 5, the chi-squared statistic can be calculated:

$$\chi_0^2 = \frac{(55-50)^2}{50} + \frac{(42-50)^2}{50} + \frac{(52-50)^2}{50} + \frac{(39-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(52-50)^2}{50} = 6.36.$$

d.f.: Here $k = 6$, $p = 0$ (as no parameter was estimated), so $d = k - p - 1 = 5$.

P-value: The null distribution is the χ^2 distribution with $df = 5$; hence we obtain a p-value of 0.2727, the probability of getting a value of 6.36 or higher from a χ_5^2 distribution (using a software R).

Cut-off for test at 5%: We observe that $\chi^2_{5;0.05} = 11.07$.

Conclusion: As $\chi^2_{5;0.05} = 11.07 > \chi_0^2 = 6.36$, at a 5% level of significance there is no real evidence to suggest that the data do not follow a Poisson distribution, so we accept the null at 5%level of significance.

Example 3.7.2 Fitting a Poisson Distribution: Let X be the number of defects in printed circuit boards. A random sample of $n = 60$ printed circuit boards is taken and the number of defects, $X_1, X_2, \ldots, X_{60}$, recorded. The results were as follows:

Number of Observed Defects (i)	Frequency (O_i)
0	32
1	15
2	9
3	4

Question: Does the assumption of a Poisson distribution seem appropriate as a model for these data?

$$H_0 : X \sim \text{ Poisson}$$

$$H_A : X \text{ does not follow a Poisson distribution.}$$

The parameter, say λ, of the (assumed) Poisson distribution under H_0 is unknown so must be estimated from the data by the MLE of λ, the sample mean:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=0}^{3} i \times O_i = \frac{1}{60}\left(0 \times 32 + 1 \times 15 + 2 \times 9 + 3 \times 4\right) = 0.75.$$

Using the Poisson distribution with $\lambda = 0.75$ we can compute p_i, the hypothesised probabilities associated with each class. From these we can calculate the expected frequencies (under the null hypothesis):

$$p_0 = P\left(X = 0\right) = \frac{e^{-0.75}\left(0.75\right)^0}{0!} = 0.472 \Rightarrow E_0 = 0.472 \times 60 = 28.32$$

$$p_1 = P\left(X = 1\right) = \frac{e^{-0.75}\left(0.75\right)^1}{1!} = 0.354 \Rightarrow E_1 = 0.354 \times 60 = 21.24$$

$$p_2 = P\left(X = 2\right) = \frac{e^{-0.75}\left(0.75\right)^2}{2!} = 0.133 \Rightarrow E_1 = 0.133 \times 60 = 7.98$$

$$p_3 = P(X \geq 3) = 1 - (p_0 + p_1 + p_2) = 0.041 \Rightarrow E_3 = 0.041 \times 60 = 2.46$$

Here the chi-squared goodness of fit test is not valid as the expected frequency of the class corresponding to the value 3 (or more) is too small (< 5). If an expected frequency is too small, two or more classes can be combined. In the above example the expected frequency in the last class is less than 5, so we should combine the last two classes to get:

Number of Defects	Observed Frequency (O_i)	Expected Frequency (E_i)
0	32	28.32
1	15	21.24
≥ 2	13	10.44

The chi-squared test statistic can now be calculated:

$$\chi_0^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94.$$

The number of degrees of freedom is $k - p - 1$. Here we have $k = 3$ classes (as we combined the last two) and we have $p = 1$ because we had to estimate one parameter (the rate, λ) from the data. So, $d = 3 - 1 - 1 = 1$. For the distribution with df $= 1$, we obtain a p-value of 0.0864 (the probability of getting a value of 2.94 or higher). We conclude that at a 5% level of significance there is no real evidence to suggest that the data do not follow a Poisson distribution, so we do nt reject the null at 5% level of significance.

Alternatively, we observe that $\chi^2_{1;0.05} = 3.84 > \chi_0^2 = 2.94$, and hence we do not reject H_0, the null hypothesis at 5% level of significance.

3.8 Test for Independence

We now discuss a method to test whether two jointly distributed random variables are independent. This idea of the test is similar to the test of goodness of fit: we check if the assumption of independence provides a good fit.

Following is the set-up:

Suppose that N i.i.d. pairs with same distribution as (discrete) random variables (X, Y) are observed. Suppose that the frequencies of the different values of the pair are tabulated in a table as follows:

	X				
Y	x_1	x_2	$\cdots$	x_m	Total for Y
y_1	f_{11}	f_{21}	$\cdots$	f_{m1}	$f_{.1}$
y_2	f_{12}	f_{22}	$\cdots$	f_{m2}	$f_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_n	f_{1n}	f_{2n}	$\cdots$	f_{mn}	$f_{.n}$
Total for X	$f_{1.}$	$f_{2.}$	$\cdots$	$f_{m.}$	N

From the table, we can say that the estimates of the probabilities $P(X = x_i, Y = y_j)$ would be f_{ij}/N, and the estimates of the marginal probabilities $P(X = x_i)$ and $P(Y = y_j)$ would be $\frac{f_{\text{i.}}}{N}$ and $\frac{f_{\text{.j}}}{N}$ respectively. However, if X and Y are independent,

then $P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j)$, and hence we expect (i, j)-th cell to be around $f_{i.} \times f_{.j}/N$, i.e. $f_{ij} \approx f_{i.} \times f_{.j}/N$.

Hence, to test H_0: X and Y are independent against H_A: X and Y are not independent, we can use

$$\chi_0^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{\text{all cells}} \frac{\left(f_{ij} - \frac{f_{i.} \times f_{.j}}{N}\right)^2}{f_{i.} \times f_{.j}/N}$$

and reject for large values of χ_0^2. Notice that here we are effectively fitting the marginal probabilities: $(m-1) + (n-1)$ of them (think why!), so we have

$$d.f. = mn - ((m-1) + (n-1)) - 1 = (m-1)(n-1).$$

We now look at an example:

Example 3.8.1: A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by their gender (male or female) and by their voting preference (Conservative, Labour, or others). Results are shown in the contingency table below.

	V	oting Preferences		
	Conservative	Labour	Others	
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

So here we are going to test H_0: gender and voting preferences are independent against H_A: gender and voting preferences are not independent at a 5% level of significance.

The expected frequencies are respectively $E_{11} = 180$; $E_{21} = 180$; $E_{31} = 40$; $E_{12} = 270$; $E_{22} = 270$; $E_{32} = 60$. Hence

$$\chi_0^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40}$$

$$+ \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60} = 16.2.$$

Here, $m = 3$, $n = 2$; so $d.f. = 2 \times 1 = 2$, so the cut-off at 5% is 5.99. Since the observed value of $\chi_0^2 >$ the cut-off at 5%, we conclude that there is significant relationship between gender and voting preferences at a 5% level of significance.

3.9 Comparing Two Independent Normal Samples

In this section we compare two samples from normal distributions. To facilitate calculations, we would assume that the samples have the same variance, but possibly different expectations. We shall now discuss a method to check if the two samples are drawn from populations with the same mean, which would mean that they are from the same distribution.

Let us now formulate the problem. Consider n i.i.d. observations $X_1, X_2, \ldots, X_n$ from $N\left(\mu_1, \sigma^2\right)$ and m more i.i.d. observations $Y_1, Y_2, \ldots, Y_m$ from $N\left(\mu_2, \sigma^2\right)$, μ_1 and μ_2 unknown. We set $H_0 : \mu_1 = \mu_2$. The possible alternatives are:

$$H_{A_1} : \mu_1 \neq \mu_2$$
$$H_{A_2} : \mu_1 > \mu_2$$
$$H_{A_3} : \mu_1 < \mu_2$$

The first of these is a two-sided alternative, and the other two are one-sided alternatives. The first hypothesis is appropriate if deviations could in principle go in both direction, and the latter two are appropriate if it is believed that any deviation must be in one specific direction. In practice, such a priori information is rarely available, so it is more prudent to use the both sided alternative.

Computations:

From the given conditions, $\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$ and $\bar{Y} \sim N\left(\mu_2, \frac{\sigma^2}{m}\right)$ and since the two samples are independent, $\bar{X}$ and $\bar{Y}$ are independent, giving

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

a) If σ^2 is known, then a confidence interval for $(\mu_1 - \mu_2)$ could be based on

$$Z = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

which follows a standard normal distribution. Therefore a $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ would be

$$\left(\bar{X} - \bar{Y}\right) \pm z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}.$$

b) Generally, σ^2 is not known, and hence it would need to be estimated by

$$s_p^2 = \frac{(n-1)\, s_X^2 + (m-1)\, s_Y^2}{n + m - 2},$$

where

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \text{ and } s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2.$$

Now, $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2/\sigma^2 \sim \chi^2_{n-1}$; $\sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2/\sigma^2 \sim \chi^2_{m-1}$; and since they are independent, $\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right)/\sigma^2 \sim \chi^2_{n+m-2}$.

Further, $\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$ is independent of $\left(\bar{X}-\bar{Y}\right)$, and hence of Z, which means the statistic

$$t=\frac{Z}{s_p}=\frac{\left(\bar{X}-\bar{Y}\right)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}/\left(\frac{s_p}{\sigma}\right)=\frac{\left(\bar{X}-\bar{Y}\right)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

follows a t distribution with $(n+m-2)$ degrees of freedom.

Based on these calculations, a $100(1-\alpha)\%$ confidence interval for $(\mu_1-\mu_2)$ would be

$$\left(\bar{X}-\bar{Y}\right)\pm t_{n+m-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}.$$

c) Under H_0, we have $\mu_1=\mu_2$, i.e. $\bar{X}-\bar{Y}\sim N\left(0,\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$. Suppose σ is known. Against $H_{A_1}:\mu_1\neq\mu_2$ the level α test could be obtained from part (a) by the duality of the tests and the confidence intervals, using the confidence interval $\left(\bar{X}-\bar{Y}\right)\pm z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}$ as follows:

Do not reject H_0 if $0\in\left(\bar{X}-\bar{Y}\right)\pm z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}$, and reject otherwise;

i.e. do not reject H_0 if $\left|\bar{X}-\bar{Y}\right|\leq z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}$, and reject otherwise;

Alternative Justification: See that we should reject H_0 if $\left|\bar{X}-\bar{Y}\right|$ is large due to the form of the alternative. Now, under H_0, $\bar{X}-\bar{Y}\sim N\left(0,\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$.

That means, $Z=\frac{(\bar{X}-\bar{Y})}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim N(0,1)$. Now, the rejection region at level α must satisfy $P\left(\left|\bar{X}-\bar{Y}\right|>c\right)=\alpha ar$ (*d to be estimated by*, or $P\left(|Z|>\frac{c}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)=\alpha$, which leads to $\frac{c}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}=z_{\frac{\alpha}{2}}$, giving the rejection region and hence the acceptance region.

d) Now suppose σ is unknown. Then, against $H_{A_1}:\mu_1\neq\mu_2$ the level α test could be obtained from part (b) by the duality of the tests and the confidence intervals, using the confidence interval $\left(\bar{X}-\bar{Y}\right)\pm t_{n+m-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$ as follows:

Do not reject H_0 if $0\in\left(\bar{X}-\bar{Y}\right)\pm t_{n+m-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$, and reject otherwise.

i.e. do not reject H_0 if $\left|\bar{X}-\bar{Y}\right|\leq t_{n+m-2;\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$, and reject otherwise.

Note that none of the above tests is UMP.

Example 3.9.1: Two methods, A and B, were used to determine the latent heat of ice. The investigators wanted to find out by how much the methods differed, if they did. The following table gives the change in total heat from ice at $-0.72^{\circ}C$ to water at $0^{\circ}C$ in calories per gram:

Method A	Method B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

Here assuming the two samples to be normal with variances equal but unknown, and means μ_1 and μ_2:

a) We find a 95% confidence interval for $(\mu_1 - \mu_2)$, and then

b) test $H_0 : \mu_1 = \mu_2$ against $H_{A_1} : \mu_1 \neq \mu_2$ at 5% level of significance.

Now, if we call the first sample X and the second sample Y, then $\bar{X} = 80.02$; $s_X = 0.024$; $\bar{Y} = 79.98$; $s_Y = 0.031$, giving $s_p^2 = \frac{12 \times s_X^2 + 7 \times s_Y^2}{19} = 0.072 \Rightarrow s_p = 0.027$. Now, $\bar{X} - \bar{Y} = 0.04$, giving the confidence interval as

$$0.04 \pm t_{19;0.025} \times 0.027 \times \sqrt{\frac{1}{13} + \frac{1}{8}} = (0.015,\ 0.065)$$

since $t_{19;0.025} = 2.093$.

For the testing part, we see that the value 0 is not in the confidence interval found above, and hence we reject H_0 at 5% level of significance.

Alternatively, $\left|\bar{X} - \bar{Y}\right| = 0.04$, and

$$t_{n+m-2;\frac{\alpha}{2}} s_p \sqrt{\frac{1}{m} + \frac{1}{m}} = t_{19;0.025} \times 0.027 \times \sqrt{\frac{1}{13} + \frac{1}{8}}$$

$$= 2.093 \times 0.027 \times \sqrt{\frac{1}{13} + \frac{1}{8}} = 0.0254,$$

which means $\left|\bar{X} - \bar{Y}\right| > t_{n+m-2;0.025} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$; we should reject H_0 at 5% level of significance.

3.10 Comparing Paired Samples

In the previous section we analysed two independent samples. In many experiments, the samples are paired, for example consider the weight of some students before and after the

Easter break, or the runs scored by a group of English batsmen in the two innings of a test match.

Setup: We now consider i.i.d. pairs (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$, where we allow the possibility that X_i and Y_i can be dependent (for the same i), but consider X_i and Y_j to be independent whenever $i \neq j$. We want to check if $\mu_1 = \mu_2$, where $E(X_i) = \mu_1$, $E(Y_j) = \mu_2$ for all i and j.

We shall work with $D_i = X_i - Y_i$ and test $H_0 : \mu_1 = \mu_2$ against one of the following three possible alternatives:

$$H_{A_1} : \mu_1 \neq \mu_2$$
$$H_{A_2} : \mu_1 > \mu_2$$
$$H_{A_3} : \mu_1 < \mu_2$$

whichever is appropriate. Note that $E(D_i) = \mu_1 - \mu_2 = \mu_D$ (say), and $\text{Var}(D_i) = \text{Var}(X_i - Y_i) = \text{Var}(X_i) + \text{Var}(Y_i) - 2\,\text{Cov}(X_i, Y_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} = \sigma_D^2$ (say)

Hence, $E(\bar{D}) = \mu_1 - \mu_2$, $\text{Var}(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})$.

To make the computations easier, let us assume that the D_i's are also normal, which tells us that

$$\bar{D} \sim N\left(\mu_1 - \mu_2, \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\right) \equiv N\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

(Note that by the central limit theorem $\bar{D}$ is approximately normal with above mean and variance for large n.) Then, we can rewrite our hypotheses as:

$$H_0 : \mu_D = 0$$
$$H_{A_1} : \mu_D \neq 0$$
$$H_{A_2} : \mu_D > 0$$
$$H_{A_3} : \mu_D < 0$$

The tests could now be obtained as before by using the methods discussed in Section 3.6.

Example 3.10.1: The following table lists the number of cigarettes 11 people smoke daily on average before and after going through a rehabilitation programme. We want to check if the programme was effective.

Before (X_i)	After(Y_i)	Difference(D_i=X_i−Y_i)
27	25	2
29	25	4
37	27	10
56	44	12
46	30	16
82	67	15
57	53	4
80	53	27
61	52	9
59	60	−1
43	28	15

Here $H_0 : \mu_D = 0$, and we set $H_{A_2} : \mu_D > 0$ as we are interested in seeing if the programme was effective, which would mean a significant reduction in the smoking rate. So, we would reject the null for large values of $\bar{D}$. Assuming normality of the samples, this requires a one sided t-test as described in Section 3.6.

Here, $\bar{D} = 10.27$ and $s_D = 7.976$. Note that $s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}$. There are 11 samples, so the rejection region would be $\bar{D} > 0 + \frac{s_D t_{10;\alpha}}{\sqrt{11}}$. For $\alpha = 0.05$, the rejection region is $\bar{D} > 7.976 \times \frac{1.812}{\sqrt{11}} = 4.359$. Since the observed value $10.27 > 4.359$, we reject the null hypothesis at 5% level of significance. There is evidence to reject the hypothesis that the treatment was not effective.

Chapter 4

Regression Techniques

4.1 Introduction

Regression relates the expected values of one random variable, Y, to the known values of associated variable(s) X. The term "regression" comes from the work of the geneticist Sir Francis Galton (1822-1911), who studied the sizes of seeds, their offspring; and the heights of fathers and their sons. In both cases, he found that the offspring of larger than average size parents tended to be smaller than their parents and that the offspring of smaller than average size parents tended to be larger than their parents. He called this "regression towards mediocrity".

Practical Examples

1. To relate the effects of drug to dosage level. The dosage is controlled by the investigator and causality is implied. Other factors such as age, weight, sex could also be included. (experimental/observational).

2. To investigate the relevance of certain variables to the value of a nominated variable. (Observational or survey data – inferring causality is a problem, but is usually of prime interest).

Note that the variables X, the so-called "*independent*" variables, are considered as *fixed values in the regression setup*. To emphasise that, we shall always denote them by small letters. You may consider them as previously measured observed values of the associated variables. They are also called *explanatory or predictor variables*. The variable Y is the *dependent* variable.

4.2 Linear Regression with One Predictor Variable

Set up:

Observations: values of the variable of interest, say $Y_1, Y_2, \ldots, Y_n$.

Co-variates/Independent variables/ Predictor variable values: $x_1, x_2, \ldots, x_n$.

The aim of the linear regression exercise is to fit a line that predicts the value of the Y-variable as best possible given x. Ideally, we would expect a line $Y = a + bx$ but in

reality the linear fit is not exact, and there are errors, hence the model could be written as

$$Y_i = a + bx_i + \epsilon_i$$

where ϵ_i is the error in the linear fit for the i-th observation. Given the data we aim to obtain a *line of best fit*:

Definition 4.1. *Among all possible straight lines, the line of best fit corresponds to those values of a and b that minimise $Q = \sum_{i=1}^{n} (Y_i - a - bx_i)^2$.*

The values of a and b that satisfy the above definition are called the *least squares estimates* as they correspond to the least of the sum of squares. The estimates are the values that minimise the above sum of squares, and can be obtained by setting the following partial derivatives to 0:

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} (Y_i - a - bx_i) \tag{4.1}$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^{n} (Y_i - a - bx_i)\, x_i \ldots (2) \tag{4.2}$$

Note that from (4.1), we have for the solutions $\hat{a}$ and $\hat{b}$ that

$$\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}x_i) = 0 \Rightarrow \sum_{i=1}^{n} Y_i = n\hat{a} + \hat{b} \sum_{i=1}^{n} x_i \Rightarrow \bar{Y} = \hat{a} + \hat{b}\bar{x} \Rightarrow \hat{a} = \bar{Y} - \hat{b}\bar{x}.$$

(4.2) gives $\sum_{i=1}^{n} \left(Y_i - \hat{a} - \hat{b}x_i\right) x_i = 0$ which, after some simplifications, yields

$$\hat{b} = \frac{S_{\mathrm{XY}}}{S_{\mathrm{XX}}},$$

where

$$S_{\mathrm{XY}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \right];$$

$$S_{\mathrm{XX}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right].$$

So the line of best fit is

$$y - \bar{Y} = \frac{S_{\mathrm{XY}}}{S_{\mathrm{XX}}} (x - \bar{x}).$$

This line passes through the point of means, $(\bar{x}, \bar{Y})$.
We denote the *fitted values of Y* by $\hat{Y}$, i.e.

$$\hat{Y}_i = \hat{a} + \hat{b}x_i = \bar{Y} + \frac{S_{XY}}{S_{XX}} (x_i - \bar{x}).$$

The values $e_i = Y_i = \hat{Y}_i$ denote the *residuals* of the fitted model.

Now, note that the sample correlation coefficient can be written as

$$r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}};$$

which means that the regression line has the alternative form

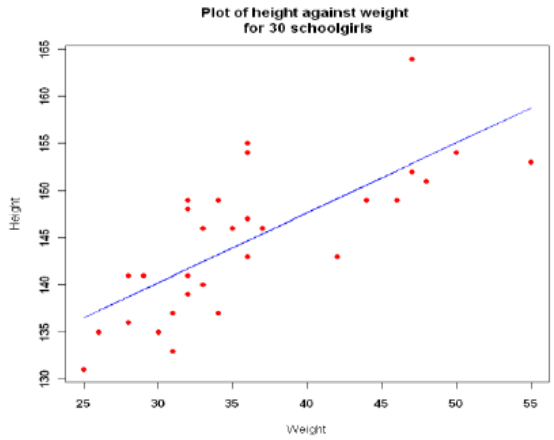
$$y = \bar{Y} + r_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}}\,(x - \bar{x})\,.$$

Example 4.2.1: We now consider the heights and weights of 30 eleven year old girls attending Heaton Middle School, Bradford UK, obtained in 1983; (so they are all a bit older, a bit taller and a bit heavier now.) We model height as the dependent variable Y and use weight as x:

Height	Weight	Height	Weight	Height	Weight
135	26	141	28	149	46
146	33	136	28	147	36
153	55	154	36	152	47
154	50	151	48	140	33
139	32	155	36	143	42
131	25	133	31	148	32
149	44	149	34	149	32
137	31	141	32	141	29
143	36	164	47	137	34
146	35	146	37	135	30

Data: (x_i, Y_i), $i = 1, 2, \ldots, 30$.
Model: $Y_i = a + bx_i + \epsilon_i$
Here, $\bar{Y} = 144.8$, $\bar{x} = 36.17$, $S_{\mathrm{XY}} = 43.97$, $S_{\mathrm{XX}} = 59.25$.
Hence, $\hat{b} = \frac{S_{XY}}{S_{XX}} = 0.742$; $\hat{a} = \bar{Y} - \hat{b}\bar{x} = 117.962$.
The fitted line is therefore

$$y \;=\; 117.962 \;+\; 0.742x$$



The blue line is the line of best fit, the regression line.

4.2.1 Properties of the Fitted Regression Line

Earlier in this section, we have defined the residuals as $e_i = Y_i - \hat{Y}_i$. Some important properties of the residuals are

1. $\sum_{i=1}^{n} e_i = 0$. Note that this means $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$.
2. $\sum_{i=1}^{n} x_i e_i = 0$.
3. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.
4. The line of regression passes through $(\bar{x}, \bar{Y})$.

Proof:

1. $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} \left(\bar{Y} + \frac{S_{XY}}{S_{XX}}(x_i - \bar{x})\right) = n\bar{Y} + \frac{S_{XY}}{S_{XX}}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = \sum_{i=1}^{n} Y_i + \frac{S_{XY}}{S_{XX}}(n\bar{x} - n\bar{x}) = \sum_{i=1}^{n} Y_i$.
2. $\sum_{i=1}^{n} x_i\left(\hat{Y}_i - Y_i\right) = \sum_{i=1}^{n} x_i\left(\bar{Y} + \frac{S_{XY}}{S_{XX}}(x_i - \bar{x}) - Y_i\right) = \sum_{i=1}^{n} x_i\left(\bar{Y} - Y_i\right) + \frac{S_{XY}}{S_{XX}}\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \left(n\bar{x}\bar{Y} - \sum_{i=1}^{n} x_i Y_i\right) + \frac{S_{XY}}{S_{XX}}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = -(n-1)S_{XY} + \frac{S_{XY}}{S_{XX}}(n-1)S_{XX} = 0$.
3. $\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} \left(\hat{a} + \hat{b}x_i\right) e_i = \hat{a}\sum_{i=1}^{n} e_i + \hat{b}\sum_{i=1}^{n} x_i e_i = \hat{a} \times 0 + \hat{b} \times 0 = 0$.
4. Left as an exercise

4.2.2 Model Description, Assumptions and Some Expectations

Note that we considered Y to be a random variable, and a, b and x as constants. That implies that the errors are also random. For a simple linear regression model, we need to assume

1. $E(\epsilon_i) = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ and
3. The ϵ_i's are independent of each other.
4. The ϵ_i's are normal random variables, i.e. $\epsilon_i \sim N\left(0, \sigma^2\right)$.

The above tells us that in the model, the observations are independent with constant variance and the errors are expected to be normal with mean zero.

The assumptions lead to the following observations:

1. $Y_i \sim N\left(a + bx_i, \sigma^2\right)$, independent of each other.
2. $\bar{Y} \sim N\left(a + b\bar{x},\ \sigma^2\right)$, so $E\left(\bar{Y}\right) = a + b\bar{x}$.
3. $E\left(\hat{b}\right) = E\left(\frac{S_{XY}}{S_{XX}}\right) = E\left(\frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) = \frac{\sum_{i=1}^{n} x_i E(Y_i) - n\bar{x}E\left(\bar{Y}\right)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} x_i(a + bx_i) - n\bar{x}(a + b\bar{x})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = a\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) + b = b$.
4. $E(\hat{a}) = E\left(\bar{Y} - \hat{b}\bar{x}\right) = E\left(\bar{Y}\right) - E\left(\hat{b}\right)\bar{x} = a + b\bar{x} - b\bar{x} = a$.

4.2.3 Estimation of the Error Variance

We denoted the error variance, $\text{Var}(\epsilon_i)$, by σ^2. From earlier sections, we know that

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(\epsilon_i-\bar{\epsilon})^2\right)=\sigma^2.$$

However, we do not observe $\epsilon_i = Y_i - (a+bx_i)$, what we observe is only an estimate of it, $e_i = Y_i - \left(\hat{a}+\hat{b}x_i\right)$. Hence, we base the estimate of σ^2 on the *sum of squared errors*

$$SSE=\sum_{i=1}^{n}(e_i-\bar{e})^2=\sum_{i=1}^{n}e_i^2.$$

Now, it can be shown that

$$\frac{\sum_{i=1}^{n}e_i^2}{\sigma^2}\sim\chi^2$$

with $(n-2)$ degrees of freedom, where the 2 degrees of freedom are "lost" in estimating a and b to obtain e_i.

Hence, the *mean squared error*

$$MSE=\sum_{i=1}^{n}e_i^2/(n-2)$$

is an unbiased estimate of σ^2. We shall later discuss a general rule that would justify this finding.

4.3 Linear Regression with More than One Predictor

Suppose we have to fit the following model:

$$Y_i=a+b_1x_{1i}+b_2x_{2i}+\cdots+b_kx_{\text{ki}}+\epsilon_i$$

Where $\epsilon_i\sim N\left(0,\sigma^2\right)$, i.i.d. This means that we now have information on k different variables, instead of just one like before.

The parameters $a, b_1, b_2, \ldots, b_k$ can be fitted by the least squares technique just as we have seen in Section 4.2. The method is just the same:

Consider $Q=\sum_{i=1}^{n}\epsilon_i^2=\sum_{i=1}^{n}(Y_i-a-b_1x_{1i}-b_2x_{2i}-\cdots-b_kx_{ki})^2$, compute the partial derivatives with respect to the $(k+1)$ parameters and equate them to zero to solve for the estimates.

Since we now estimate $(k+1)$ parameters, the estimate of σ^2 will now be $\sum_{i=1}^{n}e_i^2/(n-(k+1))$.

The details are omitted as the computations are fairly long. There is a quicker alternative way based on matrix algebra that would give us a formula to fit the linear regression models with any number of regressors without going through the partial derivative routine. Moreover, that would give us the standard error for the parameter estimates, which we can not get easily using the above derivations. Hence, we are now going to discuss the matrix method to regression.

4.3.1 Matrix Approach to Linear Regression

Consider first the model

$$Y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n,$$

with $\epsilon_i \sim N\left(0, \sigma^2\right)$**, i.i.d.**

Note that we can rewrite the model in the vector form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} a + bx_1 + \epsilon_1 \\ a + bx_2 + \epsilon_2 \\ \vdots \\ a + bx_n + \epsilon_n \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

If we denote

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$$

;

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}^T;$$

$$\beta = \begin{pmatrix} a & b \end{pmatrix}^T \quad \text{and} \quad \epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$$

then we can rewrite the above equation as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

Note that the above implies that $E(\mathbf{Y}) = \mathbf{X}\beta$, $V(\mathbf{Y}) = \sigma^2\mathbf{I}$; and further $Y_i \sim N\left(a + bx_i, \sigma^2\right)$ for $i = 1, 2, \ldots, n$.

Now, recall that for the solution of the least square equations, we obtained the following two equations

$$\sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}x_i) = 0$$

$$\sum_{i=1}^{n} \left(Y_i - \hat{a} - \hat{b}x_i\right) x_i = 0$$

The above equations could respectively be re-written in matrix terms as

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) = 0; \quad \text{or} \quad \mathbf{1}^T \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) = 0;$$

$$\text{and} \quad \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) = 0;$$

$$\text{or} \quad \mathbf{x}^T \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) = 0,$$

where $\mathbf{1}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$; $\mathbf{x}^T = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$; $\mathbf{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}$.

Together, they simplify to

$$\mathbf{X}^T\left(\mathbf{Y}-\mathbf{X}\hat{\beta}\right)=\begin{pmatrix}0\\0\end{pmatrix}$$

$$\text{or, }\ \mathbf{X}^T\mathbf{Y}=\mathbf{X}^T\mathbf{X}\beta.$$

Now, notice that

$$\mathbf{X}^T\mathbf{X}=\begin{bmatrix}\mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{x}\\ \mathbf{x}^T\mathbf{1} & \mathbf{x}^T\mathbf{x}\end{bmatrix}=\begin{bmatrix}n & \sum_{i=1}^n x_i\\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix},$$

which is an invertible matrix.

Hence,

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

We have illustrated this result for one set of predictor variables, but this holds good for general cases as well, for example, if we consider k predictor variables as follows:

$$Y_i=a+b_1x_{1i}+b_2x_{2i}+\cdots+b_kx_{ki}+\epsilon_i$$

where $\epsilon_i\sim N\left(0,\sigma^2\right)$ i.i.d., then we can still write the model in vector notations as $\mathbf{Y}=\mathbf{X}\beta+\epsilon$, only difference now being that $\mathbf{X}=\begin{bmatrix}\mathbf{1} & \mathbf{x}_1 & \cdots & \mathbf{x}_k\end{bmatrix}$; $\mathbf{x}_j=\begin{pmatrix}x_{j1} & x_{j2} & \cdots & x_{jn}\end{pmatrix}^T$; for $j=1,2,\ldots,k$ and $\beta=\begin{pmatrix}a & b_1 & b_2 & \cdots & b_k\end{pmatrix}^T$. The solution is still given by $\hat{\beta}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$.

4.3.2 Expectation, Standard Error and Approximate Distribution of the Least Square Estimates

Note that

$$E\left(\hat{\beta}\right)=E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right]=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^TE\left(\mathbf{Y}\right)=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^TE\left(\mathbf{X}\beta+\epsilon\right)$$

$$=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta=\beta$$

Hence, $\hat{\beta}$ is unbiased for β.

Further, if we write $\mathbf{A}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, then the variance-covariance matrix for $\hat{\beta}$ is given by

$$V\left(\hat{\beta}\right)=V\left(\mathbf{AY}\right)=\mathbf{A}V\left(\mathbf{Y}\right)\mathbf{A}^T=\mathbf{A}\left(\sigma^2\mathbf{I}\right)\mathbf{A}^T=\sigma^2\mathbf{A}\mathbf{A}^T=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

$$=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

This tells us that the variances of the estimates of $a,b_1,\ldots,b_k$ are respectively σ^2s^{11}, σ^2s^{22}, …, $\sigma^2s^{(k+1),(k+1)}$ where we denote the (i,j)th element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ by s^{ij}. The corresponding standard errors are the respective square roots.

Note that for practical problems the unknown σ^2 would need to be estimated, and as we discussed before, an unbiased estimate is $s_e^2=\sum_{i=1}^n e_i^2/(n-(k+1))$.

Finally, the distribution of

$$\frac{\hat{b}_j - b_j}{s_e\sqrt{s^{j+1,j+1}}}$$

is t with $n-k-1$ degrees of freedom under the present set up, for $1 \leq j \leq k$.

Similarly,

$$\frac{\hat{a} - a}{s_e\sqrt{s^{11}}}$$

is also t with $n-k-1$ degrees of freedom.

Note that b_j gives the effect of the j-th independent variable, X_j, on Y. If we want to check whether X_j has any significant effect on Y, we should test $H_0 : b_j = 0$ against the alternative $H_A : b_j \neq 0$, where *rejecting null hypothesis* H_0 would mean that $X_{\mathbf{j}}$ has significant effect on Y.

Note that the statistic $\frac{\hat{b}_j - b_j}{s_e\sqrt{s^{j+1,j+1}}}$ reduces to $\frac{\hat{b}_j}{s_e\sqrt{s^{j+1,j+1}}}$ if $H_0 : b_j = 0$ is true, so under H_0, $\frac{\hat{b}_j}{s_e\sqrt{s^{j+1,j+1}}} \sim t_{n-k-1}$.

Hence, we could use this statistic to test $H_0 : b_j = 0$ against the alternative $H_A : b_j \neq 0$, rejecting H_0 for large values of

$$\left| \frac{\hat{b}_j}{s_e\sqrt{s^{j+1,j+1}}} \right| .$$

Clearly, the rejection region for a level α test for $H_0 : b_j = 0$ against $H_A : b_j \neq 0$ would be

$$\left| \frac{\hat{b}_j}{s_e\sqrt{s^{j+1,j+1}}} \right| > t_{n-k-1;\frac{\alpha}{2}}$$

A similar check is possible for the intercept "a" by testing $H_0 : a = 0$ against $H_A : a \neq 0$.

But it is less important as all it checks whether the intercept is zero or not, and that is a much less interesting problem that deciding whether to keep an actual variable in the model.

Example 4.3.1: Consider the following data set which provides measurements of the girth, height and volume of timber in 31 felled black cherry trees.

Girth	Height	Volume	Girth	Height	Volume
8.3	70	10.3	8.6	65	10.3
8.8	63	10.2	10.5	72	16.4
10.7	81	18.8	10.8	83	19.7
11.0	66	15.6	11.0	75	18.2
11.1	80	22.6	11.2	75	19.9
11.3	79	24.2	11.4	76	21.0
11.4	76	21.4	11.7	69	21.3
12.0	75	19.1	12.9	74	22.2
12.9	85	33.8	13.3	86	27.4
13.7	71	25.7	13.8	64	24.9
14.0	78	34.5	14.2	80	31.7
14.5	74	36.3	16.0	72	38.3
16.3	77	42.6	17.3	81	55.4
17.5	82	55.7	17.9	80	58.3
18.0	80	51.5	18.0	80	51.0
20.6	87	77.0			

The following table reports the summary results of regression of tree volumes on their girth and height using R.

Model: Volume $= a + b_1 \times$ Girth $+ b_2 \times$ Height $+ \epsilon$

Fitted model:

```
            Estimate  Std. Error  t value   Pr(>|t|)
Intercept   -57.9877      8.6382   -6.713   2.75e-07 ***
Girth         4.7082      0.2643   17.816   <2.0e-16 ***
Height        0.3393      0.1302    2.607     0.0145 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948, Adjusted R-squared: 0.9442
```

The "Estimate" column reports the least square estimates of the regression coefficients, the next column computes the standard error estimates. Note that in this particular problem we have $n = 31$ as there are 31 observations. Further, $k = 2$. Hence, the standard errors are computed based on $s_e^2 = \sum_{i=1}^{n} e_i^2/28$. The value of s_e is found to be 3.882, as reported above.

Further, $\frac{\hat{b}_j}{s_e\sqrt{s^{j+1,j+1}}} \sim t_{28}$ approximately under $H_0 : b_j = 0$ for $j = 1, 2$. Similarly, $\frac{\hat{a}}{s_e\sqrt{s^{11}}} \sim t_{28}$ under $H_0 : a = 0$.

The reported "t-values" are the statistics mentioned above, obtained by dividing the estimates by their corresponding estimated standard error. For example, the reported estimated standard error for $\hat{a}$ is $s_e\sqrt{s^{11}} = 8.6382$, and the corresponding t-value is $\frac{\hat{a}}{s_e\sqrt{s^{11}}} = \frac{-57.9877}{8.6382} = -6.713$.

The p-values, indicated by "Pr(¡—t—)", are computed based on the corresponding null hypotheses $H_0 : b_j = 0$, $j = 1, 2$ or $H_0 : a = 0$ against both sided alternatives, and they indicate the probability of obtaining a test statistic as extreme as the one obtained given the null distribution, here t_{28}. So, the p-value reported in the first column is $P(|t_{28}| > 6.713) = 2.75e - 07$.

4.4 Choice and Assessment of Model

When we fit a regression model with certain independent variables, there still remains the question whether the model makes sense. There could be two problems: (a) some of the independent variables used in the model might not really have any significant linear effect on the dependent variable or (b) as a whole the chosen group of independent variables may be superfluous. Hence, we need some measures to decide on which variables to keep in the model, and in general, assess the goodness of the fit. The p-values as obtained in the previous section give us some idea about importance of variables. Here is an alternative measure.

4.4.1 Choice of Regression Model: Coefficient of Determination and Adjusted R^2

We start with a measure of overall fit that can be used to compare various models:

Definition 4.2. *The overall fit of a linear regression model can be assessed using the coefficient of determination, denoted by R^2 and defined as*

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

where $SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$, *and* $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.

It can be shown that $0 \leq R^2 \leq 1$. R^2 measures how well a model has fit. The higher the value of R^2, the better, supposedly, is the fit. However, one problem is that the value of R^2 increases whenever a new variable is introduced in the model, regardless whether the new variable is useful or not.

Definition 4.3. *The adjusted R^2 is defined as*

$$\tilde{R}^2 = 1 - \frac{SSE}{SST}\frac{n-1}{n-k-1} = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right).$$

So, the adjusted R^2, or $\tilde{R}^2$, adjusts for the number of independent variables in the model. Unlike the R^2, the $\tilde{R}^2$ increases only if the new term improves the model more than would be expected by chance. The $\tilde{R}^2$ can be negative, and will always be less than or equal to the R^2.

The idea behind $\tilde{R}^2$: If a newly introduced independent variable does not add any extra information, then we have increased k by 1, and have not increased R^2 by much. As a result, $\tilde{R}^2$ might decrease, although R^2 will increase. Hence, it provides a method to check if adding an extra independent variable to a model makes sense.

When we compare different linear models, we would pick up the model with the highest $\tilde{R}^2$ as the "best". Note that $\tilde{R}^2$ can be negative.

Example 4.4.1: Consider the Cherry tree data discussed in Example 4.3.1. Using volume as the dependent variable and using the linear mode methods we fit the following models:

(1) Volume $= a + b_1 \times$ Girth

(2) Volume $= a + b_2 \times$ Height

(3) Volume $= a + b_1 \times$ Girth$+b_2 \times$ Height

Note that for the first two models $k = 1$, and for the final model $k = 2$. Hence, $\tilde{R}^2 = 1 - \frac{SSE}{SST}\frac{30}{30-k}$.

We get the value of $\tilde{R}^2$ to be respectively 0.933, 0.336 and 0.944 for the above 3 models. Hence, based on $\tilde{R}^2$, (3), the largest model, is the best.

4.4.2 Model Assessment using Plots

A linear regression model assumes the following four conditions:

1. $E(\epsilon_i) = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$, constant for all i
3. The ϵ_i's are independent of each other and
4. The ϵ_i's are normal random variables, i.e. $\epsilon_i \sim N\left(0, \sigma^2\right)$.

The first of the conditions is automatically satisfied by the method chosen to fit the model, but the other conditions need to be checked for to ensure that the fitted model is good. If it is found that the fitted model violates any of these conditions, the model fit cannot be considered as a good one.

These assessments are done by looking at the plot of residuals. The commonest plots of residuals are the following:

a) against index numbers i

b) against the fitted values

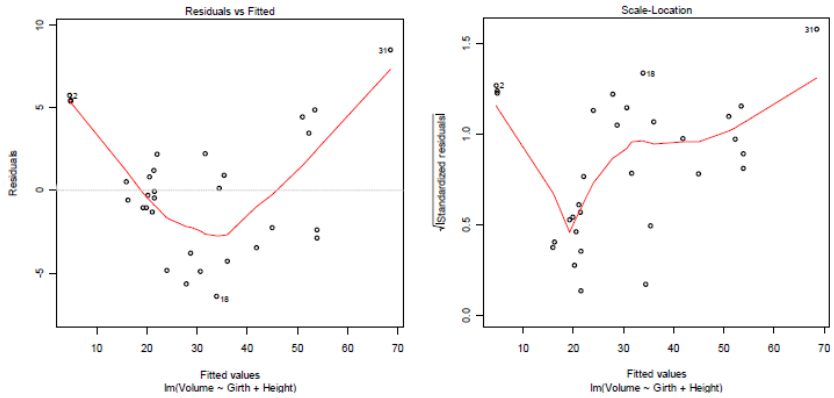
c) against normal quantiles

The aim of the first two types of plots is to check whether the errors really have constant variance and whether they are really independent. If the residuals show uneven spread, then the variance may not be constant. If they show some pattern, for example if the residuals are nearly on a straight line, or if the high values and the low values are clustered separately, that may be considered as an indication of lack of independence.

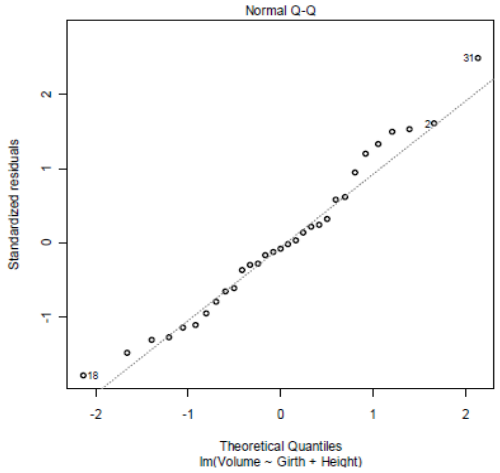
The last plot checks if the normality assumption is valid by looking at the plot of the residual quantiles against normal quantiles. If the plot is a straight line, at least

approximately, then the normality assumption is justifiable, whereas obvious departures suggest violation of the assumption.

Example 4.4.2:



Look at some of the residual plots that we obtained when we were performing the regression of tree volumes on their girth and height. The left plot above is of the residuals against the fitted values, whereas the right plot is a variation that looks at the plot of the square root of the absolute values of the "standardized" residuals ($e\text{sd}(e)$).



The error values corresponding to high fitted values seem high and that might be a pattern, but it is not really very strong. (In fact, the correlation of the fitted values and the residuals is 0.)

The plot at left is the plot against normal quantiles. Although it is not exactly a straight line, it very closely resembles one, so the fit seems ok.

It is not easy to analyse residual plots; a lot of times random behaviour seems like patterns, and vice versa.

Any good analysis requires consideration of a lot of factors; detailed discussions could be found in the textbooks.

4.5 Appendix 1: Derivatives with respect to vectors and the Derivation of the Matrix Form of the Least Squares Estimates

Definition 4.4. *Let $x \in \mathbb{R}^n$ be a column vector, and let $f : \mathbb{R}^n \to \mathbb{R}$. Then the derivative of f with respect to x is the row vector*

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

Definition 4.5. *The Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ is the square matrix of the second partial derivatives of f:*

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

Definition 4.6. *Let $x \in \mathbb{R}^n$ be a column vector, and let $f : \mathbb{R}^n \to \mathbb{R}^m$. Then the derivative of f with respect to x is the $m \times n$ matrix*

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f(x)_1}{\partial x_1} & \frac{\partial f(x)_1}{\partial x_2} & \cdots & \frac{\partial f(x)_1}{\partial x_n} \\ \frac{\partial f(x)_2}{\partial x_1} & \frac{\partial f(x)_2}{\partial x_2} & \cdots & \frac{\partial f(x)_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x)_m}{\partial x_1} & \frac{\partial f(x)_m}{\partial x_2} & \cdots & \frac{\partial f(x)_m}{\partial x_n} \end{pmatrix}.$$

This matrix is called the *Jacobian Matrix* of f.

Example 4.A.1: Let $u, x \in \mathbb{R}^n$. Suppose that u is a vector of constants. Then, $\frac{\partial \left(\sum_{i=1}^n u_i x_i\right)}{\partial x_i} = u_i \Longrightarrow \frac{\partial u^T x}{\partial x} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} = u^T$.

Example 4.A.2: Let $x \in \mathbb{R}^n$. Then $\frac{\partial \left(\sum_{i=1}^n x_i^2\right)}{\partial x_i} = 2x_i \Longrightarrow \frac{\partial x^T x}{\partial x} = \begin{pmatrix} 2x_1 & 2x_2 & \cdots & 2x_n \end{pmatrix} = 2x^T$.

Example 4.A.3: Suppose $\mathbf{A}$ is an $m \times n$ matrix and let $x \in \mathbb{R}^n$. Then writing the rows of $\mathbf{A}$ as $a_1^T, a_2^T, \ldots, a_m^T$, we have

$$\mathbf{A}x = \begin{pmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{pmatrix},$$

and hence

$$\frac{\partial \mathbf{A}x}{\partial x} = \begin{pmatrix} \frac{\partial a_1^T x}{\partial x} \\ \frac{\partial a_2^T x}{\partial x} \\ \vdots \\ \frac{\partial a_m^T x}{\partial x} \end{pmatrix} = \begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{pmatrix} = A.$$

Example 4.A.4: In the above set up, we have $\frac{\partial(x^T\mathbf{A}x)}{\partial x} = x^T\left(\mathbf{A}+\mathbf{A}^\top\right)$. The proof is left as an exercise.

Now consider our least square problem for regression: we want to minimise $f(\beta) = (Y-X\beta)^T(Y-X\beta)$ with respect to β, where $\beta \in \mathbb{R}^{k+1}$, and X is an $n\times(k+1)$ matrix, where the $(k+1)\times(k+1)$ matrix X^TX is invertible.

Now, $f(\beta) = (Y-X\beta)^T(Y-X\beta) = Y^TY - 2Y^TX\beta - \beta^TX^TX\beta$, as $Y^TX\beta = \beta^TX^TY$.

So,

$$\frac{\partial f(\beta)}{\partial \beta} = -2Y^T + \beta^T\left(X^TX + \left(X^TX\right)^T\right) = -2Y^TX + 2\beta^TX^T\mathrm{X}.$$

Hence,

$$\left.\frac{\partial f(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0 \Longrightarrow Y^TX = \hat{\beta}^TX^TX \Longleftrightarrow X^TX\hat{\beta} = X^TY \Longleftrightarrow \hat{\beta} = \left(X^TX\right)^{-1}X^T\mathrm{Y}.$$

To prove conclusively that this is the least square estimate, i.e. that this value of $\hat{\beta}$ minimises $f(\beta)$, we could compute the Hessian and show that it is negative definite. Alternatively, see that

$$(Y-X\beta)^T(Y-X\beta) = \left(Y-X\hat{\beta}+X\hat{\beta}-X\beta\right)^T\left(Y-X\hat{\beta}+X\hat{\beta}-X\beta\right)$$

$$= \left(Y-X\hat{\beta}+X\left(\hat{\beta}-\beta\right)\right)^T\left(Y-X\hat{\beta}+X\left(\hat{\beta}-\beta\right)\right)$$

$$= \left(Y-X\hat{\beta}\right)^T\left(Y-X\hat{\beta}\right) + 2\left(Y-X\hat{\beta}\right)^TX\left(\hat{\beta}-\beta\right) + \left(X\left(\hat{\beta}-\beta\right)\right)^TX\left(\hat{\beta}-\beta\right).$$

But, $\left(Y-X\hat{\beta}\right)^TX\left(\hat{\beta}-\beta\right) = \left(\hat{\beta}-\beta\right)^TX^T\left(Y-X\hat{\beta}\right) = \left(\hat{\beta}-\beta\right)^T\left(X^TY-X^TX\hat{\beta}\right) = 0$ as $X^TX\hat{\beta} = \left(X^TX\right)\left(X^TX\right)^{-1}X^TY = X^TY$.

Hence, $(Y-X\beta)^T(Y-X\beta) = \left(Y-X\hat{\beta}\right)^T\left(Y-X\hat{\beta}\right)+\left(X\left(\hat{\beta}-\beta\right)\right)^TX\left(\hat{\beta}-\beta\right) \geq \left(Y-X\hat{\beta}\right)^T\left(Y-X\hat{\beta}\right)$ for any β implying that $\hat{\beta}$ minimises $f(\beta)$.

4.6 Appendix 2: Regression Error and Fitted Values

Note that the fitted Y values will be given by

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

Let us write $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$. Then, $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. Note that $\mathbf{H}$ is a symmetric matrix.

Further, $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}$, i.e. $\mathbf{H}$ is idempotent.

Now,

$$E\left(\hat{\mathbf{Y}}\right) = E(\mathbf{H}\mathbf{Y}) = \mathbf{H}E(\mathbf{Y}) = \mathbf{H}\mathbf{X}\beta = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}\beta.$$

Also,

$$\text{Var}\left(\hat{\mathbf{Y}}\right) = \text{Var}\left(\mathbf{HY}\right) = \mathbf{H}\text{Var}\left(\mathbf{Y}\right)\mathbf{H}^{\top} = \mathbf{H}\left(\sigma^2\mathbf{I}\right)\mathbf{H}^{\top} = \sigma^2\mathbf{HH}^{\top} = \sigma^2\mathbf{HH} = \sigma^2\mathbf{H}.$$

The residuals are given by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}$. We now obtain the expectation and variances of the error values before signing off.

$$E\left(\mathbf{e}\right) = E\left[\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}\right] = \left(\mathbf{I} - \mathbf{H}\right)E\left(\mathbf{Y}\right) = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{X}\beta = \mathbf{X}\beta - \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\beta = \mathbf{X}\beta - \mathbf{X}\beta = 0.$$

Finally, we have,

$$\text{Var}\left(\mathbf{e}\right) = \text{Var}\left(\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}\right) = \left(\mathbf{I} - \mathbf{H}\right)\text{Var}\left(\mathbf{Y}\right)\left(\mathbf{I} - \mathbf{H}\right)^{\top}$$

$$= \left(\mathbf{I} - \mathbf{H}\right)\left(\sigma^2\mathbf{I}\right)\left(\mathbf{I} - \mathbf{H}\right)^{\top} = \sigma^2\left(\mathbf{I} - \mathbf{H}\right)\left(\mathbf{I} - \mathbf{H}\right)^{\top} = \sigma^2\left(\mathbf{I} - \mathbf{H} - \mathbf{H}^{\top} + \mathbf{HH}^{\top}\right) = \sigma^2\left(\mathbf{I} - \mathbf{H}\right).$$

Notes:

1. Both $\mathbf{e}$ and $\mathbf{Y}$ are "multivariate normal."
2. The variance of $\mathbf{e}$ is different from the variance of ϵ. This also shows that they are different quantities.
3. $\text{Var}\left(\mathbf{Y}\right) = \text{Var}\left(\hat{\mathbf{Y}}\right) + \text{Var}\left(\mathbf{e}\right)$.

4.7 Appendix 3: Expectation and Variance in the Multivariate Case (Compilation of Results for Reference)

1. If $\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T$ is a vector of any n random variables, then for a $n \times n$ matrix A, $E\left(A\mathbf{X}\right) = AE(\mathbf{X})$, in the sense that $E\left(\mathbf{X}\right) = \left(E\left(X_1\right), E\left(X_2\right), \ldots, E\left(X_n\right)\right)^{\top}$.
2. The above could be generalised to:

   If $\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T$ is a vector of any n random variables, then for a $k \times n$ matrix A, $E\left(A\mathbf{X}\right) = AE(\mathbf{X})$, in the sense that $E\left(\mathbf{X}\right) = \left(E\left(X_1\right), E\left(X_2\right), \ldots, E\left(X_n\right)\right)^{\top}$, for any integer k.
3. Also, $E\left(c\mathbf{X}\right) = cE\left(\mathbf{X}\right)$, for any constant c, in the sense that $\left(E\left(cX_1\right), E\left(cX_2\right), \ldots, E\left(cX_n\right)\right)^{\top} = c\left(E\left(X_1\right), E\left(X_2\right), \ldots, E\left(X_n\right)\right)^{\top}$.
4. For $\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T$,

$$\text{Var}\left(X\right) = \Sigma = \begin{bmatrix} \sigma_{X_1X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_nX_1} & \sigma_{X_nX_2} & \cdots & \sigma_{X_nX_n} \end{bmatrix}.$$

Further, the following properties are satisfied:

(a) The variances are put in the diagonal, and the covariances are the off-diagonal elements.

(b) Σ is symmetric.

(c) Σ is non-singular.

(d) Σ is positive semi-definite or non-negative definite, i.e. all principal minors are non-negative. If the random variables are non-constant, then the covariance matrix is positive-definite.

(e) If the random variables are independent, then the off-diagonal elements (the covariances) are zero, i.e. Σ is diagonal. Moreover, if the variables are i.i.d. with finite variance, then they have 0 covariance and same variance, so the covariance matrix is diagonal with all the diagonal elements equal to each other.

(f) For a linear transformation $A\mathbf{X} + b$, $\text{Var}\,(A\mathbf{X} + b) = A\Sigma A^T$, where A is $k \times n$ for some integer k and b is a $n \times 1$ vector.

(g) There exists a non-singular positive definite symmetric matrix A such that $A \times A = \Sigma$. The matrix A is known as the square root of the variance covariance matrix and denoted $\Sigma^{1/2}$.