Statistical Modelling I

Kexing Ying

January 11, 2021

Contents

1 Introduction

In this module, we will consider and analyse the relationship between measurements through the use of statistical models. This is realised in several ways including quantifying distributions, comparing distributions and predicting observations. We shall study these methods through deriving, evaluating and applying estimators, confidence intervals and hypothesis tests based, first on parametric models, and later based on the theory of linear models.

Definition 1.1 (Statistical Model). A statistical model is a specification of the distribution of Y up to an unknown parameter θ.

Definition 1.2 (Parameter Space). Given a statistical model Y up to some parameter θ, the set Θ of all possible parameter values is called the parameter space.

In this module we will assume $\Theta \subseteq \mathbb{R}^p$ for some $p \in \mathbb{N}$ so that we consider *parametric models*. A *semiparametric model* is a statistical model which parameters belong to a more general space, e.g. functions spaces.

As with last years **Probability and Statistics**, we will denote the data $\mathbf{y} = (y_1, \cdots, y_n) \in \mathbb{R}^n$ as a vector and $\mathbf{Y} = (Y_1, \cdots, Y_n)$ a random vector. In this case, the statistical model specifies the joint distribution of $Y_1, \cdots, Y_n$ up to some unknown parameter θ. If $Y, \cdots, Y_n$ are independent and identically distributed (iid.), then we call it a *random sample*.

Furthermore, in some situations, the random vector $\mathbf{Y} = (Y_1, \cdots, Y_n)$ might be dependent on random or nonrandom values $x_1, \cdots, x_n$. The x_i's are an example of covariates. An example of this could be that, in a clinical trial, some patients are given a treatment while others received a placebo. If we would like to model the outcome of the i-th patient by their survival time Y_i, it is clear that as covariate for the i-th patient, we may use the indicator function for whether or not i received the treatment as a covariate.

As another example, say we would like to ask whether or not taller people have a higher income. To answer this question we might create a statistical model in which Y_i is the income, x_i is the height and

$$Y_i = \beta_0 + x_i \beta_1 + \epsilon_i$$

for $i = 1, \cdots, n$, $\epsilon \sim N(0, \sigma^2)$ iid. and $\theta = (\beta_0, \beta_1, \sigma^2)$, $\Theta = \mathbb{R}^2 \times [0, \infty)$.

Having formulated a model, we can draw inferences from the sample. By estimating the unknown parameters we attempts to "fit the model". Through this, we receive a model that can provide us with point estimates, or better yet, tools that can help us make decisions through some combination of hypothesis tests and confidence intervals.

However, with all statistical models, we have to accept that it will not perfectly reflect reality. But, that is not the point of statistical models anyway. Statistical models are meant to be useful and in general we would like a model to

- agree with the observed data reasonably;
- be relatively simple;
- interpretable, e.g. parameters have a physical interpretation.

With these aims in mind, we might conduct sensitivity analysis in which we discard models that are not adequate for the data through a iterative process.

2 Point Estimation

From the introduction, we seen that during the process of fitting the model, we need to estimate θ in the statistical model, and furthermore, during the inference process, we need to point estimate, interval estimate or hypothesis test to address our question. We recall from first year that this can be achieved through several methods and we shall quickly review them here.

2.1 Review

We recall the following definitions.

Definition 2.1 (Realisation, Statistic, Estimate, Estimator)**.**

- Data $y_1, \cdots, y_n$ is called a realisation of $Y_1, \cdots, Y_n$.
- A function t of observable random variables is called a statistic.
- An estimate of θ is $t(y_1, \cdots, y_n)$.
- An estimator of θ is $T = t(Y_1, \cdots, Y_n)$.

Example 1. Let $Y_1, \cdots, Y_n \sim N(\mu, 1)$ iid. for some unknown $\mu \in \mathbb{R}$. There are many methods for estimating μ.

- the sample mean $\hat{\mu} = \frac{1}{n} \sum y_i$;
- the sample median;
- the k-trimmed mean where we discard the highest and lowest k observed y_i before computing the mean;
- $\cdots$

For the sample mean estimate, the corresponding estimator is $T = \bar{Y} = \frac{1}{n} \sum Y_i$.

As we can see from the example, there are many possible estimations for the same parameter. To justify the use of a specific estimator, one might use a frequentist's perspective and generate many data and tabulate the results of each estimator. Through this process, one can justify a particular estimator through observed data.

As estimators are random variables, we can formalise this idea by considering properties of its sampling distribution (that is the distribution of the estimator), e.g.

$$\mathbb{P}_\theta(T \in \mathcal{A}), \quad E_\theta(T), \quad \mathrm{Var}_\theta(T), \cdots$$

We saw this idea last year in the form of *bias* and *mean square error*. We recall the definitions here.

Definition 2.2 (Bias)**.** Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the bias of T is

$$\mathrm{bias}_\theta(T) = E_\theta(T) - \theta.$$

If $\mathrm{bias}_\theta(T) = 0$ for all $\theta \in \Theta$, then we say T is unbiased for θ.

If the parameter space is higher dimensional, say $\Theta \subseteq \mathbb{R}^k$, we may be instead be interested in the value of $g(\theta)$ for some $g : \Theta \to \mathbb{R}$. Then, we can naturally extend the definition of bias to this by

$$\text{bias}_\theta(T) = E_\theta(T) - g(\theta).$$

Example 2. Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ iid. $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Then, say if we are in μ, we may define $g : \Theta \to \mathbb{R} : (\mu, \sigma^2) \mapsto \mu$.

Definition 2.3 (Mean Square Error). Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the mean square error of T is

$$\text{MSE}_\theta(T) = E_\theta[(T - \theta)^2].$$

In addition to this, we have the standard error of a estimator

Definition 2.4 (Standard Error). Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then the standard error of T is

$$\text{SE}_\theta(T) = \sqrt{\text{Var}_\theta(T)}.$$

From last year, we saw the following proposition.

Proposition 1. Let T be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. Then

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + (\text{bias}_\theta(T))^2.$$

If we restrict out estimators to be unbiased, often times, we find that the remaining possible estimators well-behaved and we can often find the best estimators by minimising the MSE. However, a biased estimator might have a small MSE than an unbiased estimator (recall sample variance), and it is not necessarily true that such an estimator even exists.

2.2 Cramér-Rao Lower Bound

As the mean square error provide us with a method of quantifying how good an estimator is, we are motivated by minimising the mean square error for a family of estimators. That is, if $\theta \in \Theta$ is a parameter, then is there an estimator T of θ such that for all estimators of θ, S,

$$\text{MSE}_\theta(T) \leq \text{MSE}_\theta(S).$$

Unfortunately, the answer to this question is in general, no, however, for unbiased estimators, the answer is often yes. Indeed, if T is an unbiased estimator, then,

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) = \text{bias}_\theta(T)^2 = \text{Var}_\theta(T),$$

so it suffices to minimise the variance.

Theorem 1 (Cramér-Rao Lower Bound). Suppose $T = T(X)$ is an unbiased estimator for $\theta \in \Theta \subseteq \mathbb{R}$ based on $X = (X_1, \cdots, X_n)$ with joint pdf $f_\theta(x)$. Then under mild regularity conditions (which is elaborated on in the proof below),

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

where

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

and we call $I(\theta)$ the *Fisher information* of the sample.

By computing, we find the Fisher information to equal the following.

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

Indeed,

$$\begin{aligned}
E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right] &= E_\theta \left[\frac{\partial}{\partial \theta} \frac{f_\theta'(X)}{f_\theta(X)} \right] \\
&= E_\theta \left[-\frac{f_\theta'(X)}{f_\theta^2(X)} f_\theta'(X) + \frac{f_\theta''(X)}{f_\theta(X)} \right] \\
&= E_\theta \left[-\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] + E_\theta \left[\frac{f_\theta''(X)}{f_\theta(X)} \right].
\end{aligned}$$

So, the result follows as,

$$\begin{aligned}
E_\theta \left[\frac{f_\theta''(X)}{f_\theta(X)} \right] &= \int_{x \in A} \frac{f_\theta''(x)}{f_\theta(x)} f_\theta(x) \mathrm{d}x \\
&= \int_{x \in A} f_\theta''(x) \mathrm{d}x = \frac{\partial^2}{\partial \theta^2} \int_{x \in A} f_\theta(x) \mathrm{d}x = 0,
\end{aligned}$$

where we denoted A as the support of f_θ. This is a useful identity whenever the second derivative is easy to compute.

Corollary 1.1. Suppose $X_1, \cdots, X_n$ is a random sample. Then, is $f_\theta^{(1)}$ is the pdf of single observation, then

$$I_f(\theta) = n I_{f_\theta^{(1)}}(\theta).$$

Proof. Since a random sample is iid. $f_\theta(x) = \prod f_\theta^{(1)}$ and so

$$I_f(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right] = \sum_{i=1}^n -E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta^{(1)}(X_i) \right) = n I_{f_\theta^{(1)}}(\theta).$$

$\square$

From this, we can conclude that the Fisher information is proportional to the sample size.

Example 3. Let us find the Fisher information for the random sample $X_1, \cdots, X_n \sim \mathrm{Bern}(\theta)$.

By the above corollary, we have $I_f(\theta) = n I_{f_\theta^{(1)}}(\theta)$. So, since the pmf of a Bernoulli random variable is $f_\theta^{(1)}(x) = \theta^x (1-\theta)^{1-x}$, we have

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

hence,

$$I_{f_\theta^{(1)}}(\theta) = E\left[\left(\frac{x-\theta}{\theta(1-\theta)}\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}\operatorname{Var}(X) = \frac{1}{\theta(1-\theta)}.$$

Thus, the Fisher information of the random sample is just $I_f(\theta) = n/\theta(1-\theta)$.

With the Fisher information, we can apply the Cramér-Rao lower bound theorem allowing us to conclude that an unbiased estimator T for θ has variance $\operatorname{Var}(T) \geq \theta(1-\theta)/n = \operatorname{Var}(\bar{X})$. This allows us to conclude that the sample mean $\bar{X}$ minimises the mean square error among unbiased estimators for θ.

Let us now prove the Cramér-Rao lower bound theorem.

Proof. (Cramér-Rao lower bound theorem). Let us first specify the regularity conditions for the Cramér-Rao lower bound theorem.

- Assume that the set $A := \operatorname{supp} f_\theta = \{x \in \mathbb{R}^n \mid f_\theta(x) > 0\}$ is independent of θ.

- Θ is an open interval in $\mathbb{R}$.

- For all $\theta \in \Theta$ there exists $\frac{\partial f_\theta}{\partial\theta}$.

- Differentiation and integration commutes (for the specific cases where it is used).

As we saw last year, the space of random variables form an inner product space with the inner product

$$\langle X, Y \rangle = E[XY],$$

and so, the Cauchy-Schwarz inequality applies. That is for all random variables X, Y

$$[E(XY)]^2 \leq E(X^2)E(Y^2).$$

So, we have

$$\operatorname{Var}_\theta(T)I_f(\theta) = E_\theta[(T - E_\theta T)^2]E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right]$$

$$\geq \left(E_\theta\left[(T - E_\theta(T))\frac{\partial}{\partial\theta}\log f_\theta(X)\right]\right)^2.$$

Thus, it suffices to show that the expectation on the right hand side evaluates to 1.

$$E_\theta\left[(T - E_\theta(T))\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = E_\theta\left[(T - E_\theta(T))\frac{\frac{\partial}{\partial\theta}f_\theta(X)}{f_\theta(X)}\right]$$

$$= \int_{x\in A}(T(x) - E_\theta(T))\frac{\frac{\partial}{\partial\theta}f_\theta(x)}{f_\theta(X)}f_\theta(x)\mathrm{d}x$$

$$= \int_{x\in A}T(x)\frac{\partial}{\partial\theta}f_\theta(x)\mathrm{d}x - \int_{x\in A}E_\theta(T)\frac{\partial}{\partial\theta}f_\theta(x)\mathrm{d}x$$

$$= \frac{\partial}{\partial\theta}\int_{x\in A}T(x)f_\theta(x)\mathrm{d}x - E_\theta(T)\frac{\partial}{\partial\theta}\int_{x\in A}f_\theta(x)\mathrm{d}x$$

$$= \frac{\partial}{\partial\theta}E_\theta(T) - 0 = \frac{\partial}{\partial\theta}\theta = 1.$$

$\square$

2.3 Asymptotic Properties of Estimators

While the Cramér-Rao lower bound theorem provides us with a lower bound for the variance for non-biased estimators, as we have previously seen, it is not always true that there exists an unbiased estimator. So, rather than giving up, we instead study the estimators as the sample size becomes large.

As we have seen, evaluating an estimator $T = T(X_1, \cdots, X_n)$ of θ depends on its *sampling distribution*. From the sampling distribution, one can possibly find properties about the estimator such as is bias, mean square error and so on. However, it is not necessarily true that an estimator has a closed form. Indeed, often times, the estimator is defined as a solution to some equation.

To simplify this, one often consider $T_n = T_n(X_1, \cdots, X_n)$ as a sequence of random variables indexed by $n \in \mathbb{N}$ and consider the stochastic convergence of the variables in question. We recall from last term's probability course, there are three different notions of convergence for random variables,

- convergence in probability;

- convergence almost surely (almost everywhere);

- convergence in distribution.

Let us quickly define them here again.

Definition 2.5 (Convergence in Probability)**.** Let $(X_n)_{n=1}^\infty$ be a sequence of random variables. Then, (X_n) converges to the random variable X in probability if for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Definition 2.6 (Convergence Almost Surely)**.** Let $(X_n)_{n=1}^\infty$ be a sequence of random variables. Then, (X_n) converges to the random variable X almost surely if

$$\mathbb{P}\left(\lim_{n \to \infty} X_n = X \right) = 1,$$

where $\lim_{n \to \infty} X_n = X$ is denoting the event

$$\{\omega \in \Omega \mid X_n(\omega) \to X(\omega)\}.$$

With other words, $X_n \to X$ almost surely, if the set of points ω such that $X_n(\omega)$ does not converge to $X(\omega)$ has measure 0.

Definition 2.7. Let $(X_n)_{n=1}^\infty$ be a sequence of random variables. Then, (X_n) converges to the random variable X with cdf F_X in distribution if

$$\lim_{n \to \infty} \mathbb{P}(X_n \leq x) = F_X(x),$$

for all x at which F_X is continuous.

We also recall the following chain of implications,

$$X_n \to_{\mathrm{as}} X \implies X_n \to_{\mathrm{p}} X \implies X_n \to_{\mathrm{d}} X.$$

If $X = c$ is a constant, then

$$X_n \to_{\mathrm{p}} X \iff X_n \to_{\mathrm{d}} X.$$

We apply this notion onto estimators.

Definition 2.8 (Consistency). A sequence of estimators $(T_n)_{n=1}^{\infty}$ for $g(\theta)$ is called (weakly) consistent if for all $\theta \in \Theta$, $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|T_n - g(\theta)| > \epsilon) = 0.$$

While it is possible to prove consistency for certain estimators, it is often a non-trivial task. Instead, we often look at whether a sequence of estimators are asymptotically unbiased and prove the a special class of these estimators are consistent.

Definition 2.9 (Asymptotically Unbiased Estimators). A sequence of estimators $(T_n)_{n=1}^{\infty}$ for $g(\theta)$ is called asymptotically unbiased if for all $\theta \in \Theta$,

$$E_\theta(T_n) \to g(\theta).$$

We see that $E_\theta(T_n)$ is simply a value and this is simply the convergence for real sequences.

Before moving on to prove results about estimators, let us quickly recall the Markov inequality.

Proposition 2. Let X be a random variable with $X(\Omega) \subseteq [0, \infty)$, then for all $a \in \mathbb{R}$,

$$\mathbb{P}(|X| \geq a) \leq \frac{E(|X|)}{a}.$$

Proof. See first year notes. $\qquad\square$

Lemma 2.1 (MSE Consistency). Let $(T_n)_{n=1}^{\infty}$ be asymptotically unbiased for $g(\theta)$ for all $\theta \in \Theta$. Then, if $\mathrm{Var}_\theta(T_n) \to 0$ as $n \to \infty$, T_n is consistent for $g(\theta)$.

Proof. Let $\epsilon > 0$, then, by Markov's inequality,

$$\mathbb{P}_\theta(|T_n - g(\theta)| \geq \epsilon) = \mathbb{P}_\theta((T_n - g(\theta))^2 \geq \epsilon^2) \leq \frac{1}{\epsilon^2} E_\theta(T_n - g(\theta))^2$$

$$= \frac{\mathrm{MSE}_\theta(T_n)}{\epsilon^2} = \frac{1}{\epsilon}(\mathrm{Var}_\theta(T_n) + (E_\theta(T_n) - g(\theta))^2).$$

Since the right hand side tends to 0 as $n \to \infty$, so is the left hand side. $\qquad\square$

Thus, to show that a sequence of estimators is consistent, it suffices to show that it is asymptotically unbiased and its variance tends to 0.

However, while consistency is a nice property for a sequence of estimators to have, it is a very minimal requirement. So, in order to derive hypothesis tests and confidence intervals,

we also need the sampling distribution of T_n. As we have seen previously, the sample mean estimators T_n for a normal distribution $N(\theta, 1)$ has distribution $T_n \sim N(\theta, 1/n)$, and so, by centring and scaling, we have

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

for all $n \geq 1$. This means we can work with the CDF of T_n allowing us to easily analyse the behaviours of these estimators. However, this is in general not the case and in most cases, we cannot derive easily the distributions of the estimators. Nonetheless, often, we may approximate their distribution with a normal distribution.

Definition 2.10 (Asymptotically Normal). A sequence of estimators T_n for $\theta \in \mathbb{R}$ is asymptotically normal if, for some $\sigma^2(\theta)$,

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

in distribution.

From last term, we recall the central limit theorem (CLT).

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

Proof. See the *probability for statistics* course. $\qquad\square$

The central limit theorem allows us to conclude that a large class of estimators are asymptotically normal. Indeed, sample means and estimators which can be written as a combination of sample means under weak conditions are certainly asymptotically normal. However, we would also consider other estimators.

Lemma 2.2 (Slutsky's lemma). Let X_n, X and Y_n be random variables (or random vectors). If $X_n \to X$ in distribution and $Y_n \to c$ in probability for some constant c, then

- $X_n + Y_n \to X + c$ in distribution;

- $Y_n X_n \to cX$ in distribution;

- $Y_n^{-1} X_n \to c^{-1} X$ in distribution if $c \neq 0$.

Proof. See the *probability for statistics* course. $\qquad\square$

Another useful result for determining whether or not a sequence of estimators are asymptotically normal is the δ-method. The δ-method allows us to consider whether or not the transformation of a asymptotically normal estimator remains asymptotically normal.

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

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

and $g : \Theta \subseteq \mathbb{R} \to \mathbb{R}$ is a differentiable function with $g'(\theta) \neq 0$. Then

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

Proof. Since g is differentiable,

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + o((T_n - \theta)^2),$$

where R is the remainder. So

$$\sqrt{n}(g(T_n) - g(\theta)) = g'\sqrt{n}(\theta)(T_n - \theta) + o((T_n - \theta)^2).$$

Thus, assuming the remainder is negligible, we have

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

$\square$

Lastly, a useful result we will often use (perhaps implicitly) is the continuous mapping theorem. Alike sequential continuity for metric spaces, the continuous mapping theorem will allow us to preserve stochastic convergence under continuous mappings.

Theorem 4 (Continuous Mapping Theorem)**.** Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. Then if $X_n \to X$, then $g(X_n) \to g(X)$ for all three notions of convergence, i.e. convergence in distribution, in probability and almost surely.

2.4 Maximum Likelihood Estimation

We recall from first year the maximum likelihood estimator, that is the estimator that maximises the probability of observing the given realisations.

Definition 2.11 (Likelihood Function)**.** Given the realisation $\mathbf{x}$ of the random object $\mathbf{X}$, the likelihood function for θ is

$$L(\theta) = L(\theta \mid \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x} \mid \theta).$$

Definition 2.12 (Maximum Likelihood Estimator)**.** The maximum likelihood estimator of $\theta \in \Theta^n$ is an estimator $\hat{\theta}$ such that

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

where L is the likelihood function.

The maximum likelihood estimator is often well defined. However, it is possible to construct situations in which the MLE does not exist or is not unique. We also recall that given a strictly increasing function f, the maximum likelihood estimator can also be obtained by maximising $f \circ L$. This is most commonly seen in the log-likelihood function where we maximise $\log L$.

Maximum likelihood estimators has some nice properties. In short, maximum likelihood estimators are functionally invariant, consistent and asymptotically normal.

Proposition 3. If g is a bijective function and if $\hat{\theta}$ is a MLE of θ, then $\hat{\phi} = g(\hat{\theta})$ is a MLE of $\phi = g(\theta)$.

Proof. Let us denote $\tilde{L}$ for the likelihood function of ϕ, then $\tilde{L} = L \circ g^{-1}$ and so,

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

$\square$

Suppose we now relax the bijective condition on g. If g is not surjective, then there exists $\phi \in \psi$ such that $\phi \notin g(\Theta)$, and so, for these values no model is defined. If this is the case it does not make sense to speak of the likelihood of these parameters and so, we define their likelihood to be 0. With that, we see that the original proposition remains true.

On the other hand, if g is not injective, then ϕ does not uniquely identify θ and so, there might exists multiply θ such that $g(\theta) = \phi$. However, by defining the induced likelihood for all θ,

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

we see that the invariance over functions is retained.

Proposition 4. Let $X_1, \cdots$ be iid. observations with pdf $f_\theta(x)$ where $\theta \in \Theta$ and Θ is an open interval. Furthermore, let $\theta_0 \in \Theta$ be some parameter. Then under regularity conditions,

- there exists a consistent sequence $(\hat{\theta}_n)_{n=1}^\infty$ of maximum likelihood estimators;

- if $(\hat{\theta}_n)_{n=1}^\infty$ is a consistent sequence of MLEs, then

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

  where $I_f(\theta)$ is the Fisher information of a sample of size 1.

We note that this proposition requires the Fisher information of a distribution which is often not known in practical situations. So, to use this result we need to estimate $I_f(\theta_0)$. In general, this can be approximated by

- $I_f(\hat{\theta})$;

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

- $-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log(f(x_i \mid \theta)) \mid_{\theta=\hat{\theta}}$.

Proof. (Sketch of the existence of consistent MLEs). Let us denote $L(\theta) := \prod_{i=1}^n f_\theta(X_i)$ and $S_n(\theta) = \frac{1}{n} \log L(\theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i)$. Since log is strictly increasing, we see that $\hat{\theta}$ maximises $L(\theta)$ if and only if it maximises $S(\theta)$. Then, by the weak law of large numbers, given iid. $Z_1, \cdots, Z_n$ where $Z_i = \log f_\theta(X_i)$, $S_n(\theta) \to E_{\theta_0}(Z_1) = E_{\theta_0}(\log f_{\theta_0}(X_i))$ in probability. So, S_n is a consistent estimator for $E_{\theta_0}(\log f_{\theta_0}(X_i))$.

Let us now define $R(\theta) := E_{\theta_0}(\log f_\theta(X_1))$ and I claim that θ_0 maximises R. Indeed, by considering $z - 1 \geq \log z$ for all $z \in \mathbb{R}^+$, we have for all θ,

$$R(\theta) - R(\theta_0) = E_{\theta_0} \left[\log \frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} \right] \leq E_{\theta_0} \left[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} - 1 \right] = \int \left[\frac{f_\theta(x)}{f_{\theta_0}(x)} - 1 \right] f_{\theta_0}(x) \mathrm{d}x.$$

But this is simply,

$$\int f_\theta(x) - f_{\theta_0}(x) \mathrm{d}x = 1 - 1 = 0,$$

11

and hence, $R(\theta) \leq R(\theta_0)$ for all $\theta \in \Theta$.

With this in mind, we have $S_n(\theta) \to R(\theta)$ pointwise and $S_n(\hat{\theta}) \to R(\theta_0)$ in probability. Then, by using analysis techniques, we may show $\hat{\theta} \to \theta_0$ in probability, completing the proof of the first claim. $\qquad\square$

This particular sketch of the proof (which classically Wald used) requires that the map $\theta \to R(\theta)$ to be continuous for some compact set $K \subseteq \Theta$ in which for all $\epsilon > 0$, $\mathbb{P}(|\hat{\theta} - \theta_0| > \epsilon, \hat{\theta} \in K) \to 0$. There is a modern approach in which one shows that

$$\sup_{\theta \in \Theta} |S_n(\theta) - R(\theta)| \to 0$$

in probability. This approach relaxes the condition some what and will be examined in the third year course *Statistical Theory*.

3 Confidence Regions & Hypothesis Testing

So far, we have being considering point estimators for single values. This does not reflect any uncertainty. Indeed, as this estimator is simply resulted from a random sample, it does not tell us how variable this estimate would be if we drew another sample. To account for this, we may, instead of estimating a single value, we provide an interval of values that contains the true parameter with a certain probability.

As an example, let us recall an example from first years statistics. Given a random sample $Y_1, \cdots, Y_n \sim N(\mu, \sigma_0^2)$ with σ_0^2 known, we would like to find the confidence intervale I such that $\mathbb{P}(\mu \in I) = 1 - \alpha$ for some $\alpha > 0$, e.g. $\alpha = 0.05$. By using the sample mean, we have $\bar{Y} = \frac{1}{n} \sum Y_i \sim N(\mu, \sigma_0^2/n)$. So, by standardising, we have

$$\frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1),$$

and so, $c_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ where $c_{\alpha/2}$ is the value such that,

$$1 - \alpha = \mathbb{P}\left(-c_{\alpha/2} < \frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1) < c_{\alpha/2}\right).$$

Hence, the $1 - \alpha$ confidence interval of μ is a realisation of the **random** interval

$$I(\mathbf{Y}) = (\bar{Y} - c_{\alpha/2}\sigma_0/\sqrt{n}, \bar{Y} + c_{\alpha/2}\sigma_0/\sqrt{n}).$$

We note that, this does not mean that given a realisation of the random interval I, $\mathbb{P}(\mu \in I) = 1 - \alpha$. Indeed, if I is realised, either $\mathbb{P}(\mu \in I) = 0$ or $\mathbb{P}(\mu \in I) = 1$.

Definition 3.1 ($1 - \alpha$ Confidence Interval)**.** A $1 - \alpha$ confidence interval for $\theta \in \Theta$ is a random interval I_θ that contains θ with probability $\geq 1 - \alpha$, that is,

$$\mathbb{P}_\theta(\theta \in I) \geq 1 - \alpha.$$

A confidence interval can be any types of interval including unbounded ones. Indeed, if the confidence interval is unbounded, we say the confidence interval is a one-sided confidence interval. An application of such an confidence interval could be that we would like to measure the pollutant in drinking water with a maximum percentage. Then, given a random sample of measurements $Y_1, \cdots, Y_n$, we would like to find a confidence interval for such that

$$\mathbb{P}(\theta \leq h(Y)) = 1 - \alpha.$$

So, the confidence interval in this case would be in the form $(-\infty, h(y)]$.

3.1 Construction of Confidence Intervals

Definition 3.2 (Pivotal Quantity)**.** A pivotal quantity for θ is a function $t(Y, \theta)$ of the data θ (and **not** any over parameters).

With the pivotal quantity for θ, $t(Y, \theta)$, we can find constants a_1, a_2 such that

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

since we know the distribution of t. In many cases, we may rearrange the terms to give

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

where $[h_1(Y), h_2(Y)]$ is a random interval. This is a $1 - \alpha$ confidence interval for θ.

As an example of a pivotal quantity, suppose we would like to construct an confidence interval for the random sample $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Then, we may define the pivotal quantity $(Y, \mu) \mapsto \frac{\bar{Y} - \mu}{S/\sqrt{n}}$ where S is the sample standard deviation. This pivotal quantity follows the Student-t distribution with $n - 1$ degrees of freedom allowing us to construct a confidence interval according the the method above.

On the other hand, if we would like to construct an confidence interval for σ^2, we can use the pivotal quantity $(Y, \sigma^2) \mapsto \frac{1}{\sigma^2} \sum (Y_i - \bar{Y})^2$ which has χ^2-distribution with $n - 1$ degrees of freedom.

However, we see that these constructions are rather specialised to normal distributions and without justification, cannot be applied to other distributions. Nonetheless, as we have discussed asymptotic behaviours of estimators, in which many estimators are asymptotically normal, we can use this fact to extend our theory of confidence intervals.

Definition 3.3 (Asymptotic Confidence Interval)**.** A sequence of random intervals I_n is called an asymptotic $1 - \alpha$ confidence interval for θ if

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

Suppose $\hat{\sigma}_n$ is consistent for $\sigma(\theta)$ and thus, $\hat{\sigma}_n \to \sigma(\theta)$ in probability for all θ. By Slutsky's lemma, we have

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

in distribution. Using the left hand side as the pivotal quantity leads us to the approximate *confidence limits*

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

where $\Phi(c_{\alpha/2}) = 1 - \alpha/2$. In general, the easiest choice for $\hat{\sigma}_n$ is simply $\mathrm{SE}(T_n)$.

Lastly, we might be interested in constructing a confidence region for more than one parameters.

Definition 3.4 (Simultaneous Confidence Interval)**.** Suppose $\theta = (\theta_1, \cdots, \theta_k)^T \in \Theta \subseteq \mathbb{R}^k$ and we have $(L_i(Y), U_i(Y))$ such that for all θ,

$$\mathbb{P}(L_i(Y) < \theta_i < U_i(Y) \mid i = 1, \cdots, k) \geq 1 - \alpha.$$

Then, $(L_i(y), U_i(y))$ is a $1 - \alpha$ simultaneous confidence interval for $\theta_1, \cdots, \theta_k$.

Theorem 5 (Bonferroni Correction for Simultaneous Confidence Intervals)**.** Suppose $[L_i, U_i]$ is a $1 - \alpha/k$ confidence interval for θ_i. Then, $\prod [L_i, U_i]$ is a $1 - \alpha$ simultaneous confidence interval for $\theta = (\theta_1, \cdots, \theta_k)^T$.

Proof.

$$\mathbb{P}(\theta_i \in [L_i, U_i] \mid i = 1, \cdots, k) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{k}\{\theta_i \notin [L_i, U_i]\}\right) \geq 1 - \sum_{i=1}^{k}\mathbb{P}(\theta_i \notin [L_i, U_i]) \geq 1 - \alpha.$$

$\square$

We note that the Bonferroni corrections are conservative and it is very possible that the resulting simultaneous confidence interval has a higher coverage probability that that is suggested by the Bonferroni correction. We see this in the example where we attempt to find the simultaneous confidence interval of two *independent* random samples $X_1, \cdots, X_n \sim N(\mu, 1)$ and $Y_1, \cdots, Y_n \sim N(\theta, 1)$. Then by the usual method, we find I, J the $1 - \alpha$-confidence intervals for μ and θ respectively. By the Bonferroni correction, $I \times J$ is a $1 - 2\alpha$-confidence region for (μ, θ) while in actuality,

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

Choosing $\alpha = 0.1$, we see that Bonferroni guarantees the coverage probability to be above 0.8 while the actual probability is $0.9^2 = 0.81$.

3.2 Hypothesis Testing

Definition 3.5 (Null and Alternative Hypothesis)**.** Given a model f_θ where $\theta \in \Theta \subseteq \mathbb{R}^d$, the null-hypothesis H_0 and the alternative hypothesis H_1 are propositions that $\theta \in \Theta_0$ and $\theta \in \Theta_1$ respectively for some Θ_0, Θ_1 a partition of Θ such that $\Theta_0 \cap \Theta_1 = \varnothing$ and $\Theta_0 \cup \Theta_1 = \Theta$.

The goal of an hypothesis test is to determine whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, or with other words, whether to accept H_0 or reject H_0 and hence accept H_1. This is normally achieved through the observation of a particular subset of the sample space and we call this sample space for which H_0 is rejected the rejection region (or critical region).

In some literature, we might find some authors reframe from using the word *accept* (such as we were told in year one). In practice however, since we are acting based on the result of these tests, it makes some practical meaning to say we accept the null-hypothesis H_0 or we accept the alternative hypothesis H_1.

As the accuracy of the hypothesis tests is arbitrary, it is possible to make errors. The below table demonstrates the two types of errors.

	$\theta \in \Theta_0 (H_0)$	$\theta \in \Theta_1 (H_1)$
$\neg$ reject H_0	✓	Type II error
reject H_0	Type I error	✓

Definition 3.6 (Level of a Test)**.** A hypothesis test is of level α for $0 < \alpha < 1$ if

$$\mathbb{P}_\theta(\text{reject } H_0) \leq \alpha$$

for all $\theta \in \Theta$.

Usually we choose $\alpha << 1$ with common values being 0.05 and 0.01. However, it is not clear whether or not these values are optimal for general experiments and often times, α is chosen to be much smaller, e.g. $\alpha \sim 10^{-6}$.

Definition 3.7 (Power). Let Θ be a parameter space and $\Theta_0 \subseteq \Theta$ and $\Theta_1 = \Theta \setminus \Theta_0$ so $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ are null and alternative hypothesis'. Suppose we can constructed some test for this hypothesis, then, the power function is the mapping

$$\beta : \Theta \to [0,1] : \theta \mapsto P_\theta(\text{reject } H_0).$$

Conceptually, if $\theta \in \Theta_0$, we would like $\beta(\theta)$ to be small while if $\theta \in \Theta_1$, we would like $\beta(\theta)$ to be large.

Definition 3.8 (p-Value). The p-value of a particular hypothesis test is

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

That is, if a test is based on the statistic T with rejection for large values of T, then

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

where t is the observed value.

In any case, we reject H_0 if and only if $p \leq \alpha$ and this results a α-level test.

The hypothesis tests are related to confidence intervals in that we can construct a test from any confidence regions.

Let Y be the random observation of the experiment and suppose $A(Y)$ is the $1 - \alpha$ confidence region for the parameter $\theta \in \Theta$, i.e.

$$\mathbb{P}_\theta(\theta \in A(Y)) \geq 1 - \alpha,$$

for all $\theta \in \Theta$. Then, by defining $H_0 : \theta \in \Theta_0$ and $H_0 = \theta \notin \Theta_0$ where Θ_0 is some subset of Θ with level α such that we reject H_0 if $\Theta_0 \cap A(y) = \varnothing$. In this case, we see that

$$\mathbb{P}_\theta(\text{Type I error}) = \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{P}_\theta(\Theta_0 \cap A(Y) = \varnothing) \leq \mathbb{P}_\theta(\theta \notin A(Y)) \leq \alpha.$$

The reverse is also possible – constructing a confidence region from a hypothesis test. Suppose that for all $\theta_0 \in \Theta$, we have a level α test ϕ_{θ_0} for $H_0^{\theta_0} : \theta = \theta_0$ and $H_1^{\theta_0} : \theta \neq \theta_0$ such that

$$\mathbb{P}_{\theta_0}(\phi_{\theta_0} \text{ reject } H_0) \leq \alpha.$$

Then, by defining
$$A := \{\theta_0 \in \Theta \mid \phi_{\theta_0} \text{ does not reject } H_0^{\theta_0}\},$$

we find A to be a $1 - \alpha$ confidence region for θ. Indeed, for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(\theta \in A) = \mathbb{P}_\theta(\phi_\theta \text{ does reject }) = 1 - P_\theta(\phi_\theta \text{ rejects}) \geq 1 - \alpha.$$

Through this method, we may construct a test for multiple parameter test through the use of simultaneous confidence regions. Indeed, if $I \times J$ is a $1 - 2\alpha$ confidence region for (μ, θ), a level 2α test of $H_0 : (\mu, \theta) = (\mu_0, \theta_0)$ against $H_1 : (\mu, \theta) \neq (\mu_0, \theta_0)$ is given by

$$R = \{(\bar{X}, \bar{Y}) \mid (\mu_0, \theta_0) \notin I \times J\},$$

where $E(X_i) = \mu$ and $E(Y_i) = \theta$.

16

3.3   Likelihood Ratio Test

For the maximum likelihood estimator, we choose the parameter with the highest likelihood as our estimator. Using a similar idea, we can create the likelihood ratio test.

Consider the hypothesis $H_0 : \theta \in \Theta_0$ again $H_1 : \theta \notin \Theta_0$. Then, by comparing the likelihood L under H_0 to the unrestricted maximum likelihood, if the unrestricted maximum likelihood is significantly larger than that under H_0, we have evidence for rejecting H_0.

Definition 3.9 (Likelihood Ratio Test Statistic). Suppose we observe the data y. The likelihood ratio test statistic is

$$t(y) = \frac{\sup_{\theta \in \Theta} L(\theta \mid y)}{\sup_{\theta \in \Theta_0} L(\theta \mid y)} = \frac{\text{ml. under } H_0 + H_1}{\text{ml. under } H_0}.$$

We see that if $t(y)$ is "large", then this statistic will indicate suppose for H_1 since it is much more likely that we observed y without assuming H_0. Thus, we reject H_0 when

$$t(y) \geq k,$$

where k is chosen such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(Y) \geq k) \leq \alpha.$$

This choice of k ensures that the hypothesis constructed this way has level α.

In general however, it is not always easy to determine k analytically since in order to do so, we will need to first determine the distribution of the likelihood ratio statistic. Thus, rather than computing k exactly, we can approximate k whenever the sample size is large.

Theorem 6. Under certain regularity conditions (in particular the subspace closure of Θ_0 has a smaller dimension than Θ),

$$2 \log t(Y) \to \chi_r^2$$

in distribution as $n \to \infty$ under H_0, where r is the number of independent restrictions on θ needed to define H_0. Alternatively, r is the number of independent parameters under the full model minus the number of independent parameters under H_0.

Proof. We will prove the univariate case and the multivariate case generalises easily.

Consider $2 \log t(Y) = 2(\log L(\hat{\theta}) - \log L(\theta_0)) = 2(l(\hat{\theta} - l(\theta_0))$. By Taylor expanding, we have

$$l(\theta_0) = l(\hat{\theta}) + l'(\hat{\theta})(\theta_0 - \hat{\theta}) + \frac{1}{2}l''(\tilde{\theta})(\theta_0 - \hat{\theta})^2,$$

for some $\tilde{\theta} \in (\theta_0, \hat{\theta})$. Now, for the maximum likelihood estimator, we have $l'(\hat{\theta}) = 0$ and so,

$$
\begin{aligned}
2(l(\hat{\theta}) - l(\theta_0)) &= -l''(\tilde{\theta})(\theta_0 - \hat{\theta})^2 \\
&= -\frac{1}{n}l''(\tilde{\theta})(\sqrt{n}(\hat{\theta} - \theta_0))^2 \\
&= I(\theta_0)(\sqrt{n}(\hat{\theta} - \theta_0))^2 + (\sqrt{n}(\hat{\theta} - \theta_0))^2\left(-\frac{1}{n}l''(\tilde{\theta}) - I(\theta_0)\right)
\end{aligned}
$$

17

Now, assuming $-\frac{1}{n}l''(\tilde{\theta}) - I(\theta_0) \to 0$ in probability, it suffices to show $I(\theta_0)(\sqrt{n}(\hat{\theta}-\theta_0))^2 \to \chi_1^2$ in distribution. But this is obvious as $\sqrt{n}(\hat{\theta}-\theta_0) \to N(0, I(\theta_0)^{-1})$ in distribution, we have

$$I(\theta_0)(\sqrt{n}(\hat{\theta}-\theta_0))^2 \to N(0,1)^2 = \chi_1^2$$

$\square$

4 General Linear Models

From last year, we saw simple linear regression where we were able to inference in random samples. We shall review this process and extend this method.

4.1 Review

Definition 4.1 (Simple Linear Model)**.** A simple linear model $(Y_i)_{i=1}^n$ is a sequence of random variables such that

$$Y_i = \beta_1 + a_i\beta_2 + \epsilon_i,$$

where

- a_i are the observable constant – "covariate";

- β_1, β_2 unknown parameters;

- $(\epsilon_i)_{i=1}^n$ a sequence of iid. random variables (errors) with $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$;

- $\sigma^2 > 0$ is another unknown parameter;

- ϵ_i is not observable.

We call Y_i the "outcome" or "response" and they are observable random variables.

As β_1, β_2 are unknown, given that we observed a_i and y_i, we need to construct estimators for β_1, β_2. This is done through the least square estimators.

Definition 4.2 (Least Squares Estimators)**.** The least squares estimators $\hat{\beta}_1, \hat{\beta}_2$ of β_1, β_2 are defined as the minimisers of

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

The observables $e_i = y_i - \hat{\beta}_1 - a_i\hat{\beta}_2$ are called the residuals and they are **not** iid. as dependence is introduces throught $\hat{\beta}_1$ and $\hat{\beta}_2$.

We remark that in linear regression models $Y_1, \cdots, Y_n$ are generally not iid. observations. independence however, will still hold if the errors $\epsilon_1, \cdots, \epsilon_n$ are independent. The distribution of Y_i need not to have the same distribution as ϵ_i since it is dependent on a_i.

As linear regression naturally leads to a connection between statistics and linear algebra, we shall also quickly review some useful results from linear algebra, especially matrix properties.

Proposition 5. Let $X \in M_{n,p}$, then $\mathrm{rank}(X^TX) = \mathrm{rank}(X)$.

Proof. By recalling the rank-nullity theorem, we have $p = \mathrm{rank}(X) + \dim \ker(X)$. Similarly, we have $p = \mathrm{rank}(X^TX) + \dim \ker(X^TX)$. So, it suffices to show that $\ker X = \ker X^TX$. Trivially, $\ker X \subseteq \ker X^TX$ so we will show $\ker X^TX \subseteq \ker X$. Let $x \in \ker X^TX$, then by definition $X^TXx = 0$ and so

$$0 = x^T0 = x^TX^TXx = (Xx)^T(Xx) = \|Xx\|^2.$$

So, by positive-definiteness, $Xx = 0$ and hence $x \in \ker X$. $\qquad\qquad\square$

Definition 4.3 (Positive-Definite)**.** A matrix $A \in \mathbb{R}^{n \times n}$ is positive-definite if for all $x \in \mathbb{R}^n \setminus \{0\}$, $x^T A x > 0$. With other words, the quadratic form induced by A is positive-definite.

Lemma 4.1. If $A \in \mathbb{R}^{n \times n}$ is symmetric, then there exists some orthogonal matrix P such that $P^T A P$ is diagonal, i.e. A is congruent to a diagonal matrix by a orthogonal transformation.

Furthermore, if A is positive-definite, then there exists a non-singular matrix Q such that $Q^T A Q = I_n$.

Proof. Standard result. $\qquad\square$

Linear regression models construct an inference based on the realisations of random variables. This can be modelled using random vectors and we shall talk about methods on how to work with random vectors here.

Definition 4.4 (Expectation of Random Vector)**.** Let $X = (X_1, \cdots, X_n)^T$ be a random vector. Then the expectation of X is simply the vector of component expectations, i.e.

$$E(X) = (EX_1, \cdots, EX_n)^T.$$

Similarly, the expectation of matrices is also defined component-wise.

This definition allows us to easily generalise the properties of expectation to random vectors.

Proposition 6. Let X, Y be n-variate random vectors. Then

- $E(X + Y) = EX + EY$;

- for al $a \in \mathbb{R}$, $E(aX) = aE(X)$;

- if A, B are deterministic matrices of suitable dimensions, $E(AX) = A(EX)$ and $E(X^T B) = (EX)^T B$.

Proof. Easy. $\qquad\square$

Definition 4.5 (Covariance of Random Vectors)**.** If X, Y are random vectors (not necessarily with same dimension), then the covariance between X, Y is

$$\mathrm{Cov}(X, Y) = (\mathrm{Cov}(X_i, Y_i))_{i,j}.$$

Furthermore, we define $\mathrm{Cov}(X)$ to be $\mathrm{Cov}(X, X)$.

We also find this definition satisfies out usual properties for covariance.

Proposition 7. If X, Y are random vectors, then

$$\mathrm{Cov}(X, Y) = E[(X - EX)(Y - EY)^T] = E[XY^T] - EX(EY)^T.$$

Furthermore,

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)^T$;

- $\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$;

- $\mathrm{Cov}(AX, BY) = A\mathrm{Cov}(X, Y)B^T$;

- $\mathrm{Cov}(AX) = A\mathrm{Cov}(X)A^T$;

- $\mathrm{Cov}(X)$ is positive-semidefinite and symmetric;

- $\mathrm{Cov}(X, Y) = 0$ if X, Y are independent.

Proof. Easy. $\qquad\square$

4.2 Introduction to Linear Model

Definition 4.6 (General Linear Model)**.** A linear model is a model in the form

$$Y = X\beta + \epsilon$$

where

- Y is an n-dimensional random vector;

- $X \in \mathbb{R}^{n \times p}$ is a known matrix (design matrix);

- $\beta \in \mathbb{R}^p$ is an unknown parameter;

- ϵ is an n-variate random vector (not observable) with $E(\epsilon) = 0$.

We in general assume $n > p$, that is, the sample size is larger than the number of parameter. While in general this is true, "high-dimensional" cases where $p > n$ can occur in situations such as genetic data, finance and electronics. In these cases we may use the ridge regression or LASSO, etc...

Straight away, we have the general linear model is more general than the simple linear regression. Indeed, described as a general model, the simple model becomes

$$Y = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1, \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

We shall introduce some commonly seen assumptions.

Definition 4.7 (Second Order Assumption)**.** The second order assumption is the assumption that the covariance matrix of ϵ is $\mathrm{Cov}(\epsilon) = \mathrm{Cov}(\epsilon_i, \epsilon_j)_{i,j} = \sigma^2 I_n$ for some $\sigma^2 > 0$.

This assumption is equivalent to saying that the errors of two different observation are uncorrelated and the variance of all errors is identical.

Definition 4.8 (Normal Theory Assumption)**.** The normal theory assumption assumes that $\epsilon \sim N(0, \sigma^2 I_n)$ for some $\sigma^2 > 0$. That is the error follows the the multivariate normal distribution with mean 0 and the covariance matrix $\sigma^2 I_n$.

We see that NTA implies SOA and we will commonly use NTA to construct tests and confidence intervals.

Definition 4.9 (Full Rank Assumption). The full rank assumption assumes the matrix X has full rank.

The full rank assumption is normally assumed implicitly since if this is not the case, we can reduce X till FR is satisfied. We will in the section denote $r = \text{rank}(X)$.

Definition 4.10 (Identifiable). Suppose we have a statistical model with unknown parameter θ. We call θ identifiable if no two different value of θ lead to the same distribution of the observed data.

For a linear model, if $r < p$ then the parameter β is not identifiable. An easy example for this statement is to consider a linear model with only one observation.

Now that we have established the assumptions, let us take a look at the methods to obtain these estimators for the parameters. As seen from year one, a common way to do this is through minimising the mean square error, that is the minimizing least square problem,

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

Using calculus, by differentiating S with respect to β_j, we have

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

Setting the derivative to 0, we have

$$X^T X\hat{\beta} = X^T Y.$$

Thus, by linear algebra, $\text{rank}(X^T X) = r$ and so, $\hat{\beta}$ has a unique solution if and only if $r = p$ with $\hat{\beta} = (X^T X)^{-1} X^T Y$. Furthermore, by matrix algebra, we find that S indeed achieves a minimum at $\hat{\beta}$.

Geometrically, we see that $X\hat{\beta}$ is orthogonal to its distance to Y. Indeed, by computing their inner product, we find,

$$(X\hat{\beta})^T (Y - X\hat{\beta}) = \hat{\beta}^T X^T = \hat{\beta}^T (X^T Y - X^T X\hat{\beta}) = 0.$$

This is intuitive since we would like $X\hat{\beta}$ to be as close as possible to Y and this is achieved if the difference is orthogonal.

Proposition 8. Assuming full rank and the second order assumption, the linear model has the least square estimator

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

and

- $\hat{\beta}$ is linear in Y, i.e. the function $y \mapsto (X^T X)^{-1} X^T y : \mathbb{R}^n \to \mathbb{R}^p$ is a linear mapping;
- $\hat{\beta}$ is an unbiased estimator for β;
- $\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

Proof.

- Trivial.
- $E(\hat{\beta}) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X \beta = \beta$.
- Letting $A := (X^T X)^{-1} X^T$, we have

$$\text{Cov}(\hat{\beta}) = \text{Cov}(AY, AY) = A\text{Cov}(Y)A^T$$
$$= \sigma^2 A A^T = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

$\square$

Consider the special case of the simple linear regression $Y = X\beta + \epsilon$, where,

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, X = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{pmatrix},$$

assuming $n \geq 2$. Assuming the second order assumption and not all a_i are equal (ensuring full rank), we have

$$X^T X = \begin{pmatrix} n & n\bar{a} \\ n\bar{a} & \sum a_i^2 \end{pmatrix},$$

and so,

$$(X^T X)^{-1} = \frac{1}{n \sum a_i^2 - n^2 \bar{a}^2} \begin{pmatrix} \sum a_i^2 & -n\bar{a} \\ -n\bar{a} & n \end{pmatrix},$$

and

$$X^T Y = \begin{pmatrix} n\bar{Y} \\ \sum a_i Y_i \end{pmatrix}.$$

Thus, we find the least square estimators for β,

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

Also, by the above proposition, we have found the covariance matrix of $\hat{\beta}$,

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

We note that in the case which $\bar{a} = 0$, the covariance matrix is diagonal and $\hat{\beta}_1 = \bar{Y}$. To achieve this, we may re-parametrise the model by letting $\gamma_1 = \beta_1 + \bar{a}\beta_2$ and $\gamma_2 = \beta_2$; so,

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

resulting in $\sum(a_i - \bar{a}) = 0$ and so, $\hat{\gamma}_1 = \bar{Y}$ and $\hat{\gamma}_2 = \hat{\beta}_2$.

Definition 4.11 (Linear Estimator). An estimator $\hat{\gamma}$ is called linear if there exists some $L \in \mathbb{R}$ such that $\hat{\gamma} = L^T Y$.

Theorem 7 (The Gauss-Markov Theorem for Full-rank Linear Models). Assume (FR) and (SOA), then for all $c \in \mathbb{R}^p$, the estimator $c^T \hat{\beta}$ has the smallest variance among all linear unbiased estimators for $c^T \beta$ where $\hat{\beta}$ is the least squares estimator for β.

Proof. Let $\hat{\gamma} = L^T Y$ be any linear unbiased estimator of $c^T \beta$. Then,

$$\text{Var}(\hat{\gamma}) = \text{Var}(L^T Y) = \text{Var}(c^T \hat{\beta} + (L^T - c^T(X^T X)^{-1} X^T)Y),$$

so, by writing $D^T := L^T - c^T(X^T X)^{-1} X^T$, we have

$$\text{Var}(\hat{\gamma}) = \text{Cov}(c^T \hat{\beta} + D^T Y) = \text{Var}(c^T \hat{\beta}) + \text{Var}(D^T Y) + 2\text{Cov}(c^T \hat{\beta}, D^T Y).$$

Now, since

$$\text{Cov}(c^T \hat{\beta}, D^T Y) = c^T(X^T X)^{-1} X^T \text{Cov}(Y)D = c^T(X^T X)^{-1}(D^T X)^T \sigma^2 = 0,$$

as $D^T X = 0$ because $0 = E(\hat{\gamma}) - E(c^T \hat{\beta}) = E(D^T Y) = D^T X \beta$ for all β. So, finally

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

$\square$

4.3 Projection matrices

We shall in this section let L be a subspace of $\mathbb{R}^n$ with $\dim L = r \leq n$.

Definition 4.12 (Projection Matrix). $P \in \mathbb{R}^{n \times n}$ is a projection matrix onto L if

- for all $x \in L$, $Px = x$;
- for all $x \in L^\perp$, $Px = 0$.

By definition, we have $\text{rank} P = \dim L = r$ since $L \oplus L^\perp = \mathbb{R}^n$.

Lemma 4.2. P is a projection matrix (of the subspace $\text{Im} P$) if and only if $P^T = P$ and $P^2 = P$, i.e. P is symmetric and idempotent.

Proof.

($\implies$) By the definition of projection matrices, for all $Px \in \text{Im} P$, $P^2 x = P(Px) = Px$ and so, $P^2 = P$ by choosing x to be the standard basis. For symmetry, consider that, for all $x, y \in \mathbb{R}^n$, there exists unique decompositions $x = Px + x_1$ and $y = Py + y_1$ such that $x_1, y_1 \in \text{Im} P^\perp$ and so,

$$x^T P^T y = (Px)^T y = (Px)^T (Py + y_1) = (Px)^T Py = (x - x_1)^T Py = x^T Py.$$

Hence, $P = P^T$ as x, y are arbitrary.

($\impliedby$) For all $Px \in \text{Im}P$, by idempotent, $P(Px) = P^2x = Px$ and hence, the first condition is satisfied. Now, let $x \in \text{Im}P^\perp$, then for all $y \in \mathbb{R}^n$,

$$0 = (Py)^T x = yP^T x = yPx.$$

Hence, $Px = 0$ since y we arbitrary and so, P is a projection matrix. $\qquad\square$

Proposition 9. The projection matrix of a subspace L is unique.

Proof. Since $L \oplus L^\perp$, for each of the standard basis e_i, e_i can be uniquely written as the sum $e_i = x_i + y_i$ where $x_i \in L$ and $y_i \in L^\perp$. But we know that $x_i = Pe_i$ where P is some projection matrix onto L, we have that P is uniquely determined. $\qquad\square$

Proposition 10. If $X = (x_1, \cdots, x_r)$ forms a basis of L, then, projection matrix P onto L is

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

Proof. We can easily check that P is symmetric and idempotent. $\qquad\square$

In the case that X is an orthonormal basis, we see that the above proposition implies that $P = XX^T$.

Proposition 11. If P is the projection matrix onto L, then $I_n - P$ is the projection matrix onto $L^\perp$.

Proof. For all $x \in L$, we have $(I_n - P)x = x - Px = 0$ since $x \in L \implies Px = x$ and hence the second condition is satisfied by the fact that $L^{\perp\perp} = L$. Now, let $x \in L^\perp$, then, $(I_n - P)x = x - Px = x$ since $Px = 0$ and hence the result follows. $\qquad\square$

Lemma 4.3. If A is an $n \times n$ projection matrix of rank r, then

- r of the eigenvalues of A are 1 and the rest are 0;

- $\text{rank}A = \text{tr}A$.

Proof. Let v be an eigenvector of A with eigenvalue λ. Then $\lambda x = Ax = A^2x = \lambda^2 x$. Thus, $\lambda = \lambda^2$ and so, $\lambda = 0$ or $\lambda = 1$. Now since A is symmetric, A is diagonalisable[1], and so A is similar to $I_r \oplus 0_{n-r}$. Thus, as the trace is preserved by similarity, the second result also hold. $\qquad\square$

4.4   Residuals, Estimation of the Variance

Definition 4.13 (Vector of Fitted Values). $\hat{Y} = X\hat{\beta}$ where $\hat{\beta}$ is a least squares estimations, is called the vector of fitted values.

In the full rank case, the vector of fitted values is simply $\hat{Y} = X(X^TX)^{-1}X^TY$.

---

[1]Consider the triangularisation theorem.

Lemma 4.4. $\hat{Y}$ is unique and

$$\hat{Y} = PY,$$

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

Because of this, P is sometimes called the hat matrix since it "puts" the hat on Y.

Proof. It suffices to show $\hat{Y} = PY$ since the projection matrix is unique.

Since $PY \in \text{Im}X$, there exists some γ such that $X\gamma = PY$ and so,

$$S(\hat{\beta}) = \|Y - PY + PY - X\hat{\beta}\|^2 = \|Y - PY\|^2 + \|PY + X\hat{\beta}\|^2 + 2(Y - PY)^T(PY - X\hat{\beta}).$$

Since $Y - PY = (I - P)Y$ which is the projector onto $\text{Im}X^\perp$, and $PY - X\hat{\beta} \in \text{Im}X$,

$$2(Y - PY)^T(PY - X\hat{\beta}) = 0,$$

and so,

$$S(\hat{\beta}) = S(\gamma) + \|PY - X\hat{\beta}\|^2 \geq S(\hat{\beta}) + \|PY - X\hat{\beta}\|^2.$$

Thus $\|PY - X\hat{\beta}\| = 0$ which means $PX = X\hat{\beta}$. $\qquad\square$

Definition 4.14 (Residuals). $e = Y - \hat{Y}$ is called the vector of residuals.

By using the above lemma, we see that $e = Y - PY = (I - P)Y = QY$ and so, the residual is simply the image of Y under the projector onto $\text{Im}P^\perp$. Furthermore, we see that

$$E(e) = E(QY) = QE(Y) = QX\beta = 0,$$

since $QX = 0$.

Suppose we have data from the model

$$Y = X\beta + Z\gamma + \epsilon,$$

where $E(\epsilon) = 0$, $Z \in \mathbb{R}^n \setminus \text{Im}X$ and γ is deterministic. Suppose now that we erroneously worked with the model

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

instead, i.e. assumed $\gamma = 0$. Then

$$E(e) = E(QY) = E(Q(X\beta + Z\gamma + \epsilon)) = QZ\gamma.$$

Thus, by plotting QZ against the residuals, we can consider the importance of Z and thus, perhaps include it in our model if the slope of the plot differs from 0.

Definition 4.15 (Residual Sum of Squares). $\text{RSS} = e^T e$ is called the residual sum of squares.

The residual sum of squares quantified the departure of the data from the modal and it is the minimum of $S(\beta)$. The residual sum of squares has many equivalent forms.

- $\text{RSS} = \sum_{i=1}^n e_i^2$;

26

- RSS $= S(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$;
- RSS $= (QY)^T QY = Y^T QY$;
- RSS $= Y^T Y - \hat{Y}^T \hat{Y}$.

Proposition 12. $\hat{\sigma}^2 := \frac{\text{RSS}}{n-r}$ is an unbiased estimator of σ^2 where r is the rank of X.

Proof. Let $Q = I - P$. Since P is a projection matrix, so is Q, and hence, RSS $= Y^T QY$. So,

$$
\begin{aligned}
E(\text{RSS}) = E(\text{tr}(Y^T QY)) &= E(\text{tr}(QYY^T)) = \text{tr}(QE(YY^T)) \\
&= \text{tr}(Q(\text{cov}(Y) + E(Y)E(Y)^T)) = \text{tr}(Q\sigma^2) + \text{tr}(QX\beta(X\beta)^2) \\
&= \sigma^2 \text{tr}(I - P) = \sigma^2(n - \text{tr}(P)) = \sigma^2(n - r).
\end{aligned}
$$

$\square$

This is a generalisation of the result that the sample variance s^2 is an unbiased estimator for σ^2 when $Y_1, \cdots Y_n$ are i.i.d with unknown mean μ and unknown variance σ^2. Indeed, by writing this as the linear model

$$
Y = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \epsilon = \mathbf{1}\mu + \epsilon,
$$

with $E(\epsilon) = 0$, and $\text{cov}(\epsilon) = \sigma^2 I$, then $P = X(X^T X)^{-1} X^T = \frac{1}{n} X X^T$ and thus,

$$
e = Y - \hat{Y} = Y - PY = Y - \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix},
$$

and hence, $\frac{\text{RSS}}{n-r} = \frac{\sum (Y_i - \bar{Y})^2}{n-1} = s^2$.

Definition 4.16 (Coefficient of Determination)**.** The coefficient of determination of the simplest model
$$
Y = \mathbf{1}\mu + \epsilon,
$$
is the statistic
$$
R^2 = 1 - \frac{\text{RSS}}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.
$$

Since we would like to minimise RSS, we would like large R^2 for our model. Intuitively, since RSS$/n$ is an estimator of σ^2 while $\frac{1}{n} \sum (Y_i - \bar{Y}^2)^2$ is an estimator of σ^2 is the model with only the intercept term. Thus, R^2 provide us with a coefficient comparing the variance from the model to the variance observed in the data.

We remark that adding columns to X, i.e. adding parameters will never decrease R^2, and so the coefficient of determination should not be used directly for model comparisons. As we shall see, one should penalise models with a larger number of parameters.

5 Linear Model with Normal Theory Assumptions

Previously, we have considered general linear models sometimes with second order assumptions. We shall in this section strengthen this assumption with the normal theory assumption (NTA), i.e. given the model

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

where $E(\epsilon) = 0$, we assume $\epsilon \sim N(0, \sigma^2 I_n)$.

5.1 Distribution Results

We define the multivariate normal distribution. Note the difference in definition in comparison to the definition found in *probability for statistics*, in particular, we had defined the multivariate normal distribution using the joint probability density function

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

for some vector μ and positive definite matrix Σ. Furthermore, it was found that Z has the mean μ and the covariance matrix Σ.

In the case that we impose the NTA, we have $\Sigma = \sigma^2 I_n$, and so, we find

$$f_Z(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i - \mu_i)^2}{2\sigma^2}\right),$$

and so, $Z_1, \cdots, Z_n$ are independent with $Z_i \sim N(\mu_i, \sigma^2)$ for $i = 1, \cdots, n$.

We may generalise the multivariate normal distribution such that Σ is permitted to be only positive semidefinite.

Definition 5.1 (Multivariate Normal Distribution)**.** An n-variate random vector Z follows a multivariate normal distribution if for $a \in \mathbb{R}^n$, the random variable $a^T Z$ follows a univariate normal distribution (the degenerate case $N(\cdots, 0)$ is allowed).

Proposition 13. Let $X_1, \cdots, X_r \sim N(0, 1)$ be i.i.d. and let $\mu \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times r}$. Then $Z = AX + \mu \sim N(\mu, AA^T)$.

Proof. Let $a \in \mathbb{R}^n$, then $a^T Z = a^T AX + a^T \mu$. Since $a^T A$ is simply a row vector, $a^T AX$ is a linear combination of X_i and so, is an univariate normal random variable. Furthermore, as $a^T \mu$ is simply a constant, the result follows. $\square$

One may also define the multivariate normal distribution using the characteristic function where we say, $Z \sim N(\mu, \Sigma)$ if its characteristic function

$$\phi : \mathbb{R}^n \to \mathbb{C} : t \mapsto E\left(\exp\left(iZ^T t\right)\right)$$

satisfies

$$\phi(t) = \exp\left(i\mu^T t - \frac{1}{2}t^T \Sigma t\right),$$

for all $t \in \mathbb{R}^n$, where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix.

Proposition 14. Let $Z \sim N(\mu, \Sigma)$, then

- $E(Z) = \mu$;

- $\text{Cov}(Z) = \Sigma$;

- if A is a deterministic matrix and b is a deterministic vector of appropriate dimensions, then
$$AZ + b \sim N(A\mu + b, A\Sigma A^T).$$

Proof. Easy. $\square$

As demonstrated above, it turns out under normal distribution, uncorrelated implies independent. We note that this is in general not true.

Lemma 5.1. For $i = 1, \cdots, k$, let $A_i \in \mathbb{R}^{n_i \times n_i}$ be positive semidefinite, and symmetric and let Z_i be an n_i-variate random vector. If $Z \sim N(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^{\sum n_i}$, $\Sigma = \bigoplus_{i=1}^{k} A_i$, then $Z_1, \cdots, Z_k$ are independent.

Proof. In the special case that all of the A_i are positive definite, the result follows as $\Sigma^{-1} = \text{diag}(A_1^{-1}, \cdots, A_k^{-1})$ and thus, the pdf of Z can be easily factored into the respective partial pdfs.

In the general case, we may use the characteristic polynomial and show that the characteristic polynomial can be factored into the individual partial characteristic polynomials. $\square$

Definition 5.2 (Non-Central χ^2-Distribution)**.** Let $Z \sim N(\mu, I_n)$ where $\mu \in \mathbb{R}^n$. Then, $U = Z^T Z = \sum_{i=1}^{n} Z_i^2$ is said to have a non-central χ^2-distribution with n degrees of freedom and non-centrality parameter
$$\delta = \sqrt{\mu^T \mu},$$
and we denote this by $U \sim \chi_n^2(\delta)$.

In order for this to be a proper definition we need to show that the distribution of U to depend on μ only through $\mu^T \mu$. One approach is to show that the moment generating function of U is
$$M_U(t) = \frac{1}{(1 - 2t)^{n/2}} \exp\left(\frac{t\delta^2}{1 - 2t}\right).$$
See the problem sheet for the full detail.

Lemma 5.2. Let $U \sim \chi_n^2(\delta)$, then $E(U) = n + \delta^2$ and $\text{Var}(U) = 2n + 4\delta^2$. Furthermore, if $U_i \sim X_n^2(\delta_i)$, are independent for $i = 1, \cdots, k$. Then $\sum i = 1^k U_i \sim \chi^2_{\sum n_i \sqrt{\sum \delta^2}}$.

Proof. Problem sheet. $\square$

Definition 5.3 (Non-Central t-Distribution)**.** Let X and U be independent random variables with $X \sim N(\delta, 1)$ and $U \sim \chi_n^2$. The distribution of
$$Y = \frac{X}{\sqrt{U/n}},$$

29

is called the non-central t-distribution with n degrees of freedom and non-centrality parameter δ and we denote this by $Y \sim t_n(\delta)$.

If $Y_n \sim t_n$ for all $n \in \mathbb{N}$; then,

$$Y_n \to N(0,1) \text{ in distribution as } n \to \infty.$$

Indeed, if $X \sim N(0,1)$ and $U_n = \sum Z_i^2$ for $Z_i \sim N(0,1)$ independent, by the weak law of large numbers, $U_n/n \to E(Z_i^2) = 1$ in probability. As $x \mapsto \sqrt{x}$ is continuous, we have $\sqrt{U_n/n} \to 1$ in probability. Thus, by Slutsky's lemma,

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

Lastly, there is also the F-distribution.

Definition 5.4 (Non-Central F-Distribution)**.** If $W_1 \sim \chi^2_{n_1}(\delta)$ and $W_2 \sim \chi^2_{n_2}$ are independent, then

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

is said to have non-central F-distribution and we write $F \sim F_{n_1,n_2}(\delta)$.

5.2 Independence Results

Lemma 5.3. Let $A \in \mathbb{R}^{n \times n}$ be a positive semidefinite symmetric matrix with rank r. THen there exists some $L \in \mathbb{R}^{n \times r}$ such that L also has rank r, $A = LL^T$ and $L^T L$ is the diagonal matrix of the nonzero eigenvalues of A.

Proof. Revision. $\qquad \square$

Lemma 5.4. Let $X \sim N(\mu, I)$, $A \in \mathbb{R}^{n \times n}$ be a positive semidefinite symmetric matrix and let B be some matrix such that $BA = 0$. Then, $X^T AX = BX$ are independent.

Proof. Lert $r = \text{rank}(A)$. By the lemma above, there exists some $L \in \mathbb{R}^{n \times r}$ such that $\text{rank} L = r$ and $A = LL^T$. Then

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

which implies independence since they are jointly normally distributed. $\qquad \square$

Lemma 5.5. If $Z \sim N(\mu, I_n)$, and A is an $n \times n$ projection matrix of rank r, then

$$Z^T AZ \sim \chi^2_r(\delta) \text{ with } \delta^2 = \mu^T \mu.$$

Proof. All nonzero eigenvalues of A are equal to 1 and so, there exists some $L \in \mathbb{R}^{n \times r}$ such that $A = LL^T$ and $L^T L = I_r$. Let $V = L^T Z$, then $V \sim N(L^T \mu, I)$ and

$$Z^T AZ = Z^T LL^T Z = V^T V \sim \chi^2_r(\delta),$$

where $\delta = (L^T \mu)^T L^T \mu = \mu^T LL^T \mu = \mu^T A\mu$. $\qquad \square$

Lemma 5.6. If $Z \sim N(\mu, I_n)$ and $A_1, A_2 \in \mathbb{R}^{n \times n}$ are projection matrices and $A_1 A_2 = 0$. Then, $Z^T A_1 Z$ and $Z^T A_2 Z$ are independent.

Proof.
$$\text{cov}(A_1 Z, A_2 Z) = A_1 \text{cov}(Z) A_2 = A_1 A_2 = 0.$$

$\square$

This result extends to the $Z^T A_1 Z, \cdots, Z^T A_k Z$ where $A_i A_j = 0$ for all $i \neq j$.

Lemma 5.7. If $A_1, \cdots, A_k$ are symmetric $n \times n$ matrices such that $\sum A_i = I_n$ and if $\text{rank} A_i = r_i$, then the following are equivalent:

- $\sum r_i = n$;

- $A_i A_j = 0$ for all $i \neq j$;

- A_i is idempotent for all $i = 1, \cdots, k$.

Proof. Exercise.

$\square$

Theorem 8 (Fisher-Cochran Theorem)**.** If $A_1, \cdots, A_k$ are $n \times n$ projection matrices such that $\sum A_i = I_n$, and if $Z \sim N(\mu, I_n)$. Then $Z^T A_1 Z, \cdots, Z^T A_k Z$ are independent and

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

Proof. Direct consequence of the above lemmas.

$\square$

5.3 Linear Model with NTA Results

Recall that the NTA states $\epsilon \sim N(0, \sigma^2 I_n)$ and in particular $Y \sim N(X\beta, \sigma^2 I_n)$ in the linear model $Y = X\beta + \epsilon$.

By consulting the joint distribution of a multivariate normal distribution, the log-likelihood of the data is simply,

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

By noticing the first term is a constant while the second term is proportional to $S(\beta)$ - the least squares, we maximize the likelihood by minimising the least squares of β.

Lemma 5.8 (Distribution of RSS)**.** Assuming NTA;

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

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

Proof. Let P be the projection matrix onto span$X$. Then,

$$\text{RSS} = e^T e = ((I - P)Y)^T (I - P)Y = Y^T Q^T QY = Y^T QY,$$

where $Q := I - P$. Hence,

$$\frac{\text{RSS}}{\sigma^2} = \frac{Y^T}{\sigma} Q \frac{Y}{\sigma} = Z^T QZ,$$

where $ZY/\sigma \sim N(X\beta/\sigma, I)$ and Q is a projection matrix. Furthermore, $Q + P = I$, we have rank$Q$ + rank$P = n$ implying rank$Q = n - r$. Hence,

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

since $(X\beta/\sigma)^T QX\beta/\sigma = 0$. $\qquad\square$

Lemma 5.9. Assuming FR and NTA in a linear model, and let $c \in \mathbb{R}^p$, then

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

Proof. See official notes. $\qquad\square$

Thus far, we have derived pivotal quantities for one-dimensional parameters. If we were interested in the behaviour of more than one components of the parameters, we can uses the Bonferroni correction to obtain new tests based on these to obtain the higher dimensional confidence intervals. The F-test however, is a more complicated test about β.

Let RSS = the residual sum of squares in the full model $E(Y) = X\beta$ and let RSS_0 = the residual sum of squares in the sub-model $E(Y) = X_0 \tilde{\beta}$.

Lemma 5.10. Under $H_0 : E(Y) \in \text{span}(X_0)$,

$$F = \frac{\text{RSS}_0 - \text{RSS}}{\text{RSS}} \frac{n - r}{r - s} \sim F_{r-s, n-r},$$

where $r = \text{rank}X$ and $s = \text{rank}X_0$.

Let P be the projection matrix onto the span of X, and $Q = I - P$. Likewise, let P_0 be the projection onto the span of X_0 and $Q_0 = I - P_0$, then $\text{RSS} = Y^T QY$ and $\text{RSS}_0 = Y^T Q_0 Y$. Substituting this into the above equation, we have

$$F = \frac{Y^T (P - P_0)Y/\sigma^2}{Y^T (I - P)Y/\sigma^2} \frac{n - r}{r - s}.$$

Hence, the result follows if the numerator follows χ^2_{r-s} and the denominator follows χ^2_{n-r} and they are independent.

Proof. Let $Z = Y/\sigma$, $A_1 = I - P$, $A_2 = P - P_0$, $A_3 = P_0$ and we will use the Fisher-Cochran theorem.

32

Clearly $A_1 + A_2 + A_3 = I$ and we know that A_1, A_3 are projection matrices. Now, as $P - P_0$ is symmetric and

$$(P - P_0)^2 = P^2 + P_0^2 - PP_0 - P_0 P$$

as every column y of P_0 is an element of $\text{span} X_0$, and thus an element of $\text{span} X$, we have $Py = y$. Hence, $PP_0 = P$. Thus, $P_0 P = (P^T P_0^T)^T = (PP_0)^T = P_0^T = P_0$. Hence,

$$(P - P_0)^2 = P + P_0 - P_0 - P_0 = P - P_0,$$

and is idempotent. With that we see A_2 is also a projection matrix. So, by the Fisher-Cochran theorem we have

- $Z^T(P - P_0)Z$ and $Z^T(I - P)Z$ are independent;

- $Z^T(P - P_0)Z \sim \chi^2_{\text{rank}(P-P_0)}(EZ^T(P - P_0)EZ)$;

- $Z^T(I - P)Z \sim \chi^2_{\text{rank}(I-P)}(EZ^T(I - P)EZ)$.

Under H_0, we know that $EZ = \frac{1}{sigma} EY \in \text{span} X_0 \subseteq \text{span} X$, and thus,

$$(P - P_0)EZ = EZ - EZ = 0.$$

So, we have the centrality requirement. Clearly, the ranks corresponds to the appropriate degrees of freedom, and hence, the result follows. $\qquad\square$

In the case H_0 is not true, the proof remains valid without the centrality condition. Indeed, in all cases, the F-test applies in which the resulting distribution follows the F-distribution.

Suppose $EY = X\beta$ is a linear model satisfying FR and NTA. We are interested in finding a confidence region for β. It turns out the following quantity is a pivotal quantity for β,

$$A := \frac{(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)}{\text{RSS}} \frac{n - p}{p},$$

in the case that we can work out the distribution of A.

Indeed the numerator of the first fraction can be written as

$$(Y - X\beta)^T P(Y - X\beta).($$

By using $P = X(X^T X)^{-1} X^T$, this is equal to

$$X(\hat{\beta} - \beta)^T X(\hat{\beta} - \beta)).$$

Furthermore, RSS can be written as

$$\text{RSS} = Y^T QY = (Y - X\beta)^T Q(Y - X\beta).$$

Thus, letting $Z := \frac{1}{\sigma}(Y - X\beta)$, we have

$$A = \frac{Z^T P Z}{Z^T Q Z} \frac{n - p}{p}$$

where $Z \sim N(0, 1)$, $P + Q = I$, and so, by the Fisher-Cochran theorem, $A \sim F_{p, n-p}$. Hence, a $1 - \alpha$ confidence region R for β is defined by all $\gamma \in \mathbb{R}^p$ such that

$$\frac{(\hat{\beta} - \gamma)^T X^T X (\hat{\beta} - \gamma)}{\text{RSS}} \frac{n - p}{p} \leq F_{p, n-p, \alpha}$$

where $P(Z \geq F_{p, n-p, \alpha}) = \alpha$ for $Z \sim F_{p, n-p}$. The region R is an ellipsoid centred at $\hat{\beta}$.