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.