1 Lecture I: Introduction to Statistical Model Building

Required Reading: Judge, Hill et. al., all of Chapter 1, Chapter 2, pp. 58–87. Handout: *Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics*, Dale Poirier, pp. 121–133. Recommended: Kennedy, section on "Criteria for Estimators."

This lecture will attempt to describe what a *statistical model* is, and from there describe methods of estimating certain aspects of such a model. We start by describing three things:

1. An experiment,
2. A random sample
3. An estimator

An **experiment** is a method for generating data. This is a generic term because often, and especially in finance, data is a pre-condition for doing analysis and we more often than not find data as opposed to design and conduct an experiment to generate it. To get a sense for how serious this issue is, we quote from *Maximum Entropy Econometrics*, Judge, Golan and Miller:

> In this book we focus on recovering and processing information when the underlying sampling model is incompletely or incorrectly known and the data are limited, partial or incomplete. Furthermore, given the non-experimental model building and data generation restrictions of economics (in particular) and in the social sciences (in general), many of the economic-statistical models that form the basis for conventional estimation and inference are ill-posed or underdetermined...
>
> Given non-experimental data generation processes in economics and finance, and the corresponding data available for applied econometric analyses, we face the problem of how to make predictions for the population, or how to construct an image of it, using information at hand. Unfortunately, in many cases, the models that economists use for analysis and choice purposes contain unknowns that are unobserved and indeed are inaccessible to direct measurement.

For example, in capital asset pricing we cannot observe the all-important *equity risk premium* and a central question in financial economics concerns its value.

All this said, when working with financial data it is often useful *in a certain sense* to regard the data as having been produced by an experiment. The experiment may be regarded as having a *mechanism* for generating data, and the outcome of the experiment is a collection of data

$$(Y_1, \ldots, Y_N) = \mathbf{Y} \tag{1}$$

Each result, Y_i, is thought of as the outcome of one *trial* in the experiment. The number (or vector) Y_i thought to be a draw from some random variable, say, $\mathbf{Y_i}$. We are primarily interested in the joint distribution of the variables $(\mathbf{Y_1}, \ldots, \mathbf{Y_N})$. Often the problem is stated as a *parametric* problem in that the joint p.d.f. of the $\mathbf{Y_i}$'s is assumed to be of a known form $f(y|\theta)$, where θ is a vector of parameters. In this case, then, the objective of stastical estimation is to propose actual values for the elements of θ and to have a means of evaluating the quality of the choices made.

The following are examples of typical econometric estimation problems:

- **Forecasting**: given past daily returns $r_{t_1}, \ldots, r_{t_N}$ of a stock or index, what is the mean and standard deviation of the next period return. In this case, some model has to be proposed for the joint distribution of returns. A very simple model might be that the returns are *independet* and identically distributed as normal distributions. In this case, each return r_{t_i} has the same p.d.f, $f(\mathbf{r}|(\mu, \sigma))$, and the joint p.d.f. is $f(r_{t_1}|(\mu, \sigma)) \cdot f(r_{t_2}|(\mu, \sigma)) \cdot \ldots$.
- **Covariance**: given N assets, $A_1, \ldots, A_N$, what is the covariance matrix of the one month returns r_i of A_i.

2 Estimators

In this lecture we will discuss the following estimators:

- Method of Moments
- Maximum Likelihood Estimation

- Least Squares
- Maximum Entropy

The first three estimation techniques are discussed in Judge, Hill et. al. and will not be repeated here. We will discuss, however, the latter two.

2.1 Parametric Versus Non-Parametric Estimation

We will distinguish two different methods of estimation.

1. **Parametric.** Parametric estimation refers to proposing a particular joint distribution of the sample that is parametrized by a vector of parameters θ. For example, if you say that you sample is independent and normal, then you have implicitly parametrized your possible set of distributions to a vector sitting in $\mathbf{R^2}$.

2. **Non-parametric.** This refers, loosely speaking, to the situation in wich a sample of data probability distribution or moments are directly estimated, without specifying a particular parametrization. For example, you might estimate the mean of a distribution directly, without reference to any particular distribution. Below we will discuss both parametric and non-parametric methods of estimation.

2.2 Sample Moments

Recall that the ith moment of a distribution Y about the origin is defined as $\mu_r' = E[Y^r]$. Now, think of the probability density function as depending on a number of parameters, as in, e.g.,

1. A **normal distribution** $N(\mu, \sigma)$ is *completely characterized* by the parameters μ and σ.

2. A **chi-squared distribution** $\chi^2_{(m)}$ depends on the number of degress of freedom.

The moments of a distribution will always be some function of the parameters that determine the distribution, as in

1. If X $N(\mu, \sigma)$ then $E[X] = \mu$, $E[X^2] = \sigma^2 + \mu^2$.

2. If Y $\chi^2_{(m)}$ then $E[Y] = m$.

The idea of the method of moments is the notion of *moment matching*. Let $Y_1, \ldots, Y_N$ be a sample of data from an experiment. Define the r-th sample moment about the origin as

$$\hat{m}u'_r = \sum_{i=1}^{T} Y_i^r / T, \quad r = 1, \ldots, K$$

The idea behind the method of moments is to set up a system of K equations as above (one for each $r = 1, 2, \ldots$) relating the sample moments to the parameters, as in the case of a normal distribution where we find two equations suffice:

$$\mu'_1 = \mu$$
$$\mu'_2 = \sigma^2 + \mu^2$$

2.3 Maximum Likelihood Estimation

Class will cover example of N coin flips.

2.4 Discussion

Note in both the method of moments and the maximum likelihood estimation it is necessary to propose the full distribution of a set of data in order to perform the estimation procedure. This form of modeling can lead to specification error in that it is possible that the model could be simply wrong (and this can be quite subtle, as you will see in the special reading).

Here is an example of how mis-specification can lead to problems in estimation. Suppose we flip a coin 10 times and we propose the following model: $P(H) = p, P(T) = q, p + q = 1$. This seems straighforward enough, right? Then, if we flip the coin 10 times and we see 3 heads and 7 tails, the maximum likelihood estimate of p is $3/10$ and q is $7/10$. Now, suppose it turns out that there is always a 5% chance of the coin landing on its side. That is, suppose the correct model for the experiment is $P(H) = p, P(T) = q, P(S) = .05$. Then in this case we must perform the maximum likelihood estimation with the restriction that $p + q = .95$, thus changing the outcome of estimation procedure in an obvious way.

2.5 Least Squares

Least squares method is an estimation procedure which states that the estimate of the parameter μ'_r solves the following problem:

$$\mu'_r := min_x \sum (Y_i^r - x)^2$$

where the sum ranges over all sample points.

2.6 Maximum Entropy Criterion

Maximum entropy estimation is a lesser known, but interesting and important estimation technique. An en excellent reference is *Maximum Entropy Econometrics*, by Golan, Judge and Miller, Wiley. Quoting from their introduction the authors describe what it means for an estimation problem to be *ill-posed*:

> The problem is ill-posed if there is not enough information... to permit the receover of the desied K-dimension β parameter vector by traditional estimation methods, or in general, when an unknown function has to be inferred from insufficient information...
>
> Unfortunately, because of limited, partial data or insufficient information, many econometric problems fall in the ill-posed, underdetermined category. *Convenient assumptions*, representing information that we do not possess, are typically used to convert ill-posed problems into seemingly well-posed statistical models that can be analyzed with familiar classical or Bayesian tools. However, this approach often leads to erroneous interpretations o and treatments. In fact, in applied mathematics, statistics and econometrics, ill-posed inverse problems may be the rule rather than the exception. Consequently, it is necessary to consider the best way to proceed in analyzing these logically indeterminate situations.

The idea of an ill-posed problem will arise over and over again in our studies of financial data and statistical estimation. The idea of maximum entropy estimation is very straightforward: we imagine a there is a data generation mechanism (e.g., part of an experiment) that can generate one of K possible outcomes; suppose N trials are performed and N_i is the number of times the i-th outcome occurs so that

$$\sum N_i = N$$

Now write

$$p_k = \frac{N_k}{N}, \quad N_k = Np_k,$$

and denote by W the number of possible ways the set $(N_1, \ldots, N_k)$ might have occured:

$$W = \frac{N!}{Np_1!Np_2!\ldots Np_k!} = \frac{N!}{\prod_k N_k!}$$

Now, note that

$$lnW = lnN! - \sum_{k=1}^{K} lnN_k!.$$

For x large we have the Stirling formula:

$$\ln x! \approx x \ln x - x, \quad x \to \text{inf}$$

so that for large samples (large N) we have

$$\ln W \approx N \ln N - N - \sum_{k=1}^{K} N_k \ln N_k + \sum_{k=1}^{K} N_k = N \ln N - \sum_{k=1}^{K} N_k \ln N_k.$$

Finally, noting that as $N \to$ inf we have

$$\frac{N_k}{N} \to p_k$$

we have after a little manipulation

$$\ln W \approx -N \sum_{k=1}^{K} p_k \ln p_k.$$

This leads to

$$N^{-1} \ln W \approx -\sum_{k=1}^{K} p_k \ln p_k.$$

The right hand side is known as the *entropy* of the system, and we have shown that it is approximately proportional to the log of the number of ways that the outcomes could have occured. Thus, since log is monotonic, we have shown (approximately) that for a given sample that produces $N_1, \ldots, N_k$ outcomes of outcomes $1, \ldots, k$ respectively, maximizing the entropy of the system is equivalent (approximately) to maximizing the number of ways that might have produced the outcome.

The following simple example is taken from Golan, Judge and Miller. This example also demonstrates that nature of ill-posed and underdetermined problems and should serve as an archetypal example for you in things to come.

We suppose that we have a 6-sided die and we want to estimate that probabilities p_i of rolling an i in a given roll (we think of the experiment as rolling the die and each "trial" as one roll of the die). Now, suppose the only information you are given to make this estimate is the average value y of a large number of roles. Note:

1. If you were given the *frequency distribution* of the roles, that is, if you knew the percentage of times that a one, two, three, etc. were roled, then you could use maximum likelihood estimation or the method of moments to try to estimate these values.

2. You could try to make some extra assumptions to estimate the probabilities. For example, you might assume the die is fair and then find the distribution closest to $p_1 = p_2 = \cdots = p_6 = 1/6$ in some or other "distance". But what is the justification for the initial assumption of fairness?

3. The problem is ill-posed, there are only two pieces of information: the probabilities sum to one ($\sum p_i = 1$) and the average value of all the roles.

One way to estimate the parameters is via the maximum entropy criterion. Namely we solve the problem:

$$\text{maximize} \mathbf{H} = -\sum_{\mathbf{k=1}}^{\mathbf{6}} \mathbf{p_k} \ln(\mathbf{p_k}) \tag{2}$$

subject to

$$\sum p_k k = y$$
$$\sum p_k = 1.$$

This is easily solved using Lagrange multipliers:

$$L = -\sum_{k=1}^{6} p_k \ln(p_k) + \lambda(y - \sum_{k=1}^{6} p_k k) + \gamma(1 - \sum_{k=1}^{6} p_k).$$

Since $\mathbf{H}$ is strictly concave and for $y \in (1,6)$ the constraint set is nonempty and compact, there is a unique interior solution to this problem (one of your homework problems will be to find it).

There are a couple of things to note about this estimation technique:

1. The technique is *non-parametric*. That is, we did not propose a particular probability distribution and then estimate the parameters of the distribution. Rather, we estimated the distribution itself.

2. The technique introduced a simple criterion that was in a sense natural (find the probabilities that maximize the number of ways that the die could have produced the outcome); this one additional criterion changed the problem from ill-posed to well-posed, in that after introducing the entropy criterion the problem was reduced to having a unique solution.

3. It is possible to argue that a different criterion might be more appropriate. This could be the case if you had additional information about the problem (for example, if you thought it was likely that the die was fair). Whenever you enter a situation in which you have to introduce additional information, you automatically have to make an argument for or against a particular type of criterion. In the case of maximum entropy, researchers make arguments along a number of lines, including: the method seems natural, and the method seems to work as well or better than other methods in situations where the true distribution is known.

3 Point Estimates and Confidence Intervals

To this point, we have focussed on estimating parameters in the sense of supplying a precise *point estimate* for the value of a parameter. For example, a point estimate for the mean of a distribution is a particular number that is a "best guess" of what the mean is. In the next lecture we will discuss criteria for esitmators, what makes some estimators good and some not so good (see Kennedy for a preview), but for now, we have already seen that estimators may be proposed in a number of ways.

An alternative to proposing a point estimate for a parameter is to propose a *confidence interval*. A confidence interval is an interval (a, b) that, roughly speaking, has a known probability of containing the parameter in question. This is actually a complicated and often misunderstood concept, so we will explain it right now.

We suppose that we have a data generating mechanism D that generates data. We have an experiment that generates N trials from D (for example, imagine that D is the role of a die, so that our experiment consists of rolling a die N times). Then a *repeated experiment* is a repetition

of our experiment. This is an important notion and it will come up again in the context of studying particular properties of estimators.

Write S_{D_i} for the sample of data that arises from the i-th repeated experiment. Suppose we have an estimator for a scalar parameter θ for the joint p.d.f. $f(S|\theta)$. Then a *95% confidence interval* for θ is an interval constructed from the data itself

$$I = (a(S_{D_i}), b(S_{D_i}))$$

such that the probability (in the space of repeated experiments) of I containing θ is 95%.