Section 5*

Econ 240A - Second Half

Ingrid Haegele†
University of California, Berkeley

November 16, 2018

*These section notes rely on the notes prepared and revised by Markus Pelger, Raffaele Saggio and Seongjoo Min.

†E-mail: inha@berkeley.edu

1 Introduction

Last section, we talked about linear projections on a population level. However, in practice we usually don't know what the population regression vector is, but we use samples quantities instead. The focus of this section is therefore to understand how to perform inference on population objects given the availability of a single random sample. We start by introducing the OLS estimator and its finite sample properties. Second, we will talk about its large sample properties in context of asymptotic approximations. We will also remind ourselves how hypothesis testing works and in the last part of this section, we will introduce the concept and application of quantile regression.

2 OLS Estimator

One approach to define the Ordinary Least Squares (OLS) estimator is to impose strong assumptions, such as non-stochastic regressors, a linear CEF, normally distributed errors and homoskedasticity. This setup is often referred to as classical normal regression model. Thanks to the strong assumptions, the OLS estimators achieves unbiasedness and we can easily obtain a formula for the sampling variance of the OLS estimator, even for small samples. We start with this traditional view on the OLS estimator, since it is straightforward to understand and widely used in textbooks. However, note that for applied work, the later introduced approach based on asypmtotic approximations is much more relevant.

2.1 Setup

Let us first introduce the setup. Let $\mathbf{Y}$ be an $(N \times 1)$ vector, collecting observations of the response (or dependent) variable. Let $\mathbf{X}$ be an $N \times K$ matrix, collecting observations of the explanatory (or independent) variables. That is

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{1}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix} \tag{2}$$

Here, the rows of $\mathbf{X}$ store observations (individuals, countries, etc.) and the columns store variables. The row vector $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})^\top$ contains the values of each of the K regressors for observation i, hence is of dimension $K \times 1$.

Definition 1: (OLS estimator) The OLS estimator, denoted as $\hat{\beta}$, is defined as

$$\hat{\beta} = \underset{\beta \in \mathcal{R}^K}{argmin} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \beta)^2 = \underset{\beta \in \mathcal{R}^K}{argmin} (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) \tag{3}$$

2.2 The Classical Linear Normal Regression Model (CLNRM)

Let us impose the following assumptions

1. (*Non Stochastic Regressors*): $\mathbf{X}$ is a non-stochastic $N \times K$ matrix
2. (*Linear Expectation*): $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\beta$.
3. (*Homoskedasticity*): $\mathrm{Var}[\mathbf{Y}] = \sigma^2 I_N$.
4. (*Invertibility*): $\mathbf{X}$ is of full column rank K.
5. (*Normality*): $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta; \sigma^2 I_N)$;

Under these assumptions, in particular, under A.4 (invertibility), we have that

$$\hat{\beta} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i y_i \right) = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \left(\mathbf{X}^\top \mathbf{Y} \right) \tag{4}$$

Notice that we are implicitly assuming that the Data Generating Process (DGP) is given by

$$y_i = \mathbf{x}_i^\top \beta + e_i \iff \mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \tag{5}$$

Notice that given assumption 2, e_i represents the CEF error.

2.3 Finite Sample Moments

We are interesting in evaluating how good the OLS estimator is compard to the true (but unknown) population value. Thus, we are interested in its finite sample properties, such as whether it is bised and how large its variance is (compared to other estimators in the same class).

Notice that $\hat{\beta}$ is clearly a random variable. In particular its randomness comes from the fact that we have collected only a random sample of observations on $\mathcal{Y}$. Therefore $\hat{\beta}$ has a sampling distribution and we are interested in understanding the moments of such distribution.

2.3.1 Unbiasedness

Unbiasedness is a fundamental, but often hard to check, property. It states that the expected value of our estimator is centered around the true parameter. The expected value of an estimator represents, roughly speaking, the mean of the estimator computed over infinitely many samples. In our setup and given our assumptions

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}\right) = \mathrm{E}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{X}\beta\right) = \beta \tag{6}$$

Hence $\hat{\beta}$ is unbiased. We achieved this result thanks to the invertibility and linearity assumptions.

2.3.2 Variance

Similarly to what has been done previously, we can define the dispersion of our estimator around its sampling distribution,

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}\right) = \mathrm{Var}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathrm{Var}\,\mathbf{Y})\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \tag{7}$$

Clearly, we have used our homoskedasticity assumption and the fact that $\mathbf{X}$ is non-random.

2.3.3 Gauss Markov Theorem

It turns out that the OLS estimator is BLUE (Best Linear Unbiased Estimator). In other words, if we search within the class of linear, unbiased estimators, the variance of $\hat{\beta}$ is the

smallest in some sense. This is the so called Gauss Markov Theorem

Gauss Markov Theorem: Let assumption 1-5 hold. Then the OLS estimator, $\hat{\beta}$, has minimum variance among the class of linear estimators.

3 Large Sample Approximations

In the previous part, we started from very strong assumptions to show how one can conduct finite sample analysis of the OLS estimator. These assumptions are hard to justify and often completely unrealistic. However, we needed these assumptions because in principle the sampling distribution of $\hat{\beta}$ is a function of the joint distribution of $\{y_i, \mathbf{x}_i^\top\}$ and the sample size N. If we do not impose such assumptions, this function may be extremely complicated which makes it infeasible to analytically calculate the exact distribution of $\hat{\beta}$ except in very special cases (i.e. using the normality assumption).

The name of the game therefore is to understand whether in a setup that uses only a very primitive assumption - the observations $\{y_i, \mathbf{x}_i^\top\}$ are all $i.i.d$ - it is still possible to conduct inference for $\hat{\beta}$. The answer turns out to be yes and the main tool to achieve this goal is via approximation methods. The most widely used and versatile method is asymptotic theory, which approximates sampling distributions by taking the limit of the finite sample distribution as the sample size N tends to infinity. It is important to understand that as soon as we use the asymptotic distributions to assess the finite sample distribution of our estimators in practice, this represents an *approximation* technique.

3.1 The Analogy Principle

The analogy principle is the conceptual framework that is always invoked when using large sample approximations. This principle is extremely intuitive: Suppose we know some properties that are satisfied for the true parameter in the population. If we can find a parameter value in the sample that causes the sample to mimic the properties of the population, we might use this parameter value to estimate the true parameter. As

a concrete example, we know that the population linear predictor is defined as, under regularity conditions,

$$\beta = \mathrm{E}(\mathbf{x_i}\mathbf{x_i}^\top)^{-1}\,\mathrm{E}(\mathbf{x_i}y_i) \tag{8}$$

Given the availability of a random sample, following the analogy principle, we can replace expectations with sample averages as follows

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\mathbf{x_i}^\top\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}y_i\right)$$

Notice that this estimator coincides exactly with the OLS estimator defined previously. Consequently, via the analogy principle, we just introduced yet another way to define the OLS estimator.

3.2 Setup

In this framework (which is sometimes defined as the Neoclassical Regression Model) we impose the following assumptions.

1. $\mathbb{E}[Y^4] < \infty$
2. $\mathbb{E}\|X\|^4 < \infty$
3. $\mathbb{E}[(a'X)^2] > 0,\ \forall\ a \in \mathbb{R}^K,\ a \neq 0$
4. $\{y_i, \mathbf{x}_i^\top\}$ are all i.i.d random variables.

Notice that the last three assumption are remarkably similar to the assumptions that we have introduced when defining the best linear predictor coefficient. We augment these assumptions because to derive the asymptotic approximation in the subsequent parts we need some slightly stronger conditions which are the ones spelled out above. Recall that our linear prediction equation is given by

$$y_i = \mathbf{x}_i^\top\beta + u_i \tag{9}$$

where β was defined as $\beta = \mathrm{E}(\mathbf{x_i}\mathbf{x_i}^\top)^{-1}\,\mathrm{E}(\mathbf{x_i}y_i)$. Notice that in this context.

3.3 Consistency of Least-Squares

Recall the definition of consistency.

Definition: (Convergence in Probability): A random variable z_N converges in probability to z (where z is non random) if for any $\delta > 0$

$$\lim_{n\to\infty} \Pr(|z_n - z| \geq \delta) = 0 \tag{10}$$

We denote convergence in probability as $z_n \xrightarrow{p} z$. The tool that we are going to use to prove consistency of the least squares is the Weak Law of Large Numbers.

Theorem (Weak Law of Large Numbers): Given a set of $i.i.d$ random variable $z_1, z_2, \ldots z_N$ with $\mathrm{E}(z_i^4) < \infty$

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \xrightarrow{p} \mathrm{E}(z_i) \tag{11}$$

This result can be easily proven, given our set of assumptions, by the Chebyshev's Inequality, as we have seen in class. There is also another important theorem that we shall introduce without proof

Theorem (Slutsky Theorem, part 1): If $z_n \xrightarrow{p} z$ and $g(.)$ is a continuous function at z_N, then

$$g(z_N) \xrightarrow{p} g(z) \tag{12}$$

Now let's get back to our estimator $\hat{\beta}$ and using the linear prediction equation (8) we can write

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}\mathbf{x_i}^\top\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}y_i\right) = \beta + \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}\mathbf{x_i}^\top\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}u_i\right) \tag{13}$$

Now notice that by WLLN + Slutsky Theorem + Assumption 2 + Assumption 3

$$\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}\mathbf{x_i}^\top\right)^{-1} \xrightarrow{p} \mathrm{E}(\mathbf{x_i}\mathbf{x_i}^\top)^{-1} \tag{14}$$

Similarly, by WLLN + Assumption 1 + equation (8)

$$\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}u_i\right) \xrightarrow{p} \mathrm{E}(\mathbf{x_i}u_i) = 0 \tag{15}$$

Combining these last two results and applying again the Slutsky theorem we have that

$$\hat{\beta} = \beta + \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i} \mathbf{x_i}^\top \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i} u_i \right) \xrightarrow{p} \beta \tag{16}$$

Therefore $\hat{\beta}$ is a consistent estimator for the population projection coefficient β!

3.4 Asymptotic Distribution

Notice that the WLLN is a useful tool, but it does not give an approximation to the distribution of an estimator. In order to do that we need to define the concept of convergence in distribution

Definition: (Convergence in Distribution) Let z_n be a random vector with distribution $F_n(u) = \Pr(z_n \leq u)$. We say that z_n converges in distribution to z as $N \to \infty$, denoted $z_n \xrightarrow{d} z$, if for all u at which $F(u) = Pr(z \leq u)$ is continuous

$$\lim_{N \to \infty} F_n(u) = F(u) \tag{17}$$

The tool that we are going to use to prove convergence in distribution of least squares is the Central Limit Theorem

Theorem: (Lindeberg-Levy Central Limit Theorem) Given a set of $i.i.d$ random variable $z_1, z_2, \ldots z_N$ with $\mathrm{E}(z_i^2) < \infty$ then

$$\sqrt{N}(\bar{z} - \mathrm{E}(z_i)) \xrightarrow{d} \mathcal{N}(0; \mathrm{Var}(z_i)) \tag{18}$$

We are also going to use the following important result

Theorem: (Slutsky Theorem, part 2): If $z_n \xrightarrow{d} z$ and $w_n \xrightarrow{p} w$ as $N \to \infty$ then

$$z_n + w_n \xrightarrow{d} z + w \tag{19}$$

$$z_n w_n \xrightarrow{d} zw \tag{20}$$

$$\frac{z_n}{w_n} \xrightarrow{d} \frac{z}{w}; \text{whenever } w \neq 0 \tag{21}$$

Let's now apply these concepts. Using our previous decomposition, we can write

$$\hat{\beta}-\beta = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\mathbf{x_i}^{\top}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}u_i\right) \rightarrow \sqrt{N}(\hat{\beta}-\beta) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\mathbf{x_i}^{\top}\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x_i}u_i\right)$$

Focus on the numerator. Let $z_i = \mathbf{x}_i u_i$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x_i}u_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}z_i = \sqrt{N}\bar{z} \tag{22}$$

Notice that we know that $\mathrm{E}(z_i) = 0$. Moreover,

$$\mathrm{Var}(z_i) = \mathrm{Var}(\mathbf{x}_i u_i) = \mathrm{E}(u_i^2\mathbf{x}_i\mathbf{x}_i^{\top}) \tag{23}$$

Since we did not impose any assumptions about homoskedasticity we cannot further decompose this term. However, we know that is bounded given our regularity conditions, therefore we can apply the CLT and see that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x_i}u_i \xrightarrow{d} \mathcal{N}(0;\mathrm{E}(u_i^2\mathbf{x}_i\mathbf{x}_i^{\top})) \tag{24}$$

Finally, using Slustky theorem plus our previous results regarding the probability limit of $((1/N)\sum_{i=1}^{N}\mathbf{x_i}\mathbf{x_i}^{\top})^{-1}$

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta) = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\mathbf{x_i}^{\top}\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x_i}u_i\right) &\xrightarrow{d} \mathrm{E}(\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}\times\mathcal{N}\left[0;\mathrm{E}(u_i^2\mathbf{x}_i\mathbf{x}_i^{\top})\right] \\ &= \mathcal{N}\left[0;\mathrm{E}(\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}\,\mathrm{E}(u_i^2\mathbf{x}_i\mathbf{x}_i^{\top})\,\mathrm{E}(\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}\right]\end{aligned}$$

This result states that the sampling distribution of the least-squares estimator, after rescaling, is approximately normal when the sample size N is sufficiently large. This holds true for all joint distributions of $(y_i;\mathbf{x}_i)$ which satisfy the assumptions written at the beginning, and therefore it is broadly applicable. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of $\hat{\beta}$.

Notice that the asymptotic variance of $\hat{\beta}$ is a function of unknown moments of the joint distribution. One can simplify this expression sensibly by introducing in the setup an homoskedasticity assumption. Nevertheless, by using again the analogy principle, we can use sample analogs to estimate the asymptotic variance of $\hat{\beta}$ under the very general conditions that we have been using so far, that is

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i-\mathbf{x}_i^{\top}\hat{\beta})^2\mathbf{x}_i\mathbf{x}_i^{\top}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top}\right)^{-1} \xrightarrow{p} \mathrm{E}(\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}\,\mathrm{E}(u_i^2\mathbf{x}_i\mathbf{x}_i^{\top})\,\mathrm{E}(\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}$$

The LHS of this expression is called the White Formula and as you can see is a robust estimator for the variance $\hat{\beta}$ under our very general setup which does not rule out heteroskedasticity, i.e. it does not rule out the possibility that the variance of the prediction error, u_i, depends on the the values of $\mathbf{x}_i$.

4 Hypothesis Testing

The problem we face is to determine if a linear restriction of the form $\mathrm{R}\hat{\beta} = \mathrm{r}$ is consistent with the data. Depending on the number of restrictions given by the dimension of R we will either conduct a t-test or Wald-test. The general testing procedure is always the same and consists of several components:

1. two competing hypotheses, a favored "null" hypothesis and an alternative hypothesis
2. a test statistic, with a distribution known under the null hypothesis
3. a significance level, the tolerable probability of mistakenly rejecting the null hypothesis when it is correct
4. a critical region, the values of the test statistic deemed adverse to the null hypothesis

Remember that the test statistic falls into the critical region with probability equal to the significance level under the null hypothesis.

4.1 t-Statistic

Testing a single restriction on $\hat{\beta}$ is an important special case. A common use of the t-test occurs under the hypothesis that one of the elements of $\hat{\beta}$ is zero. For notational ease, let β_1 be that element of $\hat{\beta}$. Partitioning X conformly, X_1 is the first column of X. We have now the following null hypothesis:

$$H_0 : \beta_1 = r$$

$$H_1 : \beta_1 \neq r$$

Let $\hat{\sigma}_1^2$ be an unbiased estimator of the variance of β_1, that is we take the first diagonal element of $\hat{\sigma}^2(X^\top X)^{-1}$. The t-statistic is given by:

$$\hat{t} = \frac{\hat{\beta}_1 - r}{\hat{\sigma}_1}$$

Under the null hypothesis $\hat{t}$ has a t_{N-K} distribution. One can also consider a one-sided alternative hypothesis as H1 : $\beta_1 > r$. In this case the acceptance interval is $\{\hat{t} \in \mathbb{R} | \hat{t} \leq t_{N-K,1-\alpha}\}$. The acceptance interval for a two-sided test is given by s $\{\hat{t} \in \mathbb{R}|$ $|\hat{t}| \leq t_{N-K,1-\frac{\alpha}{2}}\}$.

Test statistics are unit free. Furthermore, a statistically significant test statistic does not imply aqualitatively large difference between the hypothesized value and the unrestricted estimator. Nor does a statistically insignificant test statistic imply a qualitatively small difference. A Stata output shows a probability value for every regressor and the probability value for the F-statistic. In significance testing, the probability value (sometimes called the p-value) is the probability of obtaining a statistic as different or more different from the parameter specifed in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming the null hypothesis is true. The lower the probability value, the stronger the evidence that the null hypothesis is false. Traditionally, the null hypothesis is rejected if the probability value is below 0.05. The probability values for the regressors are the smallest values of α of the set ($\alpha \in \mathbb{R}^+ | \hat{t} \geq t_{N_K,1-\alpha}$) under the hypothesis $b_i = 0$. In words, the probability value gives us the smallest probability for the error we make if we reject the hypothesis that our regressor is irrelevant.

4.2 Wald Statistic

In we are interested in more than only a single restriction, we can test multiple restructions using the matrix R, which is a full-rank (K - J) $\times$ K matrix in which J $\leq$. Let our hypotheses be defined as follows:

$$H_0 : R\beta = \beta_0$$

$$H_1 : R\beta \neq \beta_0$$

which for $\theta_0 = R\beta_0$ and $\hat{\theta} = R\hat{\beta}$ as:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

The so-called Wald-statistics is given by:

$$W_0 = N(\hat{\theta} - \theta_0)^\top (R\Lambda R^\top)^{-1}(\hat{\theta} - \theta)$$

which has a χ^2_{K-J} distribution under H_0. The critical region, that is the values for which we reject H_0, is $(W|W > \chi^2_{K-J,1-\alpha})$.