LECTURE NOTE 2: MEANS, T-TESTS, AND REGRESSIONS

1 Introduction

This lecture note reviews basic tools you learned in ECON 120A/B and highlights the connections among them. We start with standard procedures to estimate means and variances, followed by one- and two-sample t-tests, and finally regression. We will see that all of these tools can be implemented with regression.

2 Mean and Variance

2.1 Sample Average

By now, you have encountered estimators for the mean many times in your academic career. The most common is the sample average:

$$\hat{\mu}_X = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Here are three attractive properties of the sample average:

1. BLUE: $\bar{X}$ is the *best linear unbiased estimator*. That is to say, in the class of linear estimators, $\hat{\mu} = a_1X_1 + a_2X_2 + \cdots + a_NX_N$ for some constants $a_1, a_2, \cdots, a_N$, setting $a_i = \frac{1}{N}$ for all i attains the most precise (i.e., "best") unbiased estimator.
2. Law of large numbers: $\bar{X} \xrightarrow{p} \mu_X$. In words: the sample average converges in probability to the population mean. Thus, in addition to being unbiased and precise, the sample average is also consistent.
3. Central limit theorem: As N approaches infinity, the distribution of $\bar{X}$ approaches a normal distribution with $E[\bar{X}] = \mu_X$ and $V[\bar{X}] = \frac{1}{N}\sigma_X^2$. As a result, in large samples, $\frac{\sqrt{N}}{\sigma_X}\left(\bar{X} - \mu_X\right)$ is approximately distributed $\mathcal{N}(0,1)$. The definition of "large sample" varies, but a common rule of thumb cutoff is a sample size of 30.

We can verify that the sample average is an unbiased estimator:

$$E[\bar{X}] = E[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \frac{1}{N}N\mu_X = \mu_X$$

The second equality holds because we can pull a constant out of an expectation. The third equality holds because X_i are *identically distributed*, so that $E[X_i] = \mu_X$ for all i.

We can also derive the variance of the sample average:

$$V[\bar{X}] = V[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N^2}\sum_{i=1}^{N} V[X_i] = \frac{1}{N^2} N\sigma_X^2 = \frac{1}{N}\sigma_X^2$$

The second equality holds because when we pull a constant out of a variance, we square it, and because X_i are *independent*, so that $V[X_1 + X_2 + \cdots + X_N] = V[X_1] + V[X_2] + \cdots + V[X_N]$, with no covariance terms. The third equality holds because X_i are *identically distributed*, so that $V[X_i] = \sigma_X^2$ for all i.

2.2 Sample Variance

By now, you have also encountered estimators for the variance many times in your academic career. But here too, we will do it once more to highlight two points: the *degrees of freedom adjustment* and the *bias-variance tradeoff*.

We know that $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the best linear unbiased estimator for $E[X]$. By analogy, we might be inclined to use $\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2$ as an estimator ot the variance, since $V[X] = E\left[X^2\right] - (E[X])^2$. Let's check whether $\hat{\sigma}_X^2$ is unbiased:

$$\begin{aligned}
E\left[\hat{\sigma}_X^2\right] &= E\left[\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} E\left[X_i^2\right] - E\left[\bar{X}^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(V\left[X_i\right] + E\left[X_i\right]^2\right) - \left(V\left[\bar{X}\right] + E\left[\bar{X}\right]^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(\sigma_X^2 - \mu_X^2\right) - \left(\frac{\sigma_X^2}{N} - \mu_X^2\right) \\
&= \sigma_X^2\left(1 - \frac{1}{N}\right)
\end{aligned}$$

Thus, this estimator of the variance is biased. It is straightforward to show, however, that $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ is an unbiased estimator of σ_X^2; you can try this at home if you don't remember it from your introductory statistics class. The use of $N-1$ rather than N is called a *degrees of freedom adjustment*, and we need it because we estimated the mean in a first step. If we knew the true mean and did not need to estimate it, then we would not need the adjustment.

The unbiased estimator is not necessarily superior to the biased one above. Both estimators are consistent, and $\hat{\sigma}_X^2$ has a lower variance than the unbiased estimator. We have a *tradeoff between bias and variance*. In any case, the bias term shrinks to zero as the sample becomes large, a property known as *asymptotic unbiasedness*.

3 *t*-Test

A t-test assesses the location of a mean in a single sample or compares means across two samples. At its core, it involves computing a t-statistic:

$$t = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta}]}$$

where θ is the quantity of interest (a mean or a difference in means). θ_0 refers to the null hypothesis, which is often zero. $SE[\hat{\theta}]$ refers to the standard error of the estimated quantity of interest. Under the Central Limit Theorem, t has a standard normal distribution in large samples, so that $p = 2\left(1 - \Phi\left[t\right]\right)$ for a two-sided test of the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$.

3.1 One-Sample *t*-Test

The one-sample t-test involves inference on a single mean. Suppose we have a sample average $\bar{X}$ and wish to test whether it is different from 0. Then $\hat{\theta} = \bar{X}$ and $\theta_0 = 0$. We know from above that $V[\bar{X}] = \frac{1}{N}\sigma_X^2$, which implies that $SE[\bar{X}] = \frac{\sigma_X}{\sqrt{N}}$, the standard deviation of X divided by the root of the sample size. This standard error is defined in terms of population parameters, which we do not know. So to compute the t-statistic, we replace the population standard deviation σ_X with its sample estimate $\hat{\sigma}_X$. We then compute:

$$t = \frac{\bar{X}}{\hat{\sigma}_X/\sqrt{N}}$$

to test the null hypothesis that $\mu = 0$. If t is larger than 1.96 or smaller than -1.96 in a large sample, we conclude that is is significantly different from 0 at the 5% level.

3.2 Two-sample *t*-Test

Often, we are interested in comparing means across samples or populations: the two-sample t-test. For instance, suppose X represents hourly earnings, and we want to know the difference in means between white and Black men in the United States. We can write μ_W and μ_B for their respective means and σ_W^2 and σ_B^2 for their respective variances. The central limit theorem tells us that in large samples, $\bar{X}_W \sim \mathcal{N}[\mu_W, \sigma_W^2/N_W]$ and $\bar{X}_B \sim \mathcal{N}[\mu_B, \sigma_B^2/N_B]$. You'll recall from 120A that the sum (or difference) of two normally distributed random variables is itself normally distributed. (Technically, this statement is true if the variables are *jointly* normally distributed or independent, as is the case here.) Furthermore, we know that $V\left[\bar{X}_W - \bar{X}_B\right] = V\left[\bar{X}_W\right] + V\left[\bar{X}_B\right] - 2cov(\bar{X}_W, \bar{X}_B)$, but since W and B are independent samples, $cov(\bar{X}_W, \bar{X}_B) = 0$. Thus, in large samples:

$$\bar{X}_W - \bar{X}_B \sim \mathcal{N}\left[\mu_W - \mu_B, \frac{\sigma_W^2}{N_W} + \frac{\sigma_B^2}{N_B}\right]$$

Because the variances σ_W^2 and σ_B^2 are unknown, we can replace them with sample variances, $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$, as above. In this case, the standard error of $\bar{X}_W - \bar{X}_B$ is:

$$SE\left[\bar{X}_W - \bar{X}_B\right] = \sqrt{\frac{\hat{\sigma}_W^2}{N_W} + \frac{\hat{\sigma}_B^2}{N_B}}$$

Now we can perform hypothesis tests as before. For instance, suppose we want to test whether black and white men have equal earnings. Our null hypothesis is $\mu_W - \mu_B = 0$, and our alternative is $\mu_W - \mu_B \neq 0$. We calculate our t-statistic as $t = \frac{\bar{X}_W - \bar{X}_B}{SE\left[\bar{X}_W - \bar{X}_B\right]}$ and, if our sample is large, compute p-values based on critical values from the standard normal distribution.

3.3 Confidence Intervals

Until now, we have been concerned with *point estimation* and *hypothesis testing*. A *point estimator* gives a unique value (the *point estimate*) that approximates the population parameter. An alternative is *interval estimation*, which provides a range of values that contain the population parameter. One such interval estimator is the *confidence interval*. The confidence interval for a parameter θ at confidence level γ (say, 95%) will contain the θ in 100γ out of every 100 samples. Note that this concept bears a close relation to hypothesis testing. In particular, to determine whether we can reject the null hypothesis $\theta = \theta_0$ in favor of the alternative $\theta \neq \theta_0$ at significance level α, we can look to see whether the confidence interval with confidence level $1 - \alpha$ contains θ_0. In large samples, the confidence interval for θ is $\hat{\theta} \pm \Phi^{-1}[1 - \frac{\alpha}{2}]SE[\hat{\theta}]$, where $\Phi^{-1}[\cdot]$ is the inverse of the standard normal cumulative distribution function. When $\gamma = 0.95$ (i.e., $\alpha = 0.05$), $\Phi^{-1}[1 - \frac{\alpha}{2}] = 1.96$. For example, the confidence interval for μ_X is $\bar{X} \pm 1.96SE[\bar{X}] = \bar{X} \pm 1.96\hat{\sigma}_X/\sqrt{N}$.

4 Regression

You can think of the mean estimator and t-tests above as special cases of ordinary least squares (OLS) regression. Alternatively, you can think of OLS regression as a generalization of estimating means.

4.1 OLS Estimator

We observe two variables, Y and X, and seek to estimate a linear function, $\hat{Y} = b_0 + b_1 X$, that provides a best fit. Several definitions of "best fit" exist, but we focus on minimizing the mean squared error. The "error" measures how far Y deviates from the value predicted by the function: $U = Y - \hat{Y} = Y - b_0 - b_1 X$. We will find the b_0 and b_1 that solve:

$$\min_{b_0, b_1} E\left[(Y - b_0 - b_1 X)^2\right] \tag{1}$$

which has the empirical analogue:

$$\min_{\hat{b}_0,\hat{b}_1} \sum_{i=1}^{N} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 \tag{2}$$

The quantity inside the summation in equation (2), $U_i = Y_i - \hat{Y}_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$, is called the "residual," and optimization seeks to minimize the sum of squared residuals. If we are careful with our language, "errors" measure deviations from the true function, while "residuals" measure deviations from the estimated function. People often mix up the terms, however.

These minimization problems underlie ordinary least squares (OLS) regression. We will denote the solutions to minimization problem (1) β_0 and β_1, and we will denote the solutions to minimization problem (2) $\hat{\beta}_0$ and $\hat{\beta}_1$.[1] Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators for the parameters of the statistical model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

OLS estimates do not necessarily have a causal interpretation. They express Y as a linear function of X. Only with additional assumptions can we say anything about the causal effect of X on Y.

In minimization problem (1), the first order condition for b_1 is $E\left[-2X\left(Y - \beta_0 - \beta_1 X\right)\right] = 0$, and the first order condition for b_0 is $E\left[-2\left(Y - \beta_0 - \beta_1 X\right)\right] = 0$. We can solve to obtain:

$$\begin{aligned} \beta_0 &= E[Y] - \beta_1 E[X] \\ \beta_1 &= \frac{cov(Y,X)}{V[X]} \end{aligned}$$

We will estimate β_0 and β_1 using their empirical analogues:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\frac{1}{N}\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{N}\sum_i (X_i - \bar{X})^2} \end{aligned}$$

In your spare time, you can check that these expressions satisfy the first order conditions of minimization problem (2): $\min_{\hat{b}_0,\hat{b}_1} \sum_i \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2$.

4.2 Gauss-Markov Theorem

The Gauss-Markov Theorem proves the estimator's optimality under certain conditions. In the classical setup for the theorem, X_i is considered fixed (i.e., non-random). This assumption is easily relaxed, and indeed we will relax it in Lecture Note 4. But it simplifies the mathematical expressions slightly, so we will

[1] So $b_0^* = \beta_0$, $b_1^* = \beta_1$, $\hat{b}_0^* = \hat{\beta}_0$, and $\hat{b}_1^* = \hat{\beta}_1$

maintain it here. The result of this assumption is that U_i is the only random variable that concerns us. The Gauss-Markov assumptions are as follows:

1. $E[U_i] = 0$ for all i
2. $V[U_i] = \sigma^2$ for all i
3. $cov(U_i, U_j) = 0$ for all $i \neq j$

These assumptions are also known as the Gauss-Markov assumptions, after the mathematicians who proved:

Gauss-Markov Theorem: *Under assumptions (1)-(3), the ordinary least squares estimator is the best linear unbiased estimator (BLUE).*

The word "best" means that the estimator is "efficient" or "minimum variance." The word "linear" means that the estimator is a linear function of $Y_1, \cdots, Y_N$. The word "unbiased" means that $E\left[\left(\hat{\beta}_0, \hat{\beta}_1\right)\right] = (\beta_0, \beta_1)$. Thus, a BLUE gives us the most precise answer that can be attained with a linear estimator. The theorem is useful because it provides conditions under which OLS is optimal. These conditions often do not hold, but the theorem helps us understand the consequences of these violations and how to fix them.

5 Equivalencies

The sample average and the t-test are embedded in OLS regression. To see this point, let us relabel hourly earnings as Y instead of X because it is more common to put Y on the left-hand side of the regression.

A regression of Y_i on a constant, with no covariates:

$$Y_i = \beta_0 + U_i$$

estimates the overall mean of Y_i and implements a one-sample t-test. The estimate $\hat{\beta}_0$ is the same as the sample average $\bar{Y} = \frac{1}{N}\sum_i Y_i$, and the standard error on $\hat{\beta}_0$ is the same as $SE[\bar{Y}] = \hat{\sigma}x/\sqrt{N}$. Thus, the t-statistic $t = \hat{\beta}_0/SE[\hat{\beta}_0]$ tests whether $\bar{Y}$ is different from zero.

In a sample of white and Black individuals, a regression of Y_i on a constant and an indicator for being Black:

$$Y_i = \beta_0 + \beta_1 BLACK_i + U_i$$

estimates the Black-white difference in the mean of Y_i and implements a two-sample t-test. The estimate $\hat{\beta}_0$ is the sample average $\bar{Y}_W$, while the estimate $\hat{\beta}_1$ is difference in sample averages $\bar{Y}_B - \bar{Y}_W$. Thus, the t-statistic $t = \hat{\beta}_1/SE[\hat{\beta}_1]$ tests whether $\bar{Y}_W$ is different from $\bar{Y}_B$.