Econ 140 - Spring 2016
Section 5

GSI: Fenella Carpena

February 18, 2016

This section note reviews Stock and Watson Chapter 5 ("Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals"). This note is not meant to be a comprehensive review of the book, but I have tried to focus on the concepts which I think are most important.

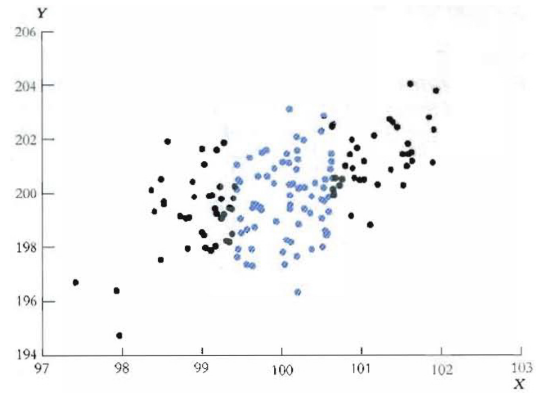
1 Hypothesis Testing for Regression Coefficients

- Let us now consider how to make inferences, using our sample, about the relationship between X and Y in the population. For instance, some questions we will try to answer in this section are:
  - Is the observed relationship between X and Y in the sample strong enough to conclude that it also holds in the population?
  - How can we use the sample statistics $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to determine a plausible range of values for β_0 and β_1 of the population regression line?
- Recall that earlier in the semester, we used the sample mean $\overline{X}$ to say something about the population mean μ. To do this, we needed to know the standard error of $\overline{X}$, that is $SE(\overline{X}) = s/\sqrt{n}$, so that we can construct the following: (1) confidence interval, (2) test statistic (e.g., t-stat), and (3) p-value. These are the same 3 tools that we will be using for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, so we will need to know $SE(\widehat{\beta}_0)$ and $SE(\widehat{\beta}_1)$ in this case as well.
- **Why do $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have standard errors? Aren't they just constants?** No, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are NOT constants. They are random variables. We already discussed this last week, but I repeat it again here since it is very important to understand this point. There is variability in $\widehat{\beta}_0$ and $\widehat{\beta}_1$ because if we draw a different sample, we would get a different value of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The standard errors describe this sample-to-sample variability of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
- In a regression with a single regressor, the formula for (homoskedastic) standard errors of $\widehat{\beta}_1$ is

$$SE(\widehat{\beta}_1) = \frac{s_{\widehat{u}}}{\sqrt{n-1}} \cdot \frac{1}{s_X}$$

  where $s_{\widehat{u}}$ is the standard error of the regression (covered in the last section). n is the sample size, and s_X is the sample standard deviation of X. **Note that a smaller $SE(\widehat{\beta}_1)$ means that our estimate of the population β_1 is more precise.** Here are several important features of $SE(\widehat{\beta}_1)$:
  - **As $s_{\widehat{u}}$ falls, $SE(\widehat{\beta}_1)$ falls.** This is because $s_{\widehat{u}}$ is our estimate for the standard deviation of the population errors u, and if this standard deviation is small, then the data will have a tighter scatter around the population regression line. Hence, its slope will also be estimated more precisely.
  - **As n increases, $SE(\widehat{\beta}_1)$ falls.** That is, when our sample size is large, our estimate $\widehat{\beta}_1$ will be close the true population coefficient β_1 with high probability. This is because the standard deviation of $\widehat{\beta}_1$ decreases to zero as n increases, so the distribution of the $\widehat{\beta}_1$ will be tightly centered around its mean β_1 when n is large. Another way to think about this is that the larger the sample, the more information we have about the population, so the more precise our estimate for β_1 is.

- **As s_X increases, $SE(\widehat{\beta}_1)$ falls.** To get a better sense of why this is true, consider the figure below which presents a scatterplot of 150 artificial data points on X and Y. Suppose you were asked to draw a line as accurate as possible through *either* the blue dots or the black dots—which would you choose? It would be easier to draw a precise line through the black dots, which have a larger variance than the blue dots. Similarly, the larger the variance of X, the more precise is $\widehat{\beta}_1$.



- There is also a formula for $SE(\widehat{\beta}_0)$ which you can find in the textbook, Appendix 5.1. However I will not focus on it here since in most applications, we are likely to be more interested in $\widehat{\beta}_1$ than $\widehat{\beta}_0$ (e.g., we often want to know how Y changes when X changes).

- In practice, statistics software such as Stata calculate $SE(\widehat{\beta}_0)$ and $SE(\widehat{\beta}_1)$ automatically, so they are provided to you in the output of the regression table.

1.1 Hypothesis Tests Using t-statistic

- **Step 1:** Construct a t-statistic. As before, the general formula for the t-statistic that we learned in Section 3 Notes also applies when conducting hypothesis tests for β_0 and β_1. That general formula is:

$$t\text{-}stat = \frac{sample\ statistic - hypothesized\ value}{se(sample\ statistic)}.$$

For example, if $\widehat{\beta}_0 = 1.86534$, $SE(\widehat{\beta}_0) = 0.40083$, and the null and alternative hypotheses are $H_0 : \beta_0 = 2$, $H_a : \beta_0 \neq 2$ (a two-sided hypothesis test), then

$$t\text{-}stat = \frac{1.86534 - 2}{0.40083} = -0.336.$$

- **Step 2:** Find the critical value. This critical value depends on: (1) whether we have a two-sided or one-sided test, and (2) the significance level α (given in the problem or chosen by you).
  - One-sided test where H_a has the format *population parameter > "hypothesized value"* (e.g. $H_a : \beta_1 > 4$): The one-sided critical value is z_α from the Normal table such that $P(Z > z_\alpha) = \alpha$. In words, this means that we are looking for the value z in the Normal table where the area of the PDF to the **right** of z is equal to α.
  - One-sided test where H_a has the format *population parameter < "hypothesized value"* (e.g. $H_a : \beta_1 < 2$): The one-sided critical value is $-z_\alpha$ from the Normal table such that $P(Z < -z_\alpha) = \alpha$. In words, this means that we are looking for the value $-z$ in the Normal table where the area of the PDF to the **left** of $-z$ is equal to α.
  - Two-sided test: The two-sided critical value is $z_{\alpha/2}$ from the Normal table such that $P(|Z| > z_{\alpha/2}) = \alpha$. In words, this means that we are looking for the value z in the Normal table where the area of the PDF to the right of z plus the area to the left of $-z$ sums up to α.
  - Note: A two-sided test has a format like $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$ (that is, H_a has a "not equal to" sign), while a one- sided test has a format like $H_0 : \beta_1 = 0$, $H_a : \beta_1 < 0$.

- **Step 3:** Compare the t-statistic with the critical value from the previous step, and apply the following rejection rule.

- One-sided test where H_a has the format *population parameter* $>$ *"hypothesized value"* (e.g. H_a : $\beta_1 > 0$): **Reject** H_0 if *t-stat* $>$ *one-sided critical value.* Otherwise, we **fail to reject** H_0.
- One-sided test where H_a has the format *population parameter* $<$ *"hypothesized value"* (e.g. H_a : $\beta_1 < 0$): **Reject** H_0 if *t-stat* $<$ *one-sided critical value.* Otherwise, we **fail to reject** H_0.
- Two-sided test: **Reject** H_0 if $|$*t-stat*$|$ $>$ *two-sided critical value.* Otherwise, we **fail to reject** H_0**.**

1.2 Hypothesis Test: Using p-value

- **Step 1:** Construct the t-statistic as in the previous section.
- **Step 2:** Calculate the p-value, depending on whether we have a two-sided or one-sided test.
  - One-sided test where H_a has the format *population parameter* $>$ *"hypothesized value"* (e.g. H_a : $\beta_1 > 0$): $p\text{-}value = P(Z > t\text{-}stat)$
  - One-sided test where H_a has the format *population parameter* $<$ *"hypothesized value"* (e.g. H_a : $\beta_1 < 0$): $p\text{-}value = P(Z < t\text{-}stat)$
  - Two-sided test: $p\text{-}value = P(|Z| > |t\text{-}stat|) = 2 \cdot P(Z > |t\text{-}stat|)$. In words, this means that we need to take the absolute value of the t-stat, find the area to the right of the absolute value of the t-stat, and multiply that area by 2 to find the p-value.
- **Step 3:** Compare the p-value obtained in the previous step with the level of α. We **reject** H_0 if p-value $< \alpha$. Otherwise, we **fail to reject** H_0.

1.3 Confidence Intervals

- The general formula for the confidence interval that we learned previously for $\overline{X}$ also applies to the case where we are constructing confidence intervals for β_0 and β_1. That general formula is:

$$\textit{sample statistic} \pm \textit{two-sided critical value} * \textit{se(sample statistic)}$$

- The $(100 - \alpha)$**% confidence interval for the slope** β_1 is given by:

$$\widehat{\beta}_1 \pm z_{\alpha/2} * SE(\widehat{\beta}_1)$$

  where α is the significance level (given in the problem or chosen by you), $z_{\alpha/2}$ is the two-sided critical value from the Normal table (found in the same way as in Section 1.1 of these notes), and n is the sample size. Note that the confidence interval for $\widehat{\beta}_0$ follows the same formula as above, but using $\widehat{\beta}_0$ instead of $\widehat{\beta}_1$.

- **How do we interpret the CI for** β_1**?** The interpretation is the same as what we have seen in Section 2 Notes. For example, if $\alpha = 0.05$, then we have a 95% CI. This means that "we are 95% confident that the true population parameter β_1 lies between that interval."

2 Dummy Variables

- So far we have been talking about X variables that are continuous (e.g., class size, age, etc.). However, regressions can also be used when X is binary, that is, when it can take on only two values (0 and 1). Such a variable is called **dummy variable** (also known as binary variable or indicator variable).
- To fix ideas, let us consider the following regression

$$\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot MBA$$

  where *wage* is measured annually in dollars, and MBA is a dummy variable equal to 1 if a worker has an MBA, and 0 otherwise.

- In the above regression, MBA is not continuous, so it is not useful to think of $\widehat{\beta}_1$ as a slope. In this case, **how should we interpret the coefficients?** We can easily do so by looking at the two cases when $MBA = 1$ or when $MBA = 0$. Specifically, we note the following.

- If $MBA = 0$, $\widehat{wage} = \widehat{\beta}_0$.
- If $MBA = 1$, $\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1$.
- Using the above two equations, we see that:
  - $\widehat{\beta}_0$ is the average wage of workers without an MBA.
  - $\widehat{\beta}_0 + \widehat{\beta}_1$ is the sample average wage of workers with an MBA.
  - $\widehat{\beta}_1$ is the difference in the sample average wage between workers with an MBA and workers without an MBA.

- Note that a hypothesis test for a difference in means between two groups can be carried out by regressing Y on a dummy variable X which defines groups. For example, in the regression of wages on MBA shown above, the hypothesis test $H_0 : \mu_{MBA} - \mu_{NonMBA} = 0$ vs $H_1 : \mu_{MBA} - \mu_{NonMBA} \neq 0$ is equivalent to testing $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

3 Homoskedasticity and Heteroskedasticity

- We say that the population error term u_i is **homoskedastic** if $var(u_i|X_i)$ is constant for all i. Otherwise, we say that u_i is **heteroskedastic.**

- To understand what heteroskedasticity means, let's take a look at the following regression of *wages* on *schooling* (i.e., years of education)

$$wages_i = \beta_0 + \beta_1 schooling_i + u_i.$$

  Note that $var(u_i|schooling_i) = var(wages_i - \beta_0 - \beta_1 schooling_i|schooling_i) = var(wages_i|schooling_i)$. Hence, this means that:

  - If u_i is homoskedastic, both $var(u_i|schooling_i)$ and $var(wages_i|schooling_i)$ are constant.
  - In particular, homoskedasticity would imply that $var(wages_i|schooling_i)$ is the same for all schooling levels. Another way of saying this is that the variability of wages around its mean is the same regardless of educational attainment.
  - Homoskedasticity is not realistic in this case because it is likely that people with more education have wider job opportunities, which could lead to more variability in wages. In contrast, people with low education levels have fewer opportunities and probably work minimum wage jobs, so there is less dispersion of wages among the uneducated.
  - In sum, we would expect that variability in wages is higher for the highly educated, and the variability in wages is low for those with low levels of schooling. Therefore, in this example, the errors u_i are likely heteroskedastic.

- **Why do we care about heteroskedasticity and homoskedasticity?** I can think of two reasons why we care.

  - If the 3 least squares assumptions hold **AND** u_i are homoskedastic, then the OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are BLUE (Best Linear Unbiased Estimators). This is the **Gauss-Markov Theorem.**
  - Heteroskedasticity and homoskedasticity have implications for calculating $SE(\widehat{\beta}_1)$ and $SE(\widehat{\beta}_0)$. For example, if the population errors are heteroskedastic but you use homoskedastic SEs, your hypothesis tests and confidence intervals will be invalid. However, since homoskedasticity is a special case of heteroskedasticity, the heteroskedastic-robust SEs will still be valid under homoskedasticity (see table below for a summary of the different cases). Hence, **a typical rule of thumb is to always use heteroskedasticity-robust SEs**. This is the "robust" option when running a regression in Stata.

		What you actually use	
		Homoskedastic	Heteroskedastic
Truth about u_i	Homoskedastic	VALID hyp. test and CI	VALID hyp. test and CI
	Heteroskedastic	INVALID hyp. test and CI	VALID hyp. test and CI