Econ 140 - Spring 2016
Section 5

GSI: Fenella Carpena

February 18, 2016

1 Hypothesis Testing for Regression Coefficients

Exercise 1.1. (Adapted from Stock & Watson, Exercise 5.1) Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression

$$\widehat{TestScore} = \underset{(20.4)}{520.4} - \underset{(2.21)}{5.82} * CS, \quad R^2 = 0.08, \quad SER = 11.5$$

(a) Test $H_0 : \beta_1 = -1.5$ vs. $H_a : \beta_1 < -1.5$ at 5% significance level using a t-statistic.

- t-stat $= (-5.82 + 1.5)/2.21 = -1.95$
- One-sided critical value $= -1.64$
- We reject H_0 at the 5% level since $-1.95 < -1.64$

(b) Calculate the p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$. Do you reject the null hypothesis at the 5% level? At the 1% level?

- t-stat $= (-5.82 - 0)/2.21 = -2.63$
- p-value $= 2 \cdot P(Z > |-2.63|) = 2 \cdot 0.0043 = 0.0086$
- We reject H_0 at the 5% level since $0.86\% < 5\%$
- We reject H_0 at the 1% level since $0.86\% < 1\%$

(c) Calculate the p-value for the two-sided test of the null hypothesis $H_0 : \beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1.

- t-stat $= (-5.82 + 5.6)/2.21 = -0.1$
- p-value $= 2 \cdot P(Z > |-0.1|) = 2 \cdot 0.4602 = 0.9204$
- We fail to reject H_0 at the 5% level since $92.04\% > 5\%$. Hence -5.6 is contained the 95% confidence interval.

(d) Construct a 95% confidence interval for β_1.
$\widehat{\beta}_1 \pm 1.96 \cdot SE(\widehat{\beta}_1) = -5.82 \pm 1.96 \cdot 2.21 = (-10.152, -1.4884)$

(e) Construct a 99% confidence interval for β_0.
$\widehat{\beta}_0 \pm 2.58 \cdot SE(\widehat{\beta}_0) = 520.4 \pm 2.58 \cdot 20.4 = (467.7, 573.0)$

2 Dummy Variables

Exercise 2.1. Suppose that the Berkeley Undergraduate Program Office wishes to examine the potential educational benefits of providing students with an iPad. Using data from 141 students, the table below summarizes the mean and SD of the variable *collGPA* (college GPA) for students with and without an iPad.

	n	Mean	SD
Without iPad	85	2.989	0.321
With iPad	56	3.159	0.421

(a) Let $\overline{X}_1$ be the sample mean collage GPA of students with an iPad, and $\overline{X}_0$ be the sample mean collage GPA of students without an iPad. Calculate:

(i) $\overline{X}_1 - \overline{X}_0$
$3.159 - 2.989 = 0.17$

(ii) $SE(\overline{X}_1 - \overline{X}_0)$
$\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}} = \sqrt{\frac{0.421^2}{56} + \frac{0.321^2}{85}} = 0.06616$

(b) The program office does not plan to provide iPads to students unless there is evidence that having an iPad increases students' GPAs. What is the one-sided null and alternative hypothesis that the program office would like to test?
Let μ_1 bet the population mean GPA of students with an iPad. Let μ_0 be the population mean GPA of students without an iPad. The one-sided hypothesis we want to test is $H_0 : \mu_1 - \mu_0 = 0$, $H_1 : \mu_1 - \mu_0 > 0$.

(c) Carry out the test from part (b) at the 1% level. Do you reject or fail to reject the null hypothesis?
Using t-statistic:

- t-stat $= \frac{\overline{X}_1 - \overline{X}_0 - 0}{SE(\overline{X}_1 - \overline{X}_0)} = 0.17/0.06616 = 2.5695$
- One-sided critical value $= 2.33$
- We reject H_0 at the 1% level since $2.5695 > 2.33$

Using p-value:

- p-value $= P(Z > 2.569) = 0.005$
- We reject H_0 at the 1% level since $0.05\% < 1\%$

(d) Can the difference in the GPA between students with and without an iPad be attributed to iPad ownership alone? Explain why or why not.
No. For example, maybe students who have an iPad are also richer (because iPads are expensive), and richer students do better in school because they have more money and resources to spend on tutoring, etc. In this case, the difference in GPA between iPad owners and non-iPad owners could be due to wealth as well.

(e) Suppose that an analyst at the program office instead estimated a regression of college GPA on $HasIPAD$, a dummy variable equal to 1 if the student has an iPad, and 0 otherwise, obtaining the following regression results. Interpret each of the coefficients in the regression.

```
Linear regression                               Number of obs =       141
                                                F(  1,    139) =      6.57
                                                Prob > F      =    0.0114
                                                R-squared     =    0.0500
                                                Root MSE      =    .36419

------------------------------------------------------------------------------
             |               Robust
      colGPA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HasIPAD |   .1695168   .0661386     2.56   0.011      .038749    .3002846
       _cons |   2.989412   .0349104    85.63   0.000     2.920388    3.058436
------------------------------------------------------------------------------
```

- $\widehat{\beta}_0 = 2.989$. This is the sample average of students without an iPad.
- $\widehat{\beta}_0 = 0.1695$. The difference in the sample average GPA of students without an iPad. Those with an iPad have 0.170 higher GPA.

(f) Consider again the null and alternative hypothesis from part (b).

(i) How can you use the regression coefficients from part (e) to test the same null and alternative hypothesis as in part (b)?

(ii) Carry out the hypothesis test using the regression coefficients at the 1% level.

(iii) What does question (f)(i) and (f)(ii) tell us about how we can use regressions to carry out a test for the differences of means across two groups?

(i) We can test $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$.

(ii) The t-stat $= \widehat{\beta}_1 / SE(\widehat{\beta}_1) = 0.16951/0.0661 = 2.566$. The one-sided critical value is 2.33. Therefore we reject H_0 since 2.566 ¿ 2.33.

(iii) Note that the t-stat we get here is the same as the t-stat from part (c), save for rounding differences. Testing the difference of means $H_0 : \mu_1 - \mu_0 = 0$, $H_1 : \mu_1 - \mu_0 > 0$ is equivalent to testing $H_0 : \beta_1 = 0$, $H_1 : \beta_1 > 0$, where β_1 is the coefficient for $iPAD$, the variable that defined the 2 groups (iPad owners vs. non-iPad owners).

3 Heteroskedasticity and Homoskedasticity

Exercise 3.1. (Stock & Watson, Review the Concepts 5.3) Define *homoskedasticity* and *heteroskedasticity*. Provide a hypothetical empirical example in which you think the errors would be heteroskedastic and explain your reasoning.

- 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.**
- An example in which the errors are likely heteroskedastic is a regression of wages on schooling, as follows:

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

  - 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.

Exercise 3.2. (Stock & Watson, Exercise 5.5, 5.6) In the 1980s, Tennessee conducted an experiment in which kindergarten students were randomly assigned to "regular" and "small" classes and given standardized tests at the end of the year. (Regular classes contained approximately 24 students, and small classes contained approximately 15 students.) Suppose that, in the population, the standardized tests have a mean score of 925 points and a standard deviation of 75 points. Let $SmallClass$ denote a binary variable equal to 1 if the student is assigned to a small class, and 0 otherwise. A regression of $TestScore$ on $SmallClass$ yields

$$\widehat{TestScore} = \underset{(1.6)}{918.0} + \underset{(2.5)}{13.9} * SmallClass, \quad R^2 = 0.01, \quad SER = 74.6$$

(a) Do small classes improve test scores? By how much? Is this effect large? Explain.
The estimated gain from being in a small class is 13.9 points. This is a moderate increase, since 13.9 is about 1/5 of the standard deviation of test scores, which is 75.

(b) Is the estimated effect of a class size on test scores statistically significant? Carry out the test at the 5% level.
Here, we are considering $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. The t-stat $= 13.9/2.5 = 5.56$. The critical value is 1.96. Since $|5.56| > 1.96$, we reject H_0 at the 5% level. The estimated effect of class size on test scores is statistically significant.

(c) Construct a 99% confidence interval for the effect of $SmallClass$ on $TestScore$.
$13.9 \pm 2.58 \cdot 2.5 = 1.39 \pm 6.45$

(d) Do you think that the regression errors are plausibly homoskedastic? Explain.
The question asks whether the variability in test scores in large classes is the same as the variability in small classes. It is hard to say. On the one hand, teachers in small classes might be able to spend more time bringing all of the students along, reducing the poor performance of particularly unprepared students. On the other hand, most of the variability in test scores might be beyond the control of the teacher.

(e) $SE(\hat{\beta}_1)$ was computed using heteroskedastic standard errors. Suppose that the regression errors were homoskedastic: Would this affect the validity of the confidence interval constructed in part (c)? Explain.
The CI would still be valid. The SEs were computed using heteroskedastic-robust SEs, which are valid whether the true population regression errors are homoskedastic or heteroskedastic.