ECO220Y1Y, Test #4, Prof. Murdock: SOLUTIONS

March 6, 2020, 9:10 – 11:00 am

NOTE: The parts of the solutions [in brackets] are extra explanations and are not required parts of your answer.

(1) (a) $R^2 = \frac{SSR}{SST} = \frac{782410.694}{782410.694 + 1347480.86} = \frac{782410.694}{2129891.6} = 0.3673$

(b) $b_1 \pm t_{\alpha/2} s.e.(b_1) = 0.5407256 \pm 1.960 * s.e.(b_1)$

$UCL = 0.5598959 = 0.5407256 + 1.960 * s.e.(b_1) \overset{implies}{\Longrightarrow} s.e.(b_1) = 0.01$

Or alternatively, $LCL = 0.5215554 = 0.5407256 - 1.960 * s.e.(b_1) \overset{implies}{\Longrightarrow} s.e.(b_1) = 0.01$

(c) It's the P-value for the test of statistical significance of the slope coefficient: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$. Given the gigantic t test statistic of 38.98, the P-value is 0: the missing value is 0.000. In this recent study of Malawi school children, the scores of female students on the local language (Chichewa) test are a highly statistically significant predictor of their scores on the English language test: we can easily rule out these language scores being unrelated with each other for female students. [This result is significant overall because the size of the coefficient is big and hence economically significant.]

(2) (a) Use the results in Regression #3. Given that $R^2 = (r)^2$, the coefficient of correlation is 0.68 $\left(= \sqrt{0.4608}\right)$

(b) Among 137 countries included in the World Bank's database in 2014 excluding Iceland, which is an outlier, those that use an extra 1,000 kWh of energy per capita have GDP per capita that is on average about $4,232 USD higher (which is A LOT higher!). The intercept (constant term) of 475 has no interpretation in this context because no country in the world has zero energy consumption.

(c) Among 138 countries included in the World Bank's database in 2014, those that use 10 percent more energy on average have GDP per capita that is approximately 8.4 percent higher.

(d) The s_e measures the amount of scatter about the OLS line, which in this case measures how much trouble we have in predicting a country's GDP per capita knowing only its energy consumption per capita using a subset of 128 countries that exclude the top 10 energy-using countries in the world. The units are US dollars. Hence, it is $9,529, which is huge given that most countries have GDP per capita below $20,000: there is *a lot* of scatter around the OLS line. However, it is a misleading measure because there is an obvious issue with heteroscedasticity (caused by nonlinearity) – notice the fan shape of dots about the OLS line in Regression #7 – which means there is a lot of scatter for high energy consuming countries and much less for low energy consuming countries. In other words, we can make more accurate predictions for low energy consuming countries.

(3) (a)

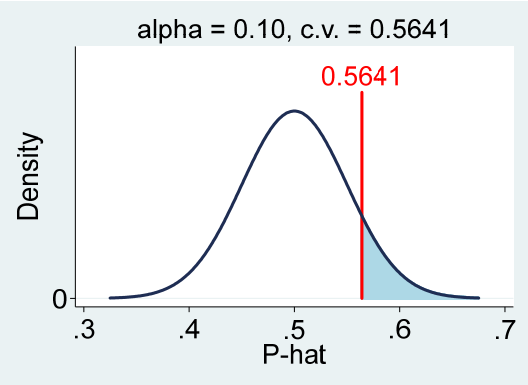
$H_0: p = 0.50$

$H_1: p > 0.50$

Find the rejection region: $P(Z > z_\alpha) = \alpha;\ P(Z > 1.282) = 0.10$

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \Rightarrow 1.282 = \frac{\hat{P} - 0.50}{\sqrt{\frac{0.50(1-0.50)}{100}}} = \frac{\hat{P} - 0.50}{0.05} \Rightarrow c.v. = 0.5641$$

The unstandardized rejection region is $(0.5641, \infty)$ or $(0.5641, 0)$. [In a random sample of 100 students, we need more than 56.4% thinking Trump is likely to win to prove that a majority the population thinks that at a 10% significance level.]

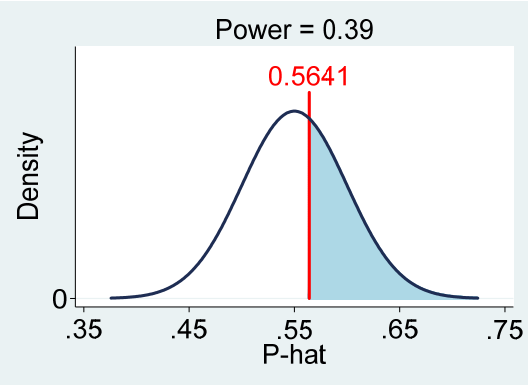


Power is the probability of BEING IN the rejection region: in other words, correctly rejecting the false null.

$Power = P(\hat{P} > 0.5641 \mid p = 0.55, n = 100) = ?$

$$Power = P(\hat{P} > 0.5641) = P\left(Z > \frac{0.5641 - 0.55}{\sqrt{\frac{0.55(1-0.55)}{100}}}\right) = P\left(Z > \frac{0.0141}{0.04975}\right) =$$

$$P(Z > 0.28) = 0.5 - 0.1103 = 0.39$$



(b) It means that even if *El Universal* is correct that a majority of Mexicans think Trump is likely to win, there is a 40% chance that it will not be able to prove it at a 5% significance level using a random sample of 1,000 people if 53% of all Mexicans think that.

(c) larger than, smaller than, larger than

(4) smaller than, larger than, smaller than, smaller than

(5)(a) Given that the question asks about the size of the difference, we must use an estimation approach and not hypothesis testing. Further, this requires an inference about the difference in means with independent samples when the variances are unequal. Going with the most common practice, we'll find a 95% confidence interval estimate of the difference. Given the huge sample sizes, the degrees of freedom are much greater than 1,000.

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(2421 - 1780) \pm 1.960\sqrt{\frac{759.7^2}{53{,}752} + \frac{1143.4^2}{11{,}702}}$$

$$(641) \pm 1.960\sqrt{\frac{577144.09}{53{,}752} + \frac{1307363.56}{11{,}702}}$$

$$641 \pm 1.960\sqrt{10.7372 + 111.7214}$$

$$641 \pm 1.960\sqrt{122.4586}$$

$$641 \pm 1.960 * 11.066$$

$$641 \pm 21.7$$

$LCL = 619$ square feet

$UCL = 663$ square feet

(b) The P-value is tiny (basically 0) and this means that there is a highly statistically significant difference between the mean number of times a house is sold in Austin versus outside Austin. However, to be significant requires that the difference is BOTH statistically significant AND big enough to care about (i.e. economically significant). There is very little difference in the mean number of times a house is sold: 1.61 in Austin and 1.68 outside Austin. [The only reason this small difference is statistically significant is because of the very large sample sizes.] In contrast, houses are much newer outside Austin: 15 years newer on average (=1987 – 1972). This big difference is also statistically significant, which means it is significant.