ECO220Y1Y, APRIL 2019, FINAL EXAM: SOLUTIONS

(1) (a)

Define Group 2 to be the *busses*: $\hat{P}_2 = \frac{88}{1{,}722} = 0.0511$

Define Group 1 to be the *subway*: $\hat{P}_1 = \frac{303+218}{9{,}342+4{,}626} = \frac{521}{13{,}968} = 0.0373$

The point estimate of the difference is $\left(\hat{P}_2 - \hat{P}_1\right) = 0.0138$. Next, obtain the CI estimate of the difference:

$$(\hat{P}_2 - \hat{P}_1) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$$

$$0.0138 \pm 1.645\sqrt{\frac{\frac{88}{1{,}722}\left(1-\frac{88}{1{,}722}\right)}{1{,}722} + \frac{\frac{521}{13{,}968}\left(1-\frac{521}{13{,}968}\right)}{13{,}968}}$$

$$0.0138 \pm 1.645 * 0.00554$$

$$0.0138 \pm 0.0091$$

The lower confidence limit (LCL) is 0.0047 and the upper confidence limit (UCL) is 0.0229.

We are 90% confident that among *all* TTC passengers the fare evasion rate is between 0.5 to 2.3 *percentage points higher* on busses compared to the subway, which is a notable difference in this context (especially given that the TTC incorrectly thought the total fare evasion rate was only 1.8%!). [**NOTE:** It <u>is valid</u> to do the entire analysis switching the definition of groups 1 and 2, so long as you are consistent: in other words, the interpretation would be how much *lower* the fare evasion rate is for the subway.]

(b)

$H_0: p = 0.144$

$H_1: p > 0.144$

$\hat{P} = \frac{603}{3{,}957} = 0.15239$

$$z = \frac{\hat{P}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.15239-0.144}{\sqrt{\frac{0.144(1-0.144)}{3{,}957}}} = \frac{0.00839}{0.00558} = 1.50$$

$$P - value = P(Z > 1.50) \approx 0.5 - 0.4332 = 0.07$$

Given this quite small P-value, at a 10% significance level we *do* have sufficient evidence to prove the shocking claim that the fare evasion rate on streetcars is in fact more than 8 times (!) what the TTC has claimed.

(2) Using the Normal approximation to the Binomial:

$$P(X > 35 \mid n = 737,\ p = 0.04) = P\left(Z > \frac{35-0.04*737}{\sqrt{737*0.04*0.96}}\right) = P\left(Z > \frac{5.52}{5.320}\right) = P(Z > 1.04) = 0.5 - 0.3508 = 0.15$$

Alternatively, with the continuity correction: $P\left(Z > \frac{35+0.5-0.04*737}{\sqrt{737*0.04*0.96}}\right) = P(Z > 1.13) = 0.5 - 0.3708 = 0.13$

Also, same answer if thing about $\hat{P}$ (instead of X): $P\left(\hat{P} > \frac{35}{737} \mid n = 737,\ p = 0.04\right) = P\left(Z > \frac{0.0475-0.04}{\sqrt{\frac{0.04*0.96}{737}}}\right) =$

$P(Z > 1.04) = 0.5 - 0.3508 = 0.15$

(3)

$H_0: p = 0.70$

$H_1: p < 0.70$

A Type II error would be the serious situation where the vaccination rate is really terrible – below 70% – but we do not have sufficient evidence in our sample to prove it. This would mean that Toronto is vulnerable to public health crisis: we should definitely be worried. The best way to improve the power (which is the complement of making a Type II error) is to increase the sample size to be much bigger than the planned 500 students born in 2010.

(4) (a) These data are time series data with 34 observations (=2017-1983) and at least two variables – fuel aridity and area burned – in addition to an identifier variable for year.

(b)

If area burned were in km^2 (not 1,000s of km^2), the value of the SST would **increase**, the value of the R^2 would **stay the same**, and the value of the s_e would **increase**.

If we dropped the observation for 2017, the value of the SST would **decrease**, the value of the R^2 would **increase**, and the value of the s_e would **decrease**.

(c) The value of b_0 is the predicted value of y when x is zero and, looking at the figure, that intercept is approximately 1.25. For years where the fuel-aridity index is average (i.e. at the 1981-2010 average) the predicted area burned in wildfires in the Western United States is about 1,250 square kilometers.

(5) $P(X < ?) = 0.20$

$P(Z < -0.84) = 0.20$

$\frac{? - 0.0112}{0.0094} = -0.84$

$? = 0.0033$, which is the value of the 20th percentile

(6) (a)

Consider The Indianapolis Museum of Art (IMA). Suppose Museum XX (MXX) has 500,000 objects and 45,000 are on display. Compared to P(Display | IMA), we can say that P(Display | MXX) is **about the same**.

Consider the entire collections (on display and in storage) of the Metropolitan Museum of Art (MET) and Denver Art Museum (DAM) combined. For one randomly selected piece from the combined collections, compared to P(MET | Not on paper), we can say that P(DAM | Not on paper) is **smaller**.

Compared to P(Not on paper | IMA), we can say that P(Not on paper | DAM) is **larger**.

[Explanation of smaller for second blank: While the DAM has a much higher fraction not on paper – P(Not on paper | DAM) ≈ 0.73 whereas P(Not on paper | MET) ≈ 0.18 – the MET collection is far more than 10 times larger than the DAM collection. Hence, the P(MET | Not on paper) = P(MET & Not on paper)/P(Not on paper) is higher than P(DAM | Not on paper) = P(DAM & Not on paper)/P(Not on paper) because P(MET & Not on paper) will be bigger than P(DAM & Not on paper).]

(b) If independent, then P(Display | Ranked D) = P(Display) and that is somewhere between 0.08 to 0.10. Of course, we do not think that the decision to display art is independent of the quality of the piece: we would expect low quality pieces to have a much lower probability of display: P(Display | Ranked D) < 0.08 (much less!).

(7) Among women living in counties in the 10th percentile of poverty, predicted life expectancy increased by 2.84 years from 1990 to 2010:

(80.4051 – 0.0301*10 + 2.9203 – 0.0083*10) – (80.4051 – 0.0301*10) = 2.9203 – 0.0083*10 = 2.84 years

Among women living in counties in the 90th percentile of poverty, predicted life expectancy increased by 2.17 years from 1990 to 2010:

(80.4051 – 0.0301*90 + 2.9203 – 0.0083*90) – (80.4051 – 0.0301*90) = 2.9203 – 0.0083*90 = 2.17 years

This means that mortality *inequality* in terms of life expectancy at birth for women actually got worse from 1990 to 2010: the gap between the rich (10th poverty percentile) and the poor (90th poverty percentile) widened because the richer counties made faster progress in increasing life expectancy.

(8) (a) Using the same approach as Pritchett and Summers (2014), as the *Supplement* explains, take these steps [note that you may combine steps 1) and 2)]:

1) Construct a new variable (column) for GDP per capita = rgdpna/pop
2) Construct a new variable (column) for the natural log of GDP per capita = ln(rgdpna/pop)
3) To obtain -0.0150951, using the 15 years of data from 1986-2000 for Romania, run a simple regression where the y-variable is ln(rgdpna/pop) and the x-variable is year: -0.0150951 is the slope coefficient for the year variable and measures the growth rate (real GPD per capita declining 1.5% annually)
4) To obtain 0.0440858, using the 15 years of data from 2000-2014 for Romania, run a simple regression where the y-variable is ln(rgdpna/pop) and the x-variable is year: 0.0440858 is the slope coefficient for the year variable and measures the growth rate (real GPD per capita rising by 4.4% annually)

(b) The very low value of the R-squared means that only 0.2 percent (less than 1 percent!) of the variation across countries in annual GDP/capita growth rates during the period from 2000-2014 can be explained by variation across countries in the those growth rates during the period from 1986-2000. There is no relationship of any kind between how fast countries grew (in terms of GDP per capita) in the period of 1986-2000 with how fast they grew in the period of 2000-2014. [This is a main conclusion of the "Asiaphoria Meets Regression to the Mean" paper: in general, we should *not* expect that countries that are currently growing fast will continue to do so in the future: past growth is a *terrible* predictor of future growth.]

(9) Use an F-test: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(0.258)^2/1}{(1-(.258)^2)/(76-1-1)} = \frac{0.0666}{(1-0.0666)/(76-1-1)} = 5.28$. The F table reveals that the correlation between negative reciprocity and patience is statistically significant at the 5% level.

[Note: In this special case where $k = 1$, the F test statistic equals the t test statistic squared. Hence, it is possible to do a TWO-tailed t test being careful to compare the t test statistic with the $t_{\alpha/2}$ critical value.]

(10) (a) In California in the years 2003 and 2009, on average households that have income that is 10% higher use approximately 1.9% more electricity. [Note that Column (1) shows a simple regression: *nothing* is being controlled for (i.e. *nothing* is being held constant/fixed).]

(b) Compared to houses in Climate zone 1, California homes in Climate zone 13 in the years of 2003 and 2009 on average used approximately 48.5% less natural gas. Regression (3) does not control for anything: it just compares houses across climate zones. In contrast, Regression (4) controls other measures of climate, building characteristics, occupant characteristics, appliances, survey year, and the year the house is constructed, which are related with both the house's location (climate zone) and natural gas use.

(c) This requires comparing the coefficient estimates for the year-constructed dummies in Column (2) with those in Column (4). We see much bigger percent reductions in *natural gas* use comparing more recently built homes with those built before 1940, after we control for climate, building characteristics, occupant characteristics, appliances, and survey year. For example, the newest homes use 35% less natural gas than the oldest homes but only 9% less electricity.