ECO220Y1Y, APRIL 2018, FINAL EXAM: SOLUTIONS

(1) (a) We need to make an inference about the difference in population proportions. To obtain the sample proportion for no taglines: $\hat{P}_{notag} = \frac{0.43*407+0.65*397}{804} = 0.5386$. To obtain the sample proportion for superfluous taglines: $\hat{P}_{tag} = \frac{0.37*394+0.51*405}{799} = 0.4410$. Hence, the point estimate of the difference is -0.0976, which says that people shown misleading advertising (superfluous taglines) are nearly 10 percentage points less likely to choose the best credit card compared to people who make the choice without being confronted with misleading advertising (no taglines).

To test for a difference (either way) requires a two-tailed test:

$H_0: (p_{notag} - p_{tag}) = 0$

$H_1: (p_{notag} - p_{tag}) \neq 0$

$z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}}$ where $\bar{P} = \frac{X_1+X_2}{n_1+n_2}$

$$\bar{P} = \frac{X_1+X_2}{n_1+n_2} = \frac{(407*0.43+397*0.65)+(394*0.37+405*0.51)}{804+799} = \frac{433.06+352.33}{804+799} = \frac{785.39}{1603} = 0.4900$$

$$z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}} = \frac{0.4410-0.5386}{\sqrt{\frac{0.49(1-0.49)}{804} + \frac{0.49(1-0.49)}{799}}} = \frac{-0.0976}{\sqrt{\frac{0.2499}{804} + \frac{0.2499}{799}}} = \frac{-0.0976}{0.02497} = -3.91$$

The difference is highly statistically significant at any conventional significance level (noting that the Standard Normal table stops at z values of 3.69 as the tail areas become so tiny), including an α of 0.001.

(b) We need to make an inference about the difference in population proportions. The point estimate of the difference is 0.14, which says that among people who saw misleading advertising (superfluous taglines) those that saw the implemental video were 14 percentage points more likely to choose the best credit card compared to people who saw the baseline video.

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

$$(0.51 - 0.37) \pm 2.576\sqrt{\frac{0.51(1-0.51)}{405} + \frac{0.37(1-0.37)}{394}}$$

$$(0.14) \pm 2.576 * 0.03477$$

0.14 ± 0.0896 which gives a LCL of 0.05 and an UCL of 0.23

For people who have to make a credit card choice while faced with misleading ads, we are 99% confident that showing them the longer (implemental) video *increases* the percent selecting the best credit card by between 5 and 23 percentage points compared to the shorter (baseline) video. (A causal interpretation is correct because these are experimental data where the key x variable – which video a person watched – is randomly assigned.) While it is clear that the longer video helps people not be distracted by misleading advertising, the width of the interval is wide: it may increase the percent making the best choice by only 5 p.p. but it could have a huge impact of 23 p.p.

(2) This requires an inference about the difference between means for *paired* data: $H_0: \mu_d = 0$ versus $H_1: \mu_d \neq 0$ where the correct test statistic is given by $t = \frac{\bar{d}}{s_d/\sqrt{n}}$.

(3) (a) In Panel A, the shape of the distribution is Uniform.

In Panel B, the shape of the distribution from $0 to $50 is positively (right) skewed.

In Panel B, the shape of the distribution from $900.01 to $1,000 is bimodal.

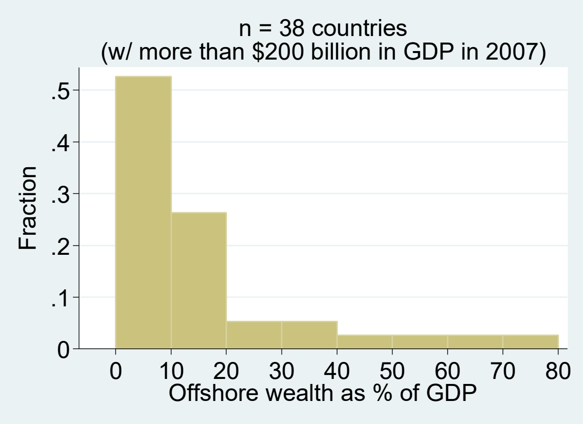
In Panel B, the shape of the distribution from $950.01 to $1,000 is negatively (left) skewed.

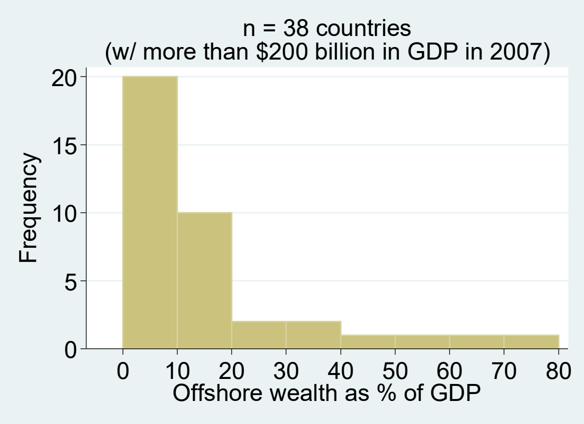
(b) Observations with a listing price less than $300.01 represent 30 percent of the subsample data. Work for 30: Since it is Uniform, we find the first percent as 100*6/20 = 30, or equivalently, 0.001*50*6 = 0.30, which is 30%.

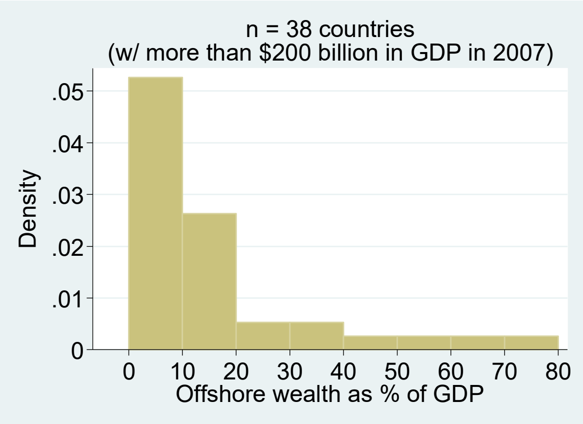
Observations with a listing price from $990.01 to $1,000 represent 3.5 percent of the subsample data. Work for 3.5: The height of the bar is around 0.0035 and the width is 10, which means 0.0035*10=0.035, which is about 3.5%.

(4) (a) The exact value of the IQR is 12.8 (=17.4 – 4.6), which is the only reasonable choice among those given. As the distance between the 75th and 25th percentiles, it measure the spread (i.e. variability) of the middle 50% of the data. There is substantial variability among countries in the % of wealth held offshore even once we exclude the bottom and top quarters of data (where all the extremes are): there is a 12.8 percentage point difference between the 75th percentile country and the 25th percentile country.

(b) Can choose to draw a relative frequency, frequency or density histogram (all three shown below), but it must be clearly labelled. Also, it is reasonable to put Ireland, which is very near the boundary of bin 1 and bin 2 into either bin. No matter how you draw the histogram, it is positively skewed.







(5) (a) We need to obtain b_1 and b_0 in $\hat{y}_i = b_0 + b_1 x_i$ where y is the firm's adaptive practice and x is the natural log of the firm's age. Plugging in: $b_1 = r\frac{s_y}{s_x} = 0.15 * \frac{1.39}{0.40} = 0.52125$ and $b_0 = \bar{Y} - b_1\bar{X} = 4.18 - 0.52125 * 1.26 = 3.523225$. Hence, the OLS equation is $\widehat{adaptive_\imath} = 3.52 + 0.52 * \ln(age)_i$. [Firms that are 10 percent older on average have adaptive practices that are 0.052 units higher on a seven point Likert scale.]

(b) $R^2 = (r)^2 = (0.14)^2 = 0.0196$

$$s_y^2 = \frac{SST}{n-1} \quad 1.39^2 = \frac{SST}{207-1} \quad SST = 398.0126$$

$$R^2 = \frac{SSR}{SST} \quad 0.0196 = \frac{SSR}{398.0126} \quad SSR = 7.80104696$$

$$SST = SSR + SSE \quad SSE = 398.0126 - 7.80104696 = 390.211553$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{390.211553}{207-2}} = 1.38$$

(c) We can use an F-test to test if there is a statistically significant correlation between the natural log of firm age and rainfall change. $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(-0.08)^2/1}{(1-(-0.08)^2)/(207-1-1)} = 1.32$. The relevant critical value from the F table is 2.71 (or 2.75 if you want to be very conservative) and hence this correlation is not close to being statistically different from zero.

(6) (a) $P(10th\ decile\ in\ 2012 \mid 1st\ decile\ in\ 2007)$, which is a conditional probability. There is a 1.5 percent chance that a Canadian taxfiler who was in poorest income decile (1st decile) in 2007 will move up to the richest income decile (10th decile) in 2012.

(b) Use the Binomial probability formula to find: $P(X \geq 7) = P(X = 7) + P(X = 8)$.

$$p(7) = \frac{8!}{7!(8-7)!}0.574^7(1-0.574)^{8-7} = 8 * 0.574^7(0.426)^1 = 0.06997$$

$$p(8) = \frac{8!}{8!(8-8)!}0.574^8(1-0.574)^{8-8} = 0.574^8 = 0.01178$$

$$P(X \geq 7) = 0.06997 + 0.01178 = 0.08175$$

(c) Two mathematically equivalent approaches are to find $P(X > 400)$ or $P(\hat{P} > 0.40)$. Either way, we use the Normal approximation (we expect more than 10 successes and 10 failures).

Approach #1: $E[X] = np = 1{,}000 * 0.416 = 416$ and $V[X] = np(1-p) = 242.944$

$$P(X > 400) = P\left(Z > \frac{400-416}{\sqrt{242.944}}\right) = P(Z > -1.0265) \approx P(Z > -1.03) = 0.5 + 0.3485 = 0.8485 \cong 0.85$$

OR

Approach #2: $E[\hat{P}] = p = 0.416$ and $V[\hat{P}] = \frac{p(1-p)}{n} = 0.000242944$

$$P(\hat{P} > 0.40) = P\left(Z > \frac{0.40-0.416}{\sqrt{0.000242944}}\right) = P(Z > -1.0265) \approx P(Z > -1.03) = 0.5 + 0.3485 = 0.8485 \cong 0.85$$

(d) All the cells in the transition matrix – the 10 rows and first 10 columns of results – would be 10. Hence, people in the 2nd decile in 2007 would have a 10% chance of immobility (remaining in the 2nd decile in 2012), a 10% chance at downward mobility, and an 80% chance of upward mobility. Canada is very different: people who are quite poor (2nd decile) in 2007 have a much greater chance of staying exactly that poor (39.4 compared to 10%) and a somewhat higher chance of becoming even poorer (13.5 compared to 10%). In Canada, the chance that someone in the 2nd decile is upwardly mobile is only 47%, which is much lower than 80%. Hence, income mobility is much less in Canada than in the hypothetical country (which has an extreme form of income mobility).

(7) (a)

$H_0: \beta_{new} = 0$

$H_1: \beta_{new} \neq 0$

$t = \frac{b_j - \beta_{j0}}{s_{b_j}} = \frac{0.020}{0.007} = 2.857$ with degrees of freedom that are far above 1,000 so we can use the Normal table as an excellent approximation when finding the P-value:

$P - value = 2 * P(t > 2.86) = 2 * (0.05 - 0.4979) = 0.0042$, which means the coefficient is highly statistically significant. (Alternatively, someone using the Student t table could say that the P-value lies between 0.002 and 0.01.)

(b) After controlling for the many housing characteristics (such as age, overall size, location) listed as explanatory variables in Table 2, houses with an additional bedroom on average have selling prices that are 5.4 percent *lower.* [Note: This does not mean that extra bedrooms are bad – there is surely a positive correlation between selling price and number of bedrooms – but rather having more bedrooms when we hold the overall size of the house, number of bathrooms, and other key variables fixed is not good (as it means tiny rooms).]

After controlling for the many housing characteristics (such as age, overall size, location) listed as explanatory variables in Table 2, houses that are 1 percent larger on average have selling prices that are 0.8 percent higher.

(c) Looking at Table 1 we can see that Energy Star homes on average sell for 22 percent higher prices than homes with no certification $\left(= \frac{326{,}940 - 267{,}685}{267{,}685} = 0.22\right)$. Hence, the simple regression coefficient would be approximately 0.22, compared to the much lower value of 0.027 in the multiple regression. Why? Table 1 also shows that Energy Star homes are on average better than homes without certification: they are much more likely to be new, are larger, and much more likely to have a garage. The multiple regression controls for these other important factors, which are correlated with the Energy Star rating (and hence are lurking/confounding/omitted/unobserved variables in the simple regression). Those variables bias the simple regression coefficient estimate (endogeneity bias) and multiple regression helps isolate the effect of the Energy Star rating on housing prices, which is much more modest: 2.7 percent higher prices, not 22 percent higher.

(d) No. Even though that coefficient is definitely not statistically significant (the t ratio would be 0.67, which falls far short of meeting even an easy 10 percent significance level and we cannot reject the null hypothesis that the coefficient is zero), we must remember that it does *not* tell us how prices differ between these two groups of houses (remember part (c)). Instead, it tells of differences after controlling for other more important house characteristics (like new, size, garage). We should expect that Austin also has higher average selling prices for Energy Star homes (that are highly statistically significant) like that observed in North Carolina. The key issue is the difference between controlling for other housing characteristics or not (not potential differences across cities).