ECO220Y, Term Test #1: SOLUTIONS

November 4, 2016, 9:10 – 11:00 am

(1) (a) Two points determine a line: use $(1990, 45)$ and $(2005, 55)$, which give (x_1, y_1) and (x_2, y_2). Use the point-slope formula: $(y - y_1) = m(x - x_1)$ where $m = \frac{\Delta y}{\Delta x} = \frac{(55-45)}{(2005-1990)} = \frac{2}{3}$ to obtain $(y - 45) = \frac{2}{3}(x - 1990)$. Hence, the OLS line is approximately $\hat{y} = -1282 + \frac{2}{3}x$, which we could write as $\widehat{days} = -1282 + \frac{2}{3}year$.

(b) The intercept has no interpretation here: we *cannot* say that in year zero time from harvest to veraison was negative 1282 days! The slope says: in the period from 1988 to 2014 in the Bordeaux region of France the days from harvest to veraison increased by two-thirds of a day each year on average for the Cabernet Franc grape.

(c) $\widehat{weeks} = -183 + 0.095year$. The (approximate) slope, 0.095, would be interpreted as, in the period from 1988 to 2014 in the Bordeaux region of France the days to veraison increased by about 0.1 weeks each year on average for the Cabernet Franc grape, which corresponds to about 1 additional week per decade.

(d) $\widehat{days} = 43 + \frac{2}{3}year$. The (approximate) intercept, 43, would be interpreted as, the OLS line predicts that in 1988 the days to veraison would be 43. A value of zero for year is now within the range of the data and plausible.

(2) (a) The data are panel (longitudinal) data. The unit of observation is a particular water station on the river in a particular year. There 515 observations in these data, corresponding to pollution measurements at 103 water-stations for each of 5 years (with some missing values for some variables for some water stations in some years). There are two identifier variables – water station name and year – and there are seven variables measuring pollution – COD, BOD, NH, petroleum, phenol, mercury and lead.

(b) For each of the five years, the standard deviation of mercury concentrations, measured a micrograms per litre $\mu g/L$, are bigger than the mean concentrations and for most years, way bigger. Because concentrations of mercury cannot logically be less than zero, the only way to obtain a big s.d. is with a tail to the right. Hence mercury concentrations must be very positively skewed (aka right skewed), with a few locations have very high concentrations.

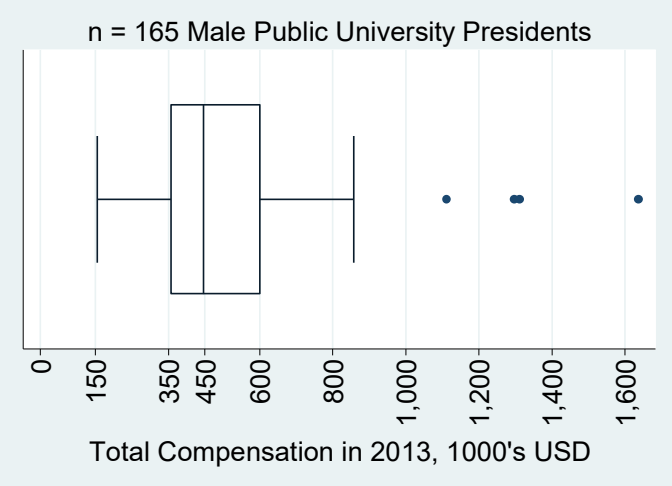
(c) The coefficient of correlation and the covariance are related: $r = \frac{s_{xy}}{s_x s_y}$. Using the sample standard deviations in the table, combine the values of r, s_x, and s_y to find s_{xy} for each pairwise combination. To find the diagonal elements of the variance-covariance matrix, square the reported standard deviations to obtain the variances.

The units of measurement of all numbers in the matrix are $(mg/L)^2$: in other words, milligrams per litre squared.

```
correlate COD BOD NH, covariance; /* Variance-Covariance Matrix */
(obs=98)

             |       COD       BOD        NH
-------------+---------------------------
         COD |      1.59
         BOD |      0.91      1.39
          NH |      0.73      0.81      2.20
```

(3)



(4) (a) $P(40\ to\ 45 \mid SUV) = \frac{0.0409}{0.2796} = 0.146$ versus $P(40\ to\ 45 \mid van) = \frac{0.0111}{0.0685} = 0.162$. Hence, there is a greater chance of getting a male 40 to 45 years old if you select a male who purchased a van.

(b) Events A and B are not independent. Independence requires that $P(A \mid B) = P(A)$ (or equivalently, that $P(B \mid A) = P(B)$). However, in this case $P(A \mid B) = \frac{0.0041}{0.0421} = 0.097$ and $P(A) = 0.074$, which are definitely not equal. (Alternatively, and equally correct, students could find $P(B \mid A) = \frac{0.0041}{0.0738} = 0.056$ and $P(B) = 0.042$, which are also definitely not equal.) (Another alternative that students may use is showing that P(A & B) is not equal to the P(A)*P(B).) Unsurprisingly, males in different age groups tend to have preferences for different kinds of vehicles. Events A and B are not disjoint (mutually exclusive): a male purchaser can be both 25 to 30 years old AND purchase a sporty vehicle: $P(A\ \&\ B) = 0.0041 \neq 0$.

(5) $n = 4$

$p = 0.5199 + 0.2105 = 0.7304$

Binomial probability: $p(x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$ for $x = 0,1,2, \ldots, n$

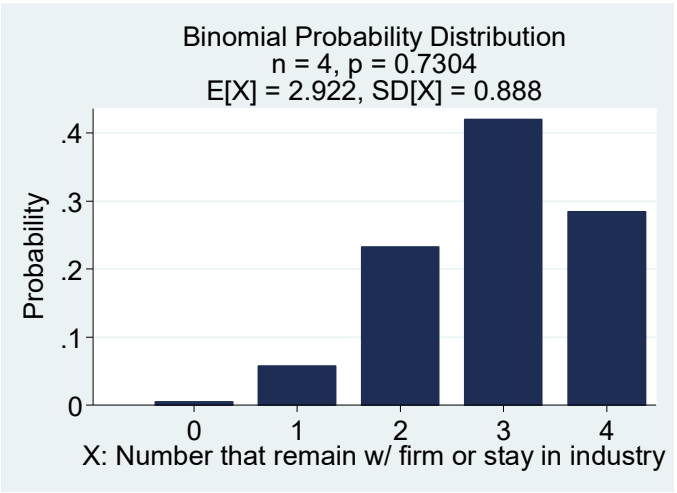
$$p(0) = \frac{4!}{0!\,(4-0)!} 0.7304^0 (1 - 0.7304)^{4-0} = 0.005$$

$$p(1) = \frac{4!}{1!\,(4-1)!} 0.7304^1 (1 - 0.7304)^{4-1} = 0.057$$

$$p(2) = \frac{4!}{2!\,(4-2)!} 0.7304^2 (1 - 0.7304)^{4-2} = 0.233$$

$$p(3) = \frac{4!}{3!\,(4-3)!} 0.7304^3 (1 - 0.7304)^{4-3} = 0.420$$

$$p(4) = \frac{4!}{4!\,(4-4)!} 0.7304^4 (1 - 0.7304)^{4-4} = 0.285$$



$mean = E[X] = np = 4 * 0.7304 = 2.922$

$variance = V[X] = np(1-p) = 4 * 0.7304 * (1 - 0.7304) = 0.788$; $standard\ deviation = \sqrt{0.788} = 0.888$

(6) (a) See Section 12.2.2 “Getting from the PWT 8.0 data to Tables 1 and 2” in the 2016/17 version of “Logarithms in Regression Analysis with Asiaphoria.”

(b) Among the 112 non-OECD member nations, on average countries with growth rates in the 1970’s that were 1 standard deviation higher had growth rates in the 1980’s that were 0.33 standard deviations higher. In other words, these is substantial regression to the mean.

(c) Among the 112 non-OECD member nations, only 5 percent of the variation in growth rates in the 2000’s across these countries is explained by variation in the growth rates in the 1990’s. In other words, these is a huge amount of scatter and growth rates in the 1990’s have virtually no predictive power for growth rates in the 2000’s.

(d) The huge discrepancy between the correlation and rank correlation suggests there is an issue with an outlier: remember that the correlation is sensitive to outliers but the rank correlation is robust to outliers. The scatter diagram confirms that there is a single outlier, South Korea, which had very strong growth in both of those consecutive decades. The presence of South Korea leads to a much stronger correlation, steeper regression coefficient, and higher R-squared than if we excluded it: the remaining 29 countries indicate no relationship at all. However, while the presence of South Korea makes the correlation look stronger, it is still quite weak and even with this outlier distorting the results, they still support the authors’ main claim. There is a huge amount of scatter and the growth rates in OECD countries in a previous decade do a poor job in forecasting the next decade’s growth (i.e. regression to the mean).