ECO220Y1Y, Test #5, Prof. Murdock: SOLUTIONS

April 5, 2019, 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) To start, answering requires repeating the analysis from Test #4, March 2019, Question (2)(a), for Canada and Japan to obtain the sample s.d. for each country.

Canada: ME is 0.0922 $\left(=\frac{7.4207-7.2363}{2}\right)$. Use $\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ with $\nu = 2{,}026$ so $t_{\alpha/2} = 1.960$ for $\alpha = 0.05$. Hence, $ME = 0.0922 = 1.960 * \frac{s}{\sqrt{2{,}027}}$. Yields $s = 2.12$. $\bar{X} = 7.3285$ $\left(=\frac{7.4207+7.2363}{2}\right)$. (Or you could get 7.328 from figure.)

Japan: ME is 0.0817 $\left(=\frac{5.9967-5.8333}{2}\right)$. Use $\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ with $\nu = 3{,}008$ so $t_{\alpha/2} = 1.960$ for $\alpha = 0.05$. Hence, $ME = 0.0817 = 1.960 * \frac{s}{\sqrt{3{,}008}}$. Yields $s = 2.29$. $\bar{X} = 5.9150$ $\left(=\frac{5.9967+5.8333}{2}\right)$. (Or you could get 5.915 from figure.)

If we call Canada group 1 and Japan group 2, we plug into: $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$. The degrees of freedom will be very large (well above 1,000) so it is not necessary to compute the exact degrees of freedom.

$(7.3285 - 5.9150) \pm 1.960\sqrt{\frac{2.12^2}{2{,}027}+\frac{2.29^2}{3{,}008}}$

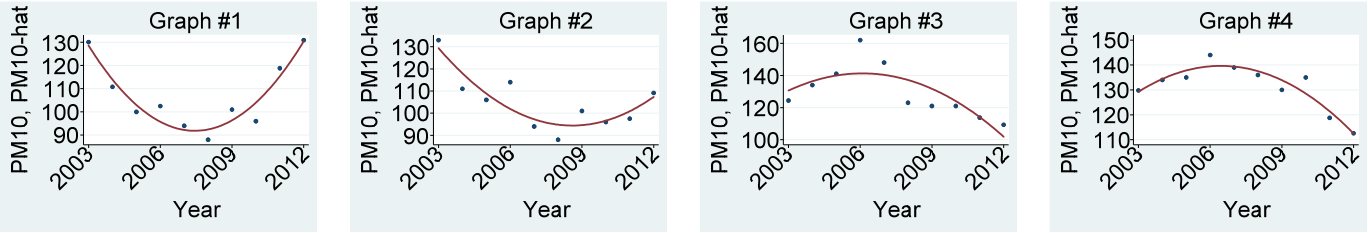
$1.4135 \pm 1.960 * 0.0629$

1.4135 ± 0.1233

$LCL = 1.29$ and $UCL = 1.54$

(b) $\widehat{happiness} = 6.414 - 0.078 * abroad$ with $n = 23{,}909$, where $-0.078 = (6.336 - 6.414)$ and $23{,}909 = 19{,}933 + 3{,}976$.

(2) (a)



Graph #2 summarizes the regression results for Tianjin. We can immediately rule out Graphs #3 and #4, which would require a *positive* coefficient on the trend and a *negative* coefficient on the trend-squared, which is the opposite of the actual results. To choose between Graphs #1 and #2, focus on last few years where there is a big difference. The given results predict that in 2012, PM10 is 107.35 (= 143.0097 – 14.73599*10 + 1.117*10^2), which means it must be Graph #2.

(b) We cannot interpret that coefficient in isolation because we cannot imagine a change in the trend variable without also changing the trend squared variable. Hence, we need to draw a graph (like part (a)) that takes both -14.73599 and 1.117 into account because they both affect the slope.

(3) In Regression #1, we simply compare the annual household electricity usage of older and newer homes in California only controlling for the year of the survey (2003 versus 2009). The coefficient in boldface tells us that the electricity usage of houses built during the period from 2001-2004 use ***39.5 percent more electricity*** than houses constructed before 1940. The extremely tiny P-value means that this difference is highly statistically significant. This looks terrible because building codes were supposed to make newer homes *more* energy efficient, but they seem to be actually using *more* electricity (*less* efficient) than older homes.

In Regression #2, we are controlling for some other potentially important reasons for higher electricity usage: the climate (how much it requires air conditioning), the size of the house, the number of people living at the house, and whether or not the house has central air conditioning. After controlling for these differences, we find that newer homes (i.e. those build between 2001-2004) actually use ***5.3 percent less electricity*** than houses constructed before 1940. Further, this difference is statistically significant at a 5% significance level.

These regressions illustrate a main point in Levinson (2016): it is not fair to simply compare older and newer homes when trying to figure out if building codes have improved energy efficiency because newer homes tend to be built in hotter parts of California, have more residents, and have central air conditioning. Hence, Regression #2 is clearly better for answering the primary research question (conveniently included in the title of the article) than the unfair comparison in Regression #1.

(4) (a) Use an F-test: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(-0.2892)^2/1}{(1-(-0.2892)^2)/(100-1-1)} = \frac{0.0836}{(1-0.0836)/(100-1-1)} = 8.94$. The F table reveals that the correlation between y and x1 is statistically significant at the 1% level.

(b) Use an F-test: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(0.1891)^2/1}{(1-(.1891)^2)/(100-1-1)} = \frac{0.0358}{(1-0.0358)/(100-1-1)} = 3.64$. The F table reveals that the correlation between y and x2 is statistically significant at the 10% level.

(c) False. A multiple regression coefficient and statistical tests of those coefficients assess how y is related to each x *after controlling for the other included x variables*. Multiple regression coefficients do NOT describe how each x is related to y in observational data. Hence, the statistical tests above (using correlations) are necessary.

(5) (a)

$H_0: \mu_T - \mu_C = 0$

$H_1: \mu_T - \mu_C \neq 0$

Depending on whether or not you assume equal variances, you use one of these two formulas. Either approach is totally acceptable in this case and leads to virtually identically results.

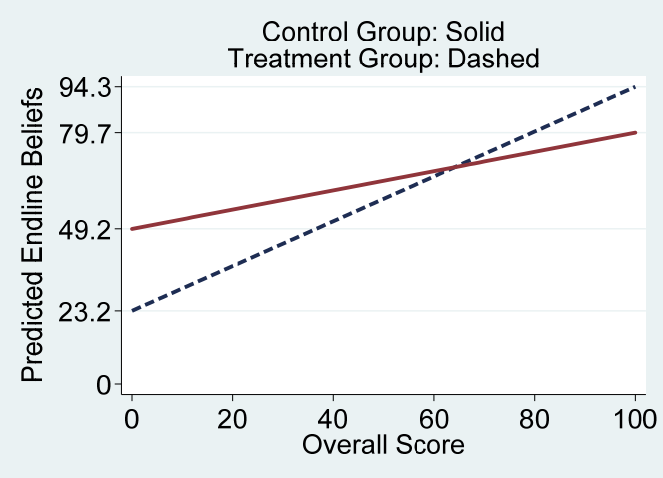
$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ or $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$

Going with the general case (unequal variances), $t = \frac{(\bar{X}_T - \bar{X}_C) - \Delta_0}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}} = \frac{(46.35731 - 47.13075) - 0}{\sqrt{\frac{307.3227}{2{,}614} + \frac{304.4582}{2{,}654}}} = -\frac{0.77344}{0.48196} = -1.6048$

The degrees of freedom will be very large, which means we can use the Standard Normal table. The P-value is approximately 0.11, which means that this difference is not statistically significant at even a 10% level.

This answer is not surprising because the students were *randomly* divided into the treatment group and the control group. There is no reason to expect any difference in the academic performance across these two groups.

(b)



Control group: $\widehat{Beliefs}_i = 49.2 + 0.305 * OverallScore_i$

Treatment group: $\widehat{Beliefs}_i = 23.2 + 0.710 * OverallScore_i$

(c) Model: $EndlineBelief_i = \beta_0 + \beta_1 Treat_i * OverallScore_i + \beta_2 OverallScore_i + \beta_3 Treat_i + \varepsilon_i$

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$t = \frac{0.405-0}{0.024} = 16.9$ with $\nu = 5{,}240$

Yes, the difference in the slopes is highly statistically significant, including at the 0.1% significance level.

(d) Regression (2) is the best regression. Because the outlier is obviously a data entry mistake in this case – it is not possible to believe that the score will be 350 points if the maximum possible score is 100 points – the best thing to do (given that we lack further information that would allow us to correct the data entry mistake) is to drop it completely, which is what Regression (2) does. The huge sample size means that a single outlier (even an extreme one like this) does not have a big effect. However, the outlier has distorted the R-squared in Regression (1). Regression (3) controls for the outlier, which means the coefficients are the same as in Regression (2), but that artificially inflates the R-squared.