**ECO220Y1Y, APRIL 2022, FINAL EXAM: SOLUTIONS**

**(1) (a)** Within one standard deviation of the mean is between 0 TRY and 27,751 TRY, remembering that a credit line limit cannot be below zero. The 90th percentile is 24,100 TRY, which means that more than 90% of the participants must have a credit line limit within one standard deviation of the mean. [The distribution is very positively skewed.]

**(b)** The standard deviation would be larger for all customers (excluding participants). This is because this distribution has a higher fraction of people at the minimum extreme of zero and the maximum extreme of 1, and this implies more variation and a higher standard deviation.

**(c)** First, note that the horizontal axis is on a log base 10 scale to address the extreme positive skew of the available credit (TRY) variable. With the log scale the distribution is bimodal: nearly 4 percent of the customers have no available credit (zero is the first major peak) whereas the median is around 10,000 TRY (which is the second major peak). Excluding the zeros, it is roughly Normal after the log transformation, although there is a slight negative skew.

**(d)**

$P(A) = \frac{18{,}666}{45{,}307} = 0.4120$

$P(B) = \frac{166+1{,}758+1{,}959+1{,}787+1{,}115}{45{,}307} = \frac{6{,}785}{45{,}307} = 0.1498$

$P(A \,\&\, B) = \frac{1{,}758}{45{,}307} = 0.0388$

$P(A \mid B) = \frac{1{,}758}{6{,}785} = 0.2591$

$P(B \mid A) = \frac{1{,}758}{18{,}666} = 0.0942$

**(2)** $\bar{X} \approx 0.245 * 0 + 0.083 * 1 + 0.139 * 2 + 0.130 * 3 + 0.071 * 4 + 0.333 * 5 = 2.7$ days per week

**(3) (a)** From Table 2, among the 46,419 people completing the survey, for the expected 1-year stock return the interquartile range is 5 percentage points.

**(b)** These are paired data so use $\bar{X}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$. We must compute $s_d$ using the Laws of Variance.

$s_d = \sqrt{V[1Y - 10Y]} = \sqrt{6.08^2 + 3.85^2 - 2 * 6.08 * 3.85 * 0.304} = \sqrt{37.55684} = 6.13$

$\bar{X}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$ with degrees of freedom $(\nu)$ of 46,418, which is essentially infinity.

$(4.64 - 6.64) \pm 2.576\frac{6.13}{\sqrt{46{,}419}}$

$-2 \pm 2.576 * 0.02845$

$-2 \pm 0.07$ and this point estimate and margin of error imply $LCL = -2.07$ and $UCL = -1.93$

**(4)** The difference between Regressions #1 and #2 is because there is a quadratic (parabolic) relationship between finish time and start order. Regression #1 forces a line when the data are curved, whereas Regression #2 correctly accommodates the curve. However, start order is an endogenous variable: it is not random which skiers ski when but instead the most skilled skiers get the opportunity to pick their preferred spot when ski conditions are expected to be the best. After controlling for skill, as measured by FIS points, in Regression #3, the start order variable (and start order squared) become less important in explaining finish time.

**(5) (a)** This is the very tiny P-value for the test of the overall statistical significance of the multiple regression.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{11} = 0$

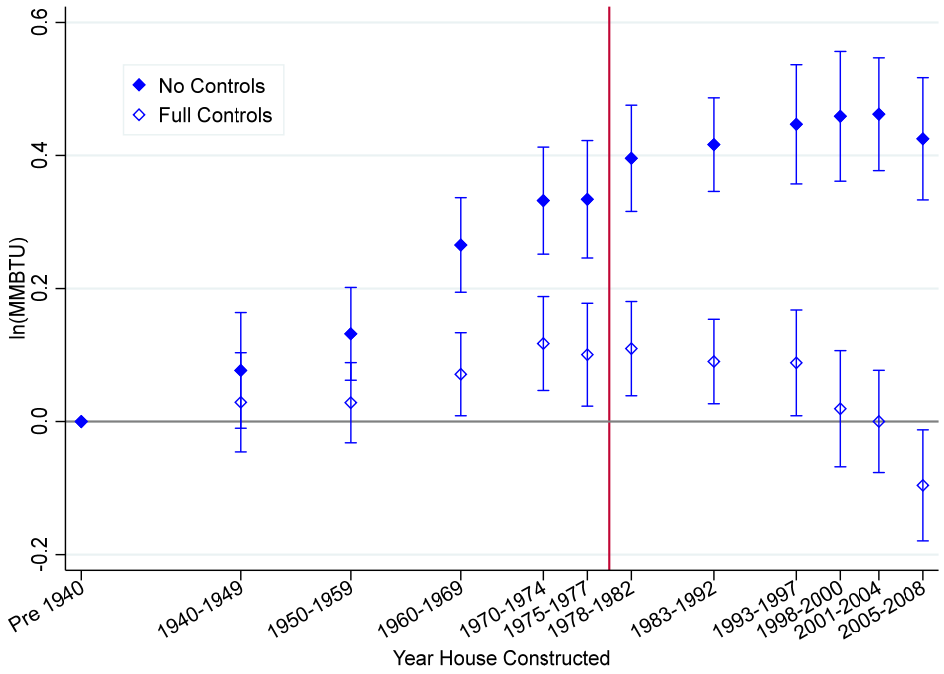
$H_1$: Not all the slope coefficients are zero

The multiple regression overall is highly statistically significant: we can very easily reject the null that states that the year the house was constructed has nothing to do with annual electricity use (logged). We have overwhelming evidence that year of construction does matter.

**(b)** The total sum of squares (SST) measures the total variation of the y variable about its mean. Because both regressions have the same y variable, which is the natural log of annual electricity use, the SST will be identical. The regression sum of squares is substantially bigger in the second regression because that regression includes important house, household, and climate variables that can help explain variation in electricity use across homes in addition to when the house was built, which are the only explanatory variables in the first regression. Hence, the R-squared is substantially larger, 0.33 versus 0.08, in the second regression.

**(c)** 2005-2008, No Controls: $b_{11} \pm t_{\alpha/2} s_{b_{11}} = 0.425060 \pm 1.960 * 0.046869 = 0.425 \pm 0.092$, yielding $LCL = 0.333$ and $UCL = 0.517$

2005-2008, Full Controls: $b_{11} \pm t_{\alpha/2} s_{b_{11}} = -0.095906 \pm 1.960 * 0.042574 = -0.096 \pm 0.083$, yielding $LCL = -0.179$ and $UCL = -0.012$



**(d)**

$H_0: \beta_{16} = 0$

$H_1: \beta_{16} \neq 0$

$t = \frac{b_{16} - \beta_{0_{16}}}{s.e.(b_{16})} = \frac{-0.095906238 - 0}{0.042573558} = -2.25$

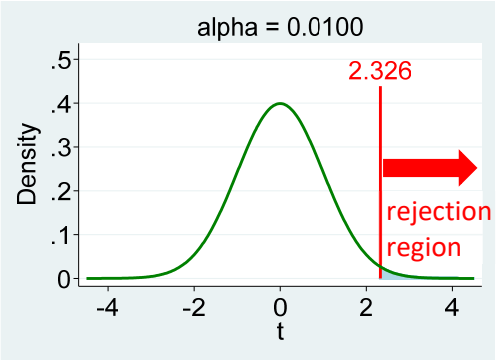
Given the very large degrees of freedom $(v = n - 16 - 1 = 4{,}438 - 16 - 1 = 4{,}421)$ we use the Normal table to compute the exact P-value:

$P - value = P(t < -2.25) + P(t > 2.25) = 2 * (0.5 - 0.4878) = 0.024$

Hence, this coefficient is statistically significant at a 5% significance level (but not a 1% significance level).

After controlling for climate, house size, number of residents, central AC, and an interaction term between central AC and house size, homes in ten counties in Southern California that were built from 2005 through 2008 on average use 9.6% less electricity annually in 2009 compared to homes constructed before 1940.

**(6) (a)** This is a difference in means for independent samples and we can either use the general case (unequal variances) or the special case of assuming equal variances. Below shows the general case, which is less work.




$H_0: (\mu_R - \mu_B) = 0$

$H_1: (\mu_R - \mu_B) > 0$

$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.91 - 3.63) - 0}{\sqrt{\frac{5.19^2}{17{,}931} + \frac{4.78^2}{23{,}474}}} = \frac{0.28}{\sqrt{0.002476}} = \frac{0.28}{0.05} = 5.6$ and given the very large sample sizes, we can treat the relevant degrees of freedom as infinity. The test statistic is clearly in the rejection region supporting the authors' claim of it being statistically significantly higher in red states at a 1% significance level.

**(b)** This is a difference in two proportions.

$H_0: (p_R - p_B) = 0$

$H_1: (p_R - p_B) \neq 0$

$z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}}$ where $\bar{P} = \frac{X_1 + X_2}{n_1 + n_2}$ and plugging in $\bar{P} = \frac{0.2801 * 18{,}103 + 0.2690 * 23{,}629}{18{,}103 + 23{,}629} = \frac{11{,}427}{41{,}732} = 0.2738$

$z = \frac{0.2801 - 0.2690}{\sqrt{\frac{0.2738(1-0.2738)}{18{,}103} + \frac{0.2738(1-0.2738)}{23{,}629}}} = \frac{0.0111}{0.0044} = 2.52$

$P - value = P(Z < -2.52) + P(Z > 2.52) = 2(0.5 - 0.4941) = 0.0118$. There is a statistically significant difference in the proportion under 40 in red versus blue states at a 5% significance level, but not at a 1% level.

**(7) (a)** Compared to swing states, on average inflation expectations are about 0.39 percentage points higher in red (Republican leaning) states during the period from 2013 to 2016 when there is a Democrat president (Barak Obama) after controlling for variation in age composition, numeracy skills, region of the US, education, income, and annual variation in inflation expectations.

**(b)** $\hat{y} = 4.793 + 0.281 + 0.653 - 0.437 - 0.432 - 0.781 = 4.077$

**(c)** In swing states in the west (region of the US), among those below 40 years of age, with low numeracy skills, with a high school degree or less, and with household income below $50,000, the average expected inflation is about 3.4 percent for 2017.