Chris Mulligan ⟨clm2186⟩
Intro Econometrics, Arkonac
Problem Set 3
Oct 3, 2012

Q1, Q2

Q1 This is HMC style problem 1.

This is the solution to HMC style problem 1. ■

Below this is a simple enumerate style more applicable to non-math classes.

1. **GPA4.dta**

   Regression Table:

Regressor	**(1)**	**(2)**	**(3)**
$hsGPA$	.456 (.094)	.455 (.093)	.461 (.090)
skipped	−.078 (.025)	−.065 (.026)	−.071 (.026)
PC	—	.129 (.06)	.137 (.059)
$bgfriend$	—	—	.086 (.054)
campus	—	—	−.124 (.079)
$Intercept$	1.580 (.325)	1.527 (.321)	1.49 (.317)
Regression summary statistics			
R^2	.2227	.2504	.2784
Regression RMSE	.331	.325	.322
n	141	141	141

   (a) Regress *colGPA* on $hsGPA$ and *skipped*.

$$colGPA = \underset{(.325)}{1.580} + \underset{(.094)}{.456} \cdot hsGPA - \underset{(.025)}{.076} \cdot skipped, \quad R^2 = .223,\ \text{SER} = .33$$

   (b) The coefficient 0.456 on $hsGPA$ means that a 1 point increase in high school GPA is associated with a 0.456 increase in college GPA, holding the number of times they skip class constant.

   (c) The t column in the Stata output has the appropriate test statistic value, in this case -3.05. As $|-3.05|$ is substantially more extreme than the 95% confidence level cutoff of 1.96 we reject the null hypothesis that the coefficient on *skipped* is zero.

   The hypothesis we're testing is whether there is a relationship between a student's attendance in class and their college GPA. If we reject the null hypothesis we conclude that there is a substantial linear relationship, i.e., skipping classes is associated with a lower grade point average.

(d) Our P-Values for regressions 1, 2, and 3 are 0.003, 0.013, and 0.007 respectively. If we're using a significance level of 99% then any P-value below .01 is signifigant, but anything above is not. As a result we conclude that in regressions 1 and 3 the coefficient on *skipped* is significantly different from zero, but on regression 2 we can't conclude that at $\alpha = 0.01$.

(e) The coefficient of $-.124$ makes sense in magnitude, but may not make sense in sign. The impact of living location seems unimportant, so a small magnitude of only about a tenth of a point makes sense. However the sign of the coefficient isn't entirely obvious. There seems to be little about living on campus that would directly cause students to perform worse. I wonder if *campus* is acting as proxy for another variable, such as indicating younger students more interested in socializing.

(f) The $bgfriend$ coefficient of .086 makes sense because the magnitude is small. I see plausible reasons for either direction on the coefficient. The coefficients on neither *campus* nor $bgfriend$ are statistically significant at the 1% level in (3), both their P-Values are over .1 so they aren't even significant at a 10% level.

2. $\overline{growth} = 1.943, s_{growth} = 1.897$
$\overline{tradeshare} = .565, s_{tradeshare} = .289$

3. (a) $\beta_{tradeshare} = 2.306$. The coefficient means that a 1% increase in $tradeshare$ is associated with an increase in growth rate of 2.306%.

(b) Regression of Growth on Tradeshare

(c) Yes, the coefficient on $tradeshare$ is significantly different from 0 at the 5% level. The P-Value Stata reports is 0.001, which is below 0.05.

(d) The 95% confidence interval for β_1 is $(.981, 3.632)$.

(e) $R^2 = .1237$, which means 12.37% of the total variation in $growth$ is explained by the change in $tradeshare$.

(f) $Corr(growth, tradeshare) = .3517$
$Corr(growth, tradeshare)^2 = .3517^2 = .1237 = R^2$
In a simple linear regression R^2 is the square of the correlation coefficient between the dependent and the independent variable.

(g) RMSE = 1.79. The Root Mean Squared Error is the standard deviation of the residuals after fitting our model. A larger RMSE indicates more variance in our error terms, and a worse model. The formula is:

$$\sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}}$$

(h) The regression error appears relatively homoskedastic. The variance of the residuals appears to decrease above approximately $tradeshare = .75$, but the number of datapoints decreases there too. As a result there may be some slight heteroskedasticity, but it doesn't appear to be severe.

(i) $\hat{\beta}_0$ and $\hat{\beta}_1$ are identical, however $SE(\hat{\beta}_i)$ are both higher without the robust option: $SE(\hat{\beta}_0)$ increased from .459 to .489; $SE(\hat{\beta}_1)$ increased from .663 to .773. As a result the P-Values are larger without the robust, and the confidence intervals are correspondingly wider.

(j) See Stata .do file and log.

Removing the outlier, Malta, makes a substantial difference to resulting regression. In a general sense the regression line has become much more flat, with a higher intercept and a smaller coefficient. However the P-Value on $\hat{\beta}_1$ is much higher, above the 5% significance level, and the R^2 is much lower. Clearly the outlier had a large affect on our OLS estimations of β_i.

(k) The outlier is Malta. It may be safe to exclude Malta due to its small size and unusual political and economic history during the period from 1960-1995. The island has only approximately half a million people. During WWII Malta was an important military location. Following the war Malta achieved independence from Great Britain in 1964, during the period under question. They later applied for EU membership. The substantial political and economic changes in Malta from 1960-1995 in addition to their small size means that it may be reasonable to exclude them as an outlier.

4. (a) The coefficient on $lorgdp60$ is -1.454.

This means that if a country had a GDP among the lowest 25% in 1960 their growth over the next 35 years was 1.454% lower than a country with a GDP in the upper 75% in 1960. In other words, poor countries grew more slowly than rich countries. This is a large affect on growth since the mean growth rate for this dataset is only 1.9%.

(b) Using the P-Value from Stata's regression output we can test test $H_o : \beta_{lorgdp60} = 0$. The P-Value is .024, below .05, so we conclude that $\beta_{lorgdp60}$ is significantly different from 0 at the 5% significance level, i.e., the growth rate was different for countries in the lowest quartile compared to the upper 3 quartiles.

(c) Difference of means for $lorgdp60$:

lorgdp60	**n**	**mean**	**s**
$X_0 = lorgdp60 = 0$	48	2.323	1.505
$X_1 = lorgdp60 = 1$	17	.869	2.467

$$\begin{aligned} &(lorgdp60 = 0) - (lorgdp60 = 1) \\ &= 2.323 - .869 \\ &= 1.454 \end{aligned}$$

t-test for difference of means:

$$\begin{aligned} t &= \frac{\overline{X}_0 - \overline{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}} \\ &= \frac{2.323 - .869}{\sqrt{\frac{1.505^2}{48} + \frac{2.467^2}{17}}} \\ &= \frac{1.454}{\sqrt{\frac{2.265}{48} + \frac{6.086}{17}}} \\ &= \frac{1.454}{\sqrt{.047 + .358}} \\ &= \frac{1.454}{.636} \\ &= 2.286 \end{aligned}$$

A t-statistic of 2.286 is above the 95% significance level of 1.96, so the difference is significant.

(d) The difference-of-means t-statistic of 2.286 is slightly smaller than the robust regression, and substantially smaller than the non-robust regression. This appears to be because the non-robust regression uses a smaller definition for $SE(\hat{\beta}_1)$ than the manual definition or the robust regression.

5. (a) $\widehat{\beta}_{age} = .4519$. Based on the Stata reported P-value of .000 the coefficient is statistically significant at any reasonable α.

   (b) $H_o : \beta_{age} = 1, H_a : \beta_{age} \neq 1$

$$\begin{aligned} t &= \frac{\widehat{\beta}_{age} - 1}{SE(\widehat{\beta}_{age})} \\ &= \frac{.451913 - 1}{.0329673} \\ &= \frac{-.548087}{.0329673} \\ &= -16.63 \end{aligned}$$

   The t-statistic of -16.63 is well below -1.96, meaning we can reject H_0 at the 95% significance level.

   (c)

$$\widehat{ahe} = \begin{cases} 13.81 & \text{if } bachelor = 0 \\ 13.81 + 6.50 = 20.31 & \text{if } bachelor = 1 \end{cases}$$

   (d) Regress *ahe* on *age* and *female*:

$$ahe = \underset{(.962)}{4.607} + \underset{(.033)}{.442} \cdot age - \underset{(.189)}{2.347} \cdot female, \quad R^2 = .0397, \text{ SER} = 8.5843$$

Tester.java

```
/**
 * COMSW1007, Kender, Fall 2012, Assignment 1, Programming Step 1
 * <p>
 * Object oriented Java Hello World application.
 * Execute this class. It news up an instance of the Greeter class and asks it
 * to say hello the constant GREETSUBJECT REPETITIONS times.
 *
 * @author Chris Mulligan clm2186@columbia.edu
 * @version 0.1
 */

public class Tester {

    private static final String GREETSUBJECT = ”World”;
    private static final int REPETITIONS = 4;

    /**
     * Main method that should be executed when you wish to say hello to the
     * world. Creates a Greeter instance with the GREETSUBJECT constant,
     * then calls greetN to do the greeting REPETITIONS times.
     *
     * @param args Unused command line arguments.
     */
    public static void main(String[] args) {
        Greeter myGreeter = new Greeter(GREETSUBJECT);
        myGreeter.greetN(REPETITIONS);
    }

}
```