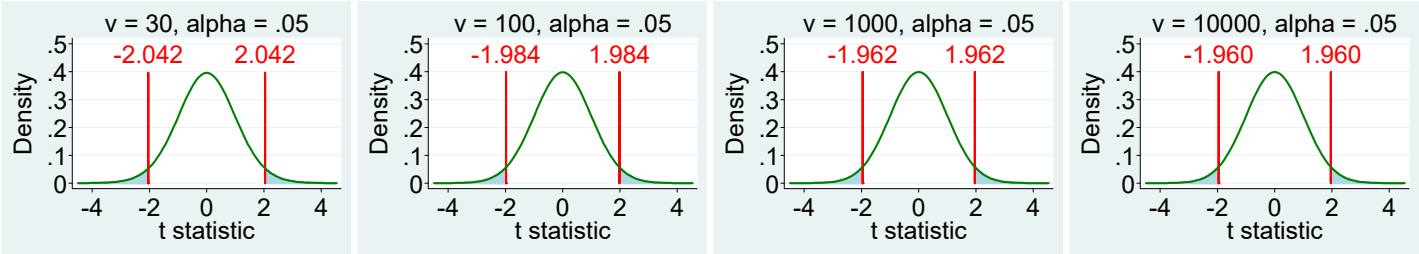
**Homework 20: ECO220Y – SOLUTIONS**

**Required Problems:**

**(1)** It is based on the conventional significance level of 5%. It is a great approximation for a range of degrees of freedom (only being inappropriate for very small degrees of freedom).




**(2) (a)** There are some outliers and heteroscedasticity. It would be useful to also report the results without the outliers. The standard deviation of the residuals is affected by both the outliers and the heteroscedasticity.

**(b)** Yes, they match up. The STATA summary reports that the $s_e$ is \$48,616. Looking at the scatter plot, the s.d. of the residuals around zero looks consistent with that (applying the Empirical Rule vertically).

**(3)** In addition to x-variables measuring the specific characteristics of the house for sale (such as square footage, property size, parking, etc.), a researcher could include x-variables of policy interest such as a variable measuring local pollution levels, proximity to a nuclear power plant, or the quality of the neighborhood school. The goal is to assess the causal link between these policy variables and housing prices. Of course, the hope is that you have controlled for any variables that are correlated with both the price of the house (the y-variable) and the policy variable (one of the x-variables). If so, the policy variable is *exogenous*. In other words, it will not be correlated with $\varepsilon$, which represents everything else that affect the price of the house aside from the included x-variables. If the policy variable is exogenous then we can interpret the coefficient on it causally and answer interesting research questions.

**(4) (a)** Two outliers with very high predicted percent body fat and very large negative residuals (i.e. the model greatly over-predicted their percent body fat) are apparent. You would look at the original x values for these two observations to investigate. One is a male that is apparently less than 3 feet tall: likely a data entry mistake. Another is tall very heavy male with the largest abdominal circumference (biggest belly) in the entire sample: his percent body fat may have been mis-measured.

**(b)** Note the differences across the results: it is important to check for outliers and address them (which does NOT always mean dropping them, even though that makes sense in this particular example).

**(5)** Notice everything is identical with the exception of the slope coefficients and standard errors associated with height and weight and the 95% CI's associated with these two coefficients. These have simply adjusted in a perfectly predictable way with the change in units. Noticed that even the t test statistic and P-value for these two coefficients are identical because the t test statistic is unit-free (it is the ratio of the coefficient and standard error, so units cancel).

**(6)** Around 0.09 millions of dollars (or \$90,000) for both graphs. You can see that this is correct looking back at the output given in Exercise 14 of Chapter 20 (which shows it is 0.0931). Of course, 0.10 or anything close would have also been a reasonable estimate. Recall that the estimate is obtained by remembering the mean of the residuals is always 0 (provided that a constant term/intercept has been included in the model) and by applying the Empirical Rule to either the first graph (vertically) or the second graph (horizontally) given in Exercise 22 of Chapter 20.

**(7)** LCL = 0.2582811 and UCL = 0.7607187. We are 95% confident that increasing the dosage of the sleeping drug by 1 mg will, on average, cause people to sleep somewhere between an extra 15.5 to 45.6 minutes.

**(8) (a)** Given that ROW is completely exogenous (it was set completely randomly) we can easily interpret it's slope. On average moving students one row further away from the front of the classroom decreases their score by about a half a percentage point, controlling for their score in ECO100. Alternatively, on average moving students ten rows further away from the front of the classroom decreases their score by about five percentage points holding ECO100 scores fixed. Notice I used the word "decreases," which implies causality. That is OK because ROW is exogenous. More caution is warranted in interpreting the slope on MARK_100. On average, a one percentage point increase in students' ECO100Y is associated with a 1.6 percentage point increase in students' ECO220Y holding row constant. Notice I used the phrase "is associated with," because I could not infer causality because MARK_100 is endogenous and hence correlated with the error. The intercept, -55.7, has no interpretation since there is no such thing as row 0 and a student would not be allowed to enroll in ECO220 if they had a 0 percent in ECO100.

Both of the slope coefficients are of the expected sign. We thought that sitting further away from the front of the room would harm students' marks (alternatively, sitting closer to the front of the room would benefit students' marks). Also, it is not surprising that students who earned high marks in ECO100 tend to earn high marks in ECO220.

**(b)** $H_0$: $\beta_1 = 0$ and $H_1$: $\beta_1 < 0$ and t = -6.82. Rejection region, $\nu = n - k - 1 = 250 - 2 - 1 = 247$ and $\alpha = 0.05$, the rejection region is t < -1.645. Hence we would reject the null hypothesis and infer that we have sufficient evidence to infer the research hypothesis is true. (Note: The research hypothesis implies a one-tailed test. STATA reports the results for a two-tailed test of statistical significance.)

**(c)** Yes. There is always regression towards the mean when we have an error term (i.e. the $\varepsilon$ in the multiple regression model). So why is the slope = 1.57 which is clearly > 1? The reason is that there is a difference in the standard deviations of marks for ECO100 and ECO220. This is not unusual because those students who are in ECO220 tend to have quite high ECO100 marks (two reasons: most ECO220 students are economics majors or commerce students (which is not true of ECO100 students) and because there is a minimum grade requirement for enrolment in ECO220). This combined with the fact that marks cannot be bigger than 100, implies the standard deviation of marks in ECO100 is smaller. Remember that the slope is only guaranteed to be less than 1 if the variables are standardized. These variables are not standardized and have substantially different means and standard deviations.

**(d)** These results are not surprising. We see that if we do not control for student's marks in ECO100 we get somewhat less precise results in terms of the standard errors and a bigger R-squared because there is more noise across students. Much of the variation in students' marks in ECO220 is explained by variation in their marks in ECO100. Of course marks in ECO100 pick up a student's study habits, effort, interest in economics, analytical skills, etc. all of which are relevant for ECO220 performance. In this simple regression we do not control for these differences across students which means that the variance of the error gets bigger: the error term becomes relatively more important. We can see this in the results by noting that the standard deviation of the residuals increases to 11.818 from 8.0958.

**(e)** Observational data could have been collected in our course. I could have marked down where students *choose* to sit. In this case ROW would be an endogenous variable. Students are free to sit where they want, which means that ROW would be subject to the choices and behaviors of individual agents (students). In the previous parts we described experimental data where the ROW was randomly assigned, which means that in the experimental data students were NOT allowed to sit where they wanted to. Now, let's explore problems that an endogenous ROW creates. We need to consider what factors would affect a student's choice of ROW (seat) and would affect their mark in the course: these are factors that would cause a violation of Assumption #6 by creating a correlation between the observed variable ROW and the other unobserved factors that are in the error term ($\varepsilon$). There are many possibilities. Here are a few: students more interested in the subject matter may sit closer to the front, students that like the professor may sit closer to the front, students that arrive on time may sit closer to the front. It is important to note these are tendencies: this does not mean that an individual student that is highly interested, likes the professor, and arrives on time would never sit in the back of the room. We're talking about on average. These tendencies mean that students that *choose* to sit at the front of the room will tend to do better *not only* because they sit near the front *but also* because of why they chose to sit at the front (interest, like of professor, arrive on time). Hence we would expect a negative correlation between ROW and the

error. Students with positive values of $\varepsilon$ likely to choose to sit near the front (ROW is low) and students with negative values of $\varepsilon$ likely to choose to sit near the back (ROW is high). This violation of Assumption #6 would tend to cause the slope to be downwardly biased. Given that we expected a negative slope on ROW, this means that we would get an estimate that was too negative: we would tend to overstate the harm from sitting in the back of the room. Put another way, we would tend to overestimate the benefit of sitting at the front. We are confounding the benefits of sitting towards the front with the other positive (from a marks-perspective) characteristics that tend to lead students to choose to sit near the front. This is problematic because my research hypothesis would be that, *other things equal*, students would benefit from moving towards the front of the room. This means that the biases created by the observational data, where other things are NOT equal, work in my favor. This is problematic because nobody is going to be convinced of my research hypothesis if I try to prove it with observational data.

**(9)** Starting with the basic summary statistics of the standardized variables, unsurprisingly, the means are all zero and the standard deviations are all one. Because standardization, which is taking each variable and subtracting its mean and then dividing by its standard deviation, removes the units of measurement, all parts of the regression output that are not unit-free will change. However, any unit-free measures, like the R-squared, t test statistics, and P-values will be identical in both regressions (i.e. with and without the variables being standardized).

Also, remember that in a simple regression standardizing both variables means that $b = r$. (See Slide 13 in Lecture 5.) Of course you must also remember that interpreting coefficients in multiple regression is different and each coefficient does *not* represent the relationship between y and that x variable, but rather the relationship between y and that x variable once you control for the other included x variables. For example, the coefficient .6310485 is interpreted as "Houses with living area that is one standard deviation higher, holding number of bedrooms, number of bathrooms, number of fireplaces, and age constant, sell for prices that are 0.61 standard deviations higher on average."

Also, regarding the value .999999994 in the standardized regression under Total MS, why isn't it 1? It is the variance of the standardized price, which should be exactly 1. The reason just has to do with machine precision. STATA cannot record values infinitely precisely so you end up with numbers that are very close to, but not exactly equal to, the theoretical values.