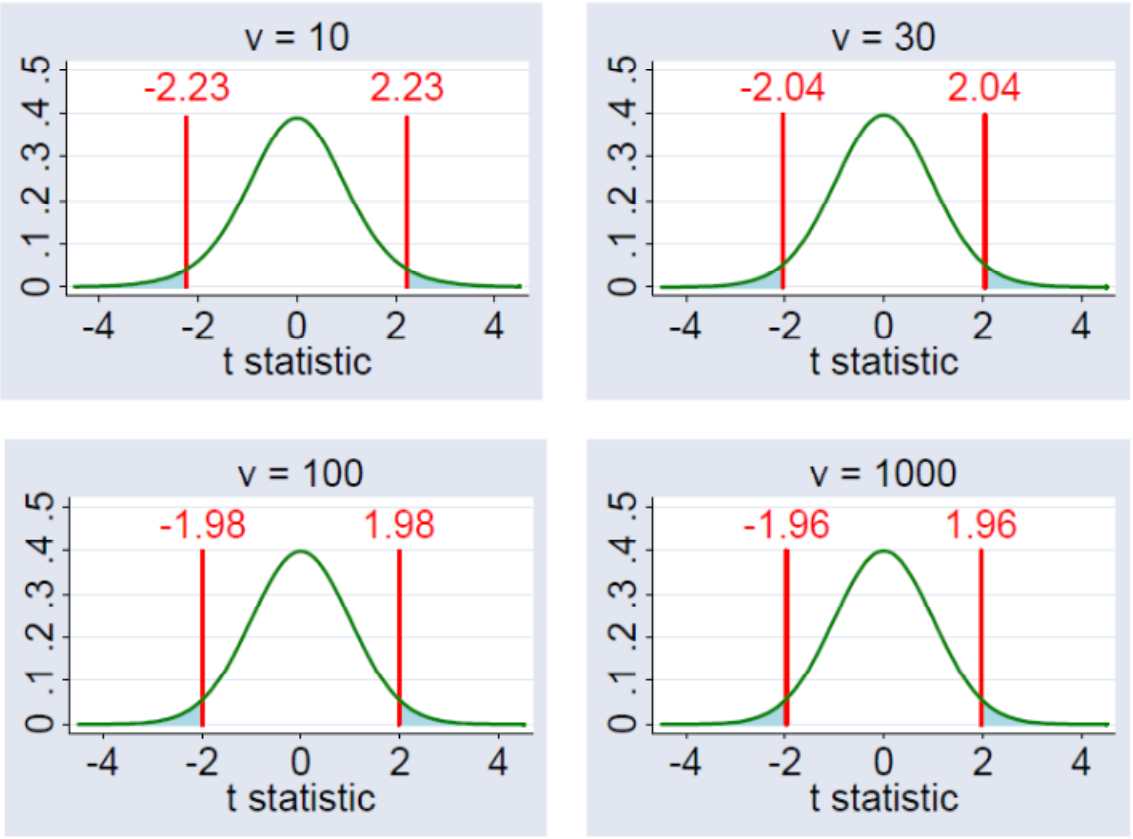
ECO220Y: Homework, Lecture 21 – SOLUTIONS

(1) It is based on the conventional significance level of 5%. It is roughly right for a range of degrees of freedom:



(2) Practice with given parameters' estimates and their standard errors and apply the "rule of thumb": reject H_0 if t>2 or t<-2.

(3)
(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. Alternatively, on average moving students ten rows further away from the front of the classroom decreases their score by about five percentage points. 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. 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 enrol 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)
$H_0 : \beta_1 = 0$

$H_A : \beta_1 < 0$

$$t = \frac{-0.4845315 - 0}{0.0710174} = -6.82\,.82$$

Rejection region, ν = n – k – 1 = 250 – 2 – 1 = 247 and α = 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) These results are not surprising. We see that if we do not control for student's marks in ECO100 we get less precise results in terms of the standard errors and 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 students study habits, effort, interest in economics, analytical skills… 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 error of estimate increases to 11.818 from 8.0958.

(d) Observational data could have been collected in our course, for example. 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 behaviours 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 #5 by creating a correlation between the observed variable ROW and the other unobserved factors that are in the error term (ε). 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 ε likely to choose to sit near the front (ROW is low) and students with negative values of ε likely to choose to sit near the back (ROW is high). This violation of Assumption #5 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 favour. This is problematic because nobody is going to be convinced of my research hypothesis (and rightly so)!

We'd have another problem with this observational data: students may not attend every lecture, which would greatly complicate the analysis and would be another source of bias (selection bias).