R Applications for Multiple Regression Model: Inference

Introductory Remarks

This file builds on the previous "R Activities". If you have not yet gone through these files, please make sure to do so before tackling the "R Applications for Multiple Regression Model: Inference".

Loading the Relevant Dataset

Before looking loading the relevant dataset, make sure to clear your work environment in R, install and/or load the relevant packages, and set your working directory. The accompanying R script does this step by step and a detailed description of how to do this was provided in the R introduction in "R Activity 1".

The dataset we will work with today is called "GPA1.dta." Instead of loading it into R as we did so far, we will call upon this dataset "within R." What does this mean? Some datasets are saved "in R" to work with as examples. Others are saved within packages to be easily accessible. In our case, we are working with a dataset that is used in the coursebook by Wooldridge. All of these datasets are available within a package called "wooldridge" in R. To load the dataset for today make sure that you have installed and loaded the "wooldridge" package and then type

```
data("gpa1") # loading the GPA data
```

The dataset should now appear in your environment under the name "gpa1."

To get a sense of what this data tells us, let's explore it using three already familiar commands, namely

```
View(gpa1) # look at data
head(gpa1) # only display first few rows
summary(gpa1) # summarize data
```

Notice that the data has 29 variables and 141 observations. We will only care about a few of these, namely "colGPA," "hsGPA," "ACT," and "skipped." "colGPA" represents the college grade point average (GPA) for each individual in our dataset, "hsGPA" is the high school GPA for that same individual, and "ACT" denotes their ACT score. Lastly, "skipped" is a variable denoting the number of classes the student skipped. We will use "colGPA" as our outcome and the other variables as regressors. In other words, for each student/individual in our data, we are interested in analyzing the relationship between their college GPA and their high school GPA, their ACT score, and a variable indicating how often they skipped class.

Mathematically, we are therefore interested in the following multivariate regression

$$colGPA_i = \beta_0 + \beta_1 hsGPA_i + \beta_2 ACT_i + \beta_3 skipped_i + \epsilon_i$$

We already know how to do this in R, namely by typing the following commands

```
multivariate_OLS <- lm(colGPA ~ hsGPA + ACT + skipped, data=gpa1) # run OLS regression
summary(multivariate_OLS) # summary of results
```

Looking at the output we find the following estimates: $\hat{\beta}_0 = 1.390$, $\hat{\beta}_1 = 0.412$, $\hat{\beta}_2 = 0.015$, $\hat{\beta}_3 = -.083$. While you can by now interpret these coefficients yourself, notice that the correlations make intuitive sense: a student's high school GPA and their ACT score are positively correlated with their college GPA, while the number of times they skip a class is negatively correlated with their college GPA. In other words, students with higher high school GPAs and ACT scores tend to have higher college GPAs and students who skip class more often tend to have lower college GPAs.

Testing a Single Hypothesis

In this R application we will focus on the standard errors of our estimates, something we have completely neglected so far. Looking at standard errors is useful when thinking about inference or, put differently, when thinking about whether our estimates are statistically significant. Consider again the output to our multivariate regression by typing

```
summary(multivariate_OLS) # summary of results
```

Notice that next to the coefficients, R automatically displays the standard errors associated with the respective coefficient *under the assumption of homoskedasticity.* Next to the standard error, you see the t-statistic associated with each coefficient. Lastly, R also displays the p-values.

Null hypothesis. When we look at coefficients, we generally want to know whether a coefficient is statistically significantly different from zero or not. To do this, we can specify a null hypothesis that the true coefficient is zero. For example, suppose we are interested in knowing whether the coefficient on "skipped" is statistically significantly different from zero. The null hypothesis in this case is

$$\beta_3 = 0$$

The two-sided alternative hypothesis in this case is

$$\beta_3 \neq 0$$

t-statistic. One way to test whether a coefficient is statistically significantly different from zero is to conduct a t-test. As we now know, R automatically calculates the t-statistics for us. In this case, we find a t-statistic of -3.20. Note that you can calculate this statistic yourself: $(-.083 - 0)/.026 = -3.20$. We now want to compare this value with some critical value, that we can find, for example, in a table of the t-distribution (as provided in Wooldridge for example). Luckily for us, R also allows us to compute the critical value as follows

```
p = .05
df = 141-4
qt(p/2, df, lower.tail=FALSE) # critical value for two-sdied test
```

$p = .05$ above is the significance level. In other words, we are here looking at a 5% significance level. You can of course change this to other significance levels such as 1% or 10%. $df = 141 - 4$ are the degrees of freedom. We know that these are equal to the sample size (141 in our case) minus the number of parameters we estimated (4 in our case). To then find the critical value, we use the "qt" function. Its first argument is the significance level p. In our case, since we are doing a two-sided test, we need to divide p by 2 (i.e. we want 2.5% on either side). The second argument are the degrees of freedom. The third argument "lower.tail = FALSE" tells R that we are interested in the "probability to the right" of p. Notice that since we are doing a two-tailed test, if you wrote "lower.tail = TRUE" you would get the same critical value, just with a minus in front (try this!). In our example, we find a critical value of 1.977.

Given this critical value, we can now compare the absolute value of our t-statistic (3.2) with 1.977. Clearly, $3.20 > 1.977$, and hence we conclude that our estimate of $-.083$ is statistically significantly different from zero at the 5% significance level and reject the null hypothesis stated above. Recall that as n (the sample size) increases, the t-distribution can be approximated by the standard normal distribution. In our case, if we feel comfortable to argue that our sample size of $n = 141$ is large enough, then we could use the critical value from the standard normal distribution. We know that for 5% significance this would be 1.96 (indeed, almost the same as what we calculated above). We would reach the same conclusion.

In case you were testing a one-sided alternative hypothesis (instead of a two-sided one), we would need to find a different critical value. We can do this in R as follows

```
qt(p, df, lower.tail=FALSE) # critival value for one-sdied test
```

In this case, since we are computing critical value for a one-tailed test, we do not divide p by 2. The rest of the command remains the same. In what follows we will continue with the two-sided hypothesis from above.

p-values. Another way to test whether a coefficient is different from zero is to compute the p-value associated with said coefficient. Again, R does this for us. By comparing the calculated p-value with 0.05, we reject the null hypothesis if the p-value is less than 0.05 (at the 5% level). For testing the hypothesis above, we find a p-value of 0.00173, which is clearly less than 0.05. So we, reassuringly, draw the same conclusion as above when using the t-statistic: we reject the null hypothesis. Note that the p-value automatically let's you "see" up to what significance level you can reject the null hypothesis. In our example, we would reject the null hypothesis at the 1% level as well.

In the lecture you were told that the p-value from a F-test as well as the one from a t-test are the same when testing a single hypothesis. We have above seen that the p-value from the t-test is 0.00173. To conduct a F-test, we can type (please make sure to install and load the package "car" before running this command)

```
linearHypothesis(multivariate_OLS, c("skipped = 0"))
```

To conduct a F-test in R we use the "linearHypothesis" function. Its first argument is our model, which in our case is called "multivariate_OLS." The second argument specifies what we want to test. In our case, we want to test whether $\beta_3 = 0$, which is expressed as the variable name associated to this parameters in R. We will learn how to test multiple hypotheses using this command below. We find a F-statistic of 10.22 and a p-value of 0.00173, which is indeed the same we found above.

Confidence intervals. A third way to check whether a coefficient is statistically significantly different from zero is to compute the confidence interval for each coefficient and to see whether 0 is in this interval. Let's consider the 95% confidence interval for the coefficient on skipped. We know that this should be approximately $[-.083 \pm 1.96 * 0.026] = [-.134, -.032]$. We could ask R to compute this for us as follows

```
confint(multivariate_OLS, 'skipped', level=0.95)
```

As you can see, we get the same confidence interval. Since zero does not lie in this confidence interval, we once again draw the same conclusion as above: we reject the null hypothesis. You can try to calculate the confidence intervals for other variables in our model and using different levels of confidence yourself.

A new null hypothesis. Suppose that instead of the above, we now want to test the following null hypothesis

$$\beta_1 + \beta_2 = 0$$

against its two-sided alternative

$$\beta_1 + \beta_2 \neq 0$$

In words, we are here testing whether the sum of the parameters on "hsGPA" and "ACT" is statistically significantly different from zero or not. Notice that this does not imply that we are testing multiple hypotheses. There is still only one hypothesis here. The difference to the above is that above we tested whether one parameter was equal to zero, whereas here we test whether the sum of two parameters is equal to zero. A simple way to find out whether we are testing a single or multiple hypotheses is to count the equal signs.

One way to go about testing this hypothesis is to follow what you learned in the lecture, i.e. to compute the t-statistic and then to compare its absolute value against the same critical value that we found above.

Alternatively, we can do this in R directly by running the following command

```
linearHypothesis(multivariate_OLS, c("hsGPA + ACT = 0"))
```

As above, we use the "linearHypothesis" function in R to conduct a F-test. While the first argument is again the same as above, the second argument looks a bit different here since we want to test a different hypothesis. Specifically, we want to test whether $\beta_1 + \beta_2 = 0$, which is expressed as the variable names associated to these parameters in R. Note that you could rewrite this command as follows to get the same result

```
linearHypothesis(multivariate_OLS, c("hsGPA =- ACT "))
```

Either command will yield an F-statistic of 22.258. While we may not know the critical value of the F-distribution by heart, R conveniently also provides us with a p-value. In this case, the p-value is $5.809e^{-.06}$,

which is basically zero (or surely less than 0.01). Hence, we conclude that we can reject the null hypothesis that the sum of the coefficients on "hsGPA" and "ACT" is zero (at the 1% level).

We have therefore seen that we can use F- and t-tests to test single hypotheses. When testing multiple hypotheses, however, you will have to rely on only the F-test. The t-test will not be applicable in that case.

Testing Multiple Hypotheses

Now, suppose we want to test the following null hyoptheses

$$\beta_1 = \beta_2 = \beta_3 = 0$$

In words, we are testing whether all our coefficients (except the intercept) are jointly zero. Be aware that we are now testing multiple hypotheses (3 to be precise).

To test multiple hypotheses, we want to conduct an F-test. Luckily, R again provides a simple way to do this, using the same function we used above

```
linearHypothesis(multivariate_OLS, c("hsGPA = 0", "ACT = 0", "skipped=0"))
```

In other words, we use the same function as above, but specify our multiple null hypotheses individually. The output provides us with an F-statistic of 13.92 and a p-value of $5.653e^{-08}$. Clearly, we reject the null hypotheses that our three estimates are jointly zero at the 1% level.

Look at the summary of our original regression again by typing

```
summary(multivariate_OLS) # summary of results
```

Notice that the last line provides you with an F-statistic and a p-value and notice that these are equivalent to the F-statistic and p-value we just calculated. Hence, if you want to test the null hypotheses that all your parameters are zero (except the intercept), you can rely on the summary output of our regression without having to compute the F-statistic. However, if you wanted to test whether β_1 and β_2 are jointly zero only (i.e. omitting β_3 in our example above), then you cannot rely on this output and have to compute the F-statistic yourself. In general, it is good practice to compute the F-statistic yourself.

Statistical Significance vs. Economic Significance vs. Causality

To conclude the R application activity, notice that we here discussed statistical significance. In other words, we checked whether some coefficient is statistically significantly different from zero. This is a statistical property. In the end, what we care about is economic significance. Statistical significance does not guarantee economic significance. For example, it is possible that some effect is statistically significant, but economically irrelevant. This will have to be decided on a case by case basis.

Notice also that statistical significance does not imply a causal relationship. Statistical significance just tells us that the correlation between two variables (e.g. between "skipped" and "colGPA" in our example) is not zero. But it does not tell us whether this is a causal effect or not.