**Lab Assignment Week 04**

**I. Objectives**: The primary objective is to examine how the error variance is estimated and how the standard errors for the regression coefficients are calculated. A secondary objective is to examine model specification questions, particularly how we balance the costs of *under-*specification (omitted variable bias) against those of *over-*specification (irrelevant variable bias and multicollinearity). In so doing, we will explore measures of regression model fit.

**II. Datasets**: *caschool.dta*

**III. Packages**: *foreign*, *car*, *stargazer*

**IV. Preparation**

1. Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.
2. Open the project from the Projects menu
3. Pull from Github for any updates
4. Navigate in the lower right file area to the Lab 3 section
5. 5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.
6. Run lines 1–7 in the R script to load three packages that you will need. If any of these are not already installed, you need to install them using the **Packages** tab or with code like in line 3.

**V. Instructions for Lab Week 04**

You will use the same dataset from Lab Week 03, exploring the factors that determine test scores in school districts in California.

Begin by loading the dataset, typing the following line, changing the directory if needed:

cadta<-read.dta(here("data","caschool.dta"))

This command is shown in line 11 of the R script. If you failed to load the *foreign* package earlier, then R will not load the Stata dataset. If you did not load the *here* package, you will need to specify the working directory manually and move the file to the correct working directory. Ensure that the package is both installed and loaded, and then re-run the line of code.

Recall that *CaliforniaTestScores.pdf* lists the variables you will use, and some that you will not. Some key demographic variables are the average income of families in the district, in thousands of dollars (avginc), the percent of students qualifying for the CalWORKs program, which is the state of California’s name for welfare (calw\_pct), the percent of students qualifying for reduced-price lunch (meal\_pct), and the percent of students who are learning English as a second language (el\_pct). Three other potentially influential variables are the number of computers per student (comp\_stu), the district’s expenditures per student (expn\_stu), and the number of students per teacher (str).

Line 12 changes R Studio’s default to *not* use scientific notation. If you do not run this line of code, then the first time you run a correlation (line 21) you will –1.79e–15, which is actual –1.79 x 10–15, or 0.00000000000000179, or virtually zero. Line 13 changes R Studio’s default to round to three digits for easier examination of the correlation matrix, speaking of which …

Run line 14 to create the *correlation* matrix for the seven explanatory variables that are listed above; if you look down the first column you can see how each explanatory variable correlates to avginc, which was an important explanatory variable in last week’s lab. You may wish to scan across individual rows or down columns to identify associations that appear particularly strong. For example, variables that measure relative affluence of a district – avginc, calw\_pct, and meal\_pct – are associated strongly but not necessarily *positively*.

If you prefer graphical illustrations of association, run lines 16-17 to display scatterplots of the first four variables (districts’ demographics) and the last three variables (districts’ policies), respectively.

***A. Omitted Variable Bias***

Start by running line 19 to estimate a baseline model, which regresses average test scores in a district (testscr) on average income (avginc) and the percent of students who are English-language learners (el\_pct).

What is the baseline model’s R2? \_\_\_\_\_\_ . What is the model’s residual standard error? \_\_\_\_\_\_ .

What are the coefficient and standard error for the second variable, el\_pct, i.e., the percent of students who are English language learners ( = \_\_\_\_\_\_, and *se*() = \_\_\_\_\_\_ ).

Run line 20 to check that the mean residual from the baseline model equals zero, and then run line 21 to check that the residuals are uncorrelated to the **included** variables.

Next, run line 22 to explore whether the residual from the baseline model is related to the two **omitted** measures of affluence: the percent of students whose families qualify for welfare (calw\_pct) and the percent of students who qualify for reduced price lunch (meal\_pct). If either correlation is sufficiently large then you might consider a variable to be “relevant” and potentially worthy of inclusion in the model’s specification.

Run line 24 to estimate a model that adds the variable calw\_pct to the specification; what is this second model’s R2? \_\_\_\_\_\_ . What is this second model’s residual standard error? \_\_\_\_\_\_ .

Write down the values of the coefficient and standard error for el\_pct, i.e., the percent of students who are English language learners ( = \_\_\_\_\_\_, and *se*() = \_\_\_\_\_\_ ) in this model.

Also write down both the coefficient and standard error for calw\_pct, i.e., the percent of students who qualify for welfare ( = \_\_\_\_\_\_, and *se*() = \_\_\_\_\_\_ ) for later reference.

Next, run line 26 to estimate a model that adds the variable meal\_pct to the specification; what is this third model’s R2? \_\_\_\_\_\_ . What is this third model’s residual standard error? \_\_\_\_\_\_ .

Write down the values of the coefficient and standard error for el\_pct, i.e., the percent of students who are English language learners ( = \_\_\_\_\_\_, and *se*() = \_\_\_\_\_\_ ) in this model.

Also write down the new coefficient and standard error for calw\_pct, i.e., the percent of students who qualify for welfare ( = \_\_\_\_\_\_, and *se*() = \_\_\_\_\_\_ ), which was in the second model as well.

You should observe that adding more indicators of affluence improved the model fit, represented by increases in the R2 and decreases in the residual standard error (a.k.a. *sigma*). Each time a variable is added, the model gets less noisy, but this has come at a cost. All the variables’ coefficients become smaller (in absolute value) in each successive model, and one variable’s coefficient becomes statistically insignificant when the percent of students who qualify for free meals (meal\_pct) is added to the model.

Run line 20 to create a correlation matrix which shows the interrelationships among these four measures of affluence as well as their correlations to districts’ test scores. The variable that we added to the specification last, meal\_pct, has the strongest bivariate correlation to test scores and has the strongest correlations to the other variables. If you inspect the equations for the multiple regression coefficients (from lecture), perhaps you can see why adding this variable changes the other coefficients and their standard errors so dramatically.

***B. Multicollinearity***

In lecture, we will discuss using the Variance Inflation Factor as a measure of the degree of multicollinearity. Run line 30 (which uses the *car* package) to calculate VIF’s for each of the variables using the third model we estimated. Which variable has the highest VIF? What was the threshold that Wooldridge asserts we should use as a rule of thumb for deciding whether to drop a variable? What was the justification for using this threshold?

Run line 32 to check how the VIF is calculated for one explanatory variable, meal\_pct. The code asks R to regress meal\_pct on the other regressors and then find the R2 of this regression. Fill in your own code in this space to check that the VIF for meal\_pct equals 1/(1-R2):

>

Suppose we decided to drop calw\_pct from the specification, because it was statistically insignificant in the last regression. For instance, you can run line 34 to estimate the “final” model:

> final <- lm(testscr ~ avginc + el\_pct + meal\_pct, cadta)

How does this final model’s fit (R2 = \_\_\_\_\_ ) compare to the third model (from line 26)?

How does this model’s residual standard error (*sigma* = \_\_\_\_\_ ) compare to the third model?

You can also re-check the variance inflation factors, as shown in line 35:

> vif(final)

If you were advising another student about a potential multicollinearity problem in their research, what would seem to be the sounder advice: dropping the least significant variable or dropping the variable with the highest VIF? Or is there another option worth considering?

The major effect of ‘near multicollinearity’ is on the standard errors of the estimated coefficients. Take some time to focus on how these numbers are calculated. The estimated error variance (designated by in Wooldridge) is the ratio of the residual sum of squares divided by *n* – *k* – 1; line 37 asks R to compute the estimated error variance, which equals \_\_\_\_\_\_\_.

Line 38 asks R to compute the square root of the previous term. The “Residual standard error” (designated by in Wooldridge) is the square root of the estimated error variance. Write down the result shown in the regression table (Residual standard error = \_\_\_\_\_), and check that you get the same result when you run line 38 and the following line of code:

> summary(final)$sigma

Run lines 40-41 to generate the ***variance-covariance matrix* *of the estimators***. (I put this in ***bold and italic*** print, because we will be using this matrix more and more over coming weeks.)

> vce <- vcov(final)

The diagonal elements of this matrix are the variances of the coefficient estimates. If you take the square root of any diagonal element, you compute the standard error for the estimated coefficient!

Try it: run line 42 to take the square root of the second diagonal element, which is avginc:

> sqrt(vce[2,2)

The result equals \_\_\_\_\_\_; the standard error for the coefficient for avginc shown by line 34 equals \_\_\_\_\_\_.

Line 44 provides a summary of the residual from the ‘final’ regression. Obviously the mean residual should equal 0, but when you type line 45, R also tells you the standard deviation of the residual (which equals \_\_\_\_\_ )… which is slightly *smaller* than the Residual standard error (which equaled \_\_\_\_\_ ); do you know why? If not, ponder the equations until you figure this out.

Line 46 asks R for a histogram of residuals.

***C. Model Fit***

Let’s review how the *R*2 statistic is calculated. Run lines 48-51 in the R script:

# Model Fit

final.res<-resid(final)

resf<-final.res^2

ssr<-sum(resf); ssr

In your own words, describe what the residual sum of squares is, and then describe what steps are being taken when you these three lines of code. Please write down the value of the residual sum of squares (SSR = \_\_\_\_\_\_ ).

Next you can, obtain the total sum of squares by running lines 53-56:

summary(cadta$testscr); ybar <- mean(cadta$testscr)

devy <-(cadta$testscr-ybar)

devysq<-devy^2

sst<-sum(devysq); sst

Please write down the number that results (SST = \_\_\_\_\_\_ ) and describe in your own words what steps were taken to arrive at that number.

Now that you have the residual sum of squares (named *ssr*) and the total sum of squares (named *sst*), fill in for yourself one line of R code could you use to find the R2.

>

***D. Multicollinearity Redux***

Let’s keep our focus on the final regression (from line 34) with three explanatory variables (avginc, el\_pct, and meal\_pct), but now with a smaller sample. The reason for doing this is that multicollinearity is sometimes referred to as “micronumerosity,” because it results from a problem with the sample – too little “free” variation in one or more explanatory variables – that can be addressed by expanding the sample. In section 3.4, Wooldridge contends that multicollinearity and small samples are related problems, as you will soon observe.

Run line 61 to subset the data, selecting just the school districts in Los Angeles County, Orange County, and Ventura County; combined these three counties have 47 districts, compared to 420 in the entire state. (You can run lines 59-60 to see how many districts are in each county.)

Run line 63 to estimate the final model but with the smaller sample. Earlier you loaded the *stargazer* package; use that package now by running line 64, which will create a table of regression results allowing you to more easily compare models using the same specifications but different samples.

Compared to the same model that used the full sample, (a) what has happened to the model fit? (b) what has happened to the estimated coefficients? (c) what has happened to the standard errors of the coefficients?

Line 65 brings back the diagnostics you used earlier to assess multicollinearity. Are there any VIFs that we should be concerned about? Compare this output to line 34’s output.

Run line 66 to examine our small sample’s correlation matrix; compare this to line 28’s output. How did the correlation between el\_pct and the other two explanatory variables (avginc and meal\_pct) change from the large sample of 420 school districts to the smaller sample of 47 school districts?

Finally, run line 67 to examine our small sample’s scatterplots. If you compare this to line 10’s output, you should observe that the scatterplots involving el\_pct and the other two explanatory variables look more line-like than cloud-like in the smaller dataset.