**Module 5 HW**

Professor Nick Williams Fall 2024

Economics 4010, University of Cincinnati

## **What are you being asked to do:**

Estimate and interpret multiple regression models, and test for significance

## Why is this important?

The MLR model is the workhorse of econometrics, used in many, if not most econometrics applications. Hence learning how to use, and more importantly, how to interpret the results from the MLR model is one of the fundamental learning objectives of the course.

## Directions

* Turn in your answers in a Word document through Canvas.
* You **DO** need to turn in a copy of your R script for this homework.
* Make sure you look at my example R scripts from the lectures and learning exercise. I am NOT asking you to use any R code that we have not already used in at least one of those R scripts.
* In some circumstance, copying results from RStudio into Word is acceptable, but note that in many instances I ask you to interpret or explain. Below I make it clear when I want to write an answer.

Make sure you read and follow my directions in the companion Word document “Practicing Professionalism” that I distributed with the Module 2 HW. **Important: Lack of producing a neat and organized homework may result in a 10-point deduction from your total score!**

* The homework will be carefully graded out of 100 total points.

## Questions

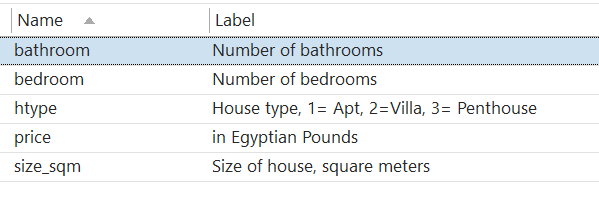
1. A sample of house sales are given in the data ***egypt\_housing\_clean.dta***. We want to use this to explore the determinants of the sales price of the house in egypt. You will do so by running three different models, which will be reported in a nice table produced by the R package “stargazer”.

The dependent variable should be *price/1000000*. The two basic models reported should be:

* Base: Include only *size\_sqm* and *bedroom* as independent variables (note that bedroom = 0 is a studio)
* Base+1: Starting from Base model, add the *bathroom* model

You will estimate these regressions for two different subsets of the data and eventually create a table with 4 columns using “stargazer” (see my R Hints for how to do this). **Turn in this table in your Word document, best presented as the last page of your homework.**

Relevant variable definitions:

**

**Since the Egyptian pound has experienced dramatic inflation over the past several years, create a new variable equal to price/1000000. That is, the price will be in millions of Egyptian pounds. Use this new variable as the dependent variable in your analysis below.**

(Not graded) Examine the distributions of the variables in the data set using appropriate summary statistics and visualizations. Make sure you understand their units which will be important for your interpretations below.

**For parts a) and b) create a subset of data that only includes apartments. You should run the Base and Base+1 regressions on this subset.**

* 1. In the Base model, interpret the coefficients on *size\_sqm* and *bedroom.*

In the base model, or the model which just contains size\_sqm and bedroom, the coefficient on size\_sqm is 0.022 and bedroom -0.414. This would indicate that for each additional 1 square meter the cost of the apartment would increase by 22000 egyptian pounds. As far as the coefficient on bedrooms, since we are holding size constant, as seen in other exercises, this has a negative effect on price, most likely due to the fact that the additional bedrooms are smaller. So a coefficient of -0.414 means that for each additional 1 bedroom, the cost of the apartment goes down by 414000 egyptian pounds.

* + 1. Also discuss their statistical significance.

Each of the pvalues for these 2 independent variables is 0.000, which means we can reject the null hypothesis that these 2 variables are statistically insignificant. (Null hypothesis is that the coefficients are statistically 0, and Alternative is that the coefficients are statistically significant). The R^2 value is 0.223 though which leaves about 78% of the variance up to other variables.

* + 1. Also discuss their substantive or economic significance.

I think that the size of the apartment does have a significant affect on its price, and should be taken into account. The number of bedrooms, while holding size constant though I think has less of an effect and significance. As we see it is a negative coefficient meaning that as there are more bedrooms (with size held constant) price actually goes down, which in reality just makes no sense.

* 1. In the Base+1 model, test the overall significance of the model (hint this is not a simple t-test of the significance of a single coefficient). Your answer should include
  + What is the null hypothesis? What is the alternative hypothesis?

H0 = the coefficients of the independent variables have no statistical significance (ie = 0)

HA = the coefficients of the independent variables do have statistical significance (ie != 0)

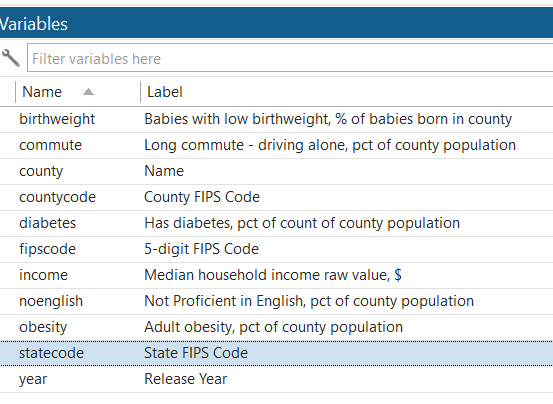
* + The algebraic formula for test statistic and its numeric value in this case.

TS = (Bj) / se(Bj) ~ TN – (k + 1)  = 811.321

* + Your decision and the reasoning behind your decision for these hypotheses

We can reject the null hypothesis, this model does have overall significance.

1. Use the data frame *county\_health\_subset\_2238.dta* to answer this question. The data frame is similar to the one used in the M03 homework, but includes some additional variables. Variables include:



* 1. Create the new variable, *incometh* which is the median income of the county in thousands of dollars.
  2. (Not graded) Examine the distributions of the variables in the data set using appropriate summary statistics and visualizations. Make sure you understand their units which will be important for your interpretations below.

The dependent variable should be *birthweight*. The two basic models reported should be:

* Base: Include only *incometh and obesity* as independent variables
* Base+1: Starting from Base model, add the *diabetes and noenglish* variables to the model
  1. In the Base model interpret the coefficient on *incometh.*

The coefficient on incometh is -0.040 which means that as income increases by 1000, babies born with a low birthweight decreases by 0.04 percent.

* + 1. Also discuss its statistical significance.

This independent variable would be considered statistically significant via the P val test due to the fact that the pval is 0.000 which means that it is lower than any level of alpha we set, thus rejecting the null hypothesis.

* + 1. Also discuss its substantive importance.

Honestly I don’t think it seems substantively that important, 0.04 % is pretty low, decrease per 1000 dollars income increase, overall it seems that wealth does not play a role in determining whether babies are overweight when born.

* 1. In the Base model, test the overall significance of the model (hint this is not a simple t-test of the significance of a single coefficient). Your answer should include
  + What is the null hypothesis? What is the alternative hypothesis?

H0 = the coefficients of the independent variables have no statistical significance (ie = 0)

HA = the coefficients of the independent variables do have statistical significance (ie != 0)

* + Your decision and the reasoning behind your decision for these hypotheses.
  1. Consider the differences between the Base and Base + 1 models
     1. In particular, what happened to the effect of *incometh* on *birthweight*?

It almost has turned to 0 now, being -0.001, which indicates that it really has no effect on birthweight variable.

* + 1. What happened to the statistical significance of *incometh*?

It has no statistical significance now, pval has jumped way up to 0.6 indicating that we cannot reject the null, meaning it likely has a stat significance of 0.

* + 1. Why might these changes have occurred? Briefly discuss.

We added in these new variables of diabetes and non English speaking. These variables picked up a lot of slack in the model in terms of error and inaccuracy.

* 1. What happens to the R2 between the two regressions? Is that what you would expect? Briefly discuss

The R^2 in the first model where we regress birthweight onto just incometh and obsesity is .212 . The 2nd model which further includes diabetes percent and non English speaking percent of each county it goes up to .456 . The R^2 value will always increase as more variables are added but it did have a bit more of a jump indicating that some variance was accounted for due to these variables being added.

## R Hints

# The “stargazer” package is an easy way to display regression results in the console.

I have a simple example of this in the script for the Module 5 Learning Exercise. There are many additional options you can explore. For one source, see:

<https://www.rdocumentation.org/packages/stargazer/versions/5.2.2/topics/stargazer> for one source.

The options I have found most useful are:

* report=”????” where ? is replaced by what you want shown in the table. This should include v = variable names, and c = coefficients, and could include “s” for standard errors, “t” for t-values and “p” for p-values.
  + For example, report=”vct” would report variable names, coefficients and t-values
  + Or, report=”vcsp” would report variable names, coefficients, standard errors and p-values
* notes = c("????") where ???? are notes such as “Standard errors in parentheses”
* digits=??? Where ??? is replaced with the number of digits after the decimal point to be displayed
* keep.stats=(“??”, “???”,…”????”) will report statistics from each regression. The most useful are probably
  + “n” – reports the number of observations for each regression
  + “rsq” – reports the R2 for each regression
  + “f” – reports the F-statistic for each regression

You may also find the following stats interesting or useful at times

* + “ser” – reports the standard error of the regression ()