ANA 510 Assignment #6: Model Building

# Part 1: Review (EDA, OLS, marginal effects)

## Introduction

This assignment is intended to give you additional help in model building. In this course and ANA 500 we have encountered a wide variety of variables. Let’s go over some definitions first.

Some properties of **random variables** are:

* A variable whose possible values are numerical outcomes of a random phenomenon,
* A variable that only takes a real value, and
* Independent and identically distributed.

There are continuous and discrete random variables. The values of a random variable will describe a probability distribution. Don’t forget about the Central Limit theorem!

But when we are building models we have to be concerned about a variety of “other” variables, e.g. **indicator** **variables**, **instruments or instrument variables**, and **target variables**. In addition, just to confuse everything there can be mediating, mitigating, moderating, control, and confounding variables.

Starting with these last variables first, **mediating variables** explain how other variables are related. Mitigating variables do exactly what they are called. **Mitigating variable** mitigate the effect of a variable or the relationship between variables. **Moderating variables** affect the strength and/or the direction of the relationship between variables. Confounding variables are probably the worst. **Confounding variables** are unmeasured variables that influence other variables, typically variables that represent a cause and an effect. **Control variables** also do exactly what they are called. They exert control something else, e.g. temperature can control plant growth.

We have not considered these types of variables in ANA 500 or ANA 510 but they do exist and you will need to think about them when you are working with real world data and models. Also, the discussion that follows does not hold for time-series data.

What we do deal with in ANA 500 and ANA 510 are the indicator, instrument, and target variables. To get to where we can understand what these variables are and how we use them let’s start from the definition of a random variable. Let’s change the property of a random variable that said it was a possible numerical value that was an outcome of a random phenomenon. Let’s say that random variables involve data pairs that are obtained by random sampling, a process that results in pairs that are independent and identically distributed. In addition, we assume that we have included all the important variables, we have use “the correct functional form,” and there are no factors that cause the error to be associated (correlated) with our independent or explanatory variable.

Let’s further assume that each of these pairs we obtain through sampling or measurement such that we also have some amount of **error**. I hope that you can tell from the way I have described this that both the independent and dependent variables in the pairs are associated with the error in the samples or measurements that we have made. In the real world, error is impossible to avoid. The response or the dependent variable, even in simple regression, is going to depend on two things: the independent variable and the error. We’re going to call independent or explanatory variables that are not associated (correlated) with error are **exogenous** variables. And we’re going to call independent or explanatory variables that are associated (correlated) with error are **endogenous** variables.

To further illustrate what is going on, let’s look at a specific example. Start from the position that we are going to study the relationship between wages and years of education. We’ll start by assuming that the variable ‘years of education’ fully describe the variation in ‘wages’. However, another variable we might want to consider is ‘intelligence’. Higher intelligence typically leads to higher levels of education which leads to higher wages. If we just consider wages in terms of level of education then we are overstating the effect of education. This means that we have bias in our analysis. It means that, given we are interested in regression, the least squares estimator is inconsistent.

If we apply the “**Method of Moments**,” we assume that we have some error. Without deriving all the mathematics behind the Method of Moments, to handle this we introduce another variable which we call an **instrumental variable**. We use the term “instrumental” because this variable is instrumental in describing the relationship between independent and dependent variable(s). It is a tool or instrument. This instrumental variable is exogenous, i.e. it is not correlated with the error. Instrumental variables have a normal distribution. I have written “variable(s)” because we can use instrumental variables in simple, multivariable, or multivariate regressions. Furthermore, we want to have “strong” instruments. However, they can be weak. We determine the strength of an instrument after controlling for the effect of any and all other exogenous variables.

We implement all this through a method called “Two Stage Least Squares Regression” (2SLS Regression). If we consider just simple regression, from the math it can be shown that one instrumental variable will result in two moment conditions which we can solve for the two unknown model parameters; the instrumental variable and the original independent variable. Essentially, we have split out the part of an endogenous independent variable associated with error into an instrumental variable and a more pure form of the independent variable, an exogenous independent variable.

As a reminder, the assumptions of simple linear regression include:

1. The value of the dependent variable for each value of independent variable, i.e. the pair is:
2. The expected value of the random error is:
3. The variance of the random error is:

(Note that the variance for the variables y and e are the same because they only vary by a constant.)

1. The covariance between any pair of random errors and is:

We can have more instrumental variables than independent or explanatory variables, i.e. surplus moment conditions. For example, we could have three sample moment conditions or three equations but only two unknowns (independent variables or regressors). In this case, we could just throw one of the instrumental variables away. However, there are ways to avoid throwing away information. For any simple regression, what we need or what is required is that we have one independent variable or regressor for each instrument we want to use at a minimum.

## Example 1: Wages (Use the wageDataset.gdt)

Let’s look at a more specific, mathematical example using the example of wages and education. We could formulate an equation for wages as a function of “level of education” or “education”. We get our data from the “wageDataset.gdt”. As usual, we start by considering the dataset and some exploratory data analysis (EDA) as follows:

 

1. How many independent variables are in the wageDataset.gdt dataset?
2. Are the wage data skewed (yes/no)? If so, are these data skewed right or left?
3. The mean wage is \_\_\_\_\_\_\_\_?
4. The median wage is \_\_\_\_\_\_\_\_?

Work through the gretl script provided for this assignment to get the detailed information on the model building associated with these. I/we know there are problems with this model. But rather than go through methods to correct or at least mitigate these problems right now let’s go through how to subset the dataset and consider the effects of things like education on these subsets. In this case, the subsets are groups, i.e. male, female, white, black, etc.

1. The coefficient of the constant or , is -6.71. Is this a meaningful number, i.e. meaningful in the sense that it is realistic or a number that could represent a real situation?
   1. No, because it represents a salary (of someone with no education). And, salaries are not negative.
   2. No, because it represents the number of years of education. And, the number of years of education cannot be negative.
   3. Yes, because it is the value at the point the line representing wage as a function of education crosses the y-axis.
   4. There is no way to know if this is a meaningful number or not.
2. How much is the estimated increased in the expected hourly wage rate for each extra year of education?
   1. 1.98
   2. -6.71
   3. 0.00
   4. -6.71+1.98 = -4.73
3. Consider the residuals as plotted against educ. What is the pattern you see in this plot?
   1. The residuals are diverging, i.e. the variance is larger for larger values of educ.
   2. The residuals are converging, i.e. the variance is smaller for smaller values of educ.
   3. The residuals are equally distributed about zero.
   4. There is no discernable pattern shown in this figure.
4. Who gets a bigger increase in wage rate for an extra year in education? Mark all the boxes that are correct. (Note that this doesn’t say anything about who actually has a higher wage rate.)
   1. Black workers’ increase is larger than white workers.
   2. White workers’ increase is larger than black workers.
   3. Female workers’ increase is larger than male workers.
   4. Male workers’ increase is larger than female workers.
5. For the original linear model, what was the marginal effect for someone with 12 years of education?
6. For the quadratic model, what is the margin effect for someone with 12 years of education?
7. For the log-model, do the residuals look correct now? Yes/no
8. Does the value representing the wage rate of someone with no education look realistic now? Yes/no
9. What are the marginal effects for wage rate in this new, log-quadratic model? (Hint: don’t forget to take the anti-log appropriately since we used the log of wage to build the model!)
   1. 14.80
   2. 0.10
   3. 0.00
   4. 1.61
10. In the comparison of the datasets used, roughly what percentage is educ, or the wage rate which is represented by the variable educ, increased?

# Part 2: Two Stage Least Squares Model Building

## Example 2: Wages cont. (Use the mroz.gdt dataset)

Now, you are working with the second dataset that has expanded variables. First, we’ll build a model using the log of wage and exper-squared. Let’s see what you/we get from that. Then, we’ll expand that model using the same dataset.

1. How many observations are in the mroz.gdt dataset?
2. How many total variables are in the dataset? Hint, don’t include the “const” variable that the program adds.
3. The variable wage (is/is not) \_\_\_\_\_\_ skewed (right/left) \_\_\_\_\_\_\_\_. Fill in the blanks with the best choice of words from the words in parentheses before the blanks.
4. Models for both datasets yield very different results for increases in the variable educ. True/False

After you have compared the same models for both datasets, build a Two-Stage Least Squares (TSLS) Regression Model using mothereduc as an instrument. The first stage is to develop the regression equation for educ as the dependent variable and look at the output from the related least squares model.

1. Why do we use Two-Stage Least Squares Regression instead of Ordinary Lease Squares Regression? Select all the answers that are correct.
   1. When endogenity is present, i.e. an explanatory variable is correlated with the error term.
   2. Because the presence of endogenity indicates that the results of an analysis will be biased.
   3. There is a second explanatory variable that is correlated to a problematic explanatory variable but not to the error term.
   4. Because you like it and Two-Stage Least Squares Regression is easy to understand and use.
2. In general, to start a Two-Stage Least Squares Regression you need to: (Select the “best” choice from the following statements.)
   1. Develop a regression equation including the instrument or instruments as explanatory variables and using the related explanatory variable as the dependent variable in the first stage.
   2. Run a Hausman test to consider whether or not the OLS estimates are consistent.
   3. Determine if the instrument or instruments you want to use are strong or weak.
   4. I don’t know how to start a Two-Stage Least Squares Regression.
3. What do you do to start the second stage of a Two-Stage Least Squares Regression, in principle?
   1. Use the output (outcome) from the first stage and substitute it for the related explanatory variable to develop a second stage regression equation without endogenity.
   2. Use the terms in the first stage to generate a structural equation.
   3. Find the instrument or instruments that are required for the desired analysis.
   4. Eliminate any surplus instruments.
4. The conditions for a valid instrument include: (Select all the following choices that are correct.)
   1. Relevant
   2. Exogenous
   3. Linear
   4. Follow a normal distribution

We want to build a model as shown below:

which will tell us about the affect of experience and education on wages. But we think that the variable educ may be correlated with the error term, i.e. it is endogenous. So we consider the other variables available and believe that a mother’s level of education and a father’s level of education may affect a woman’s level of education. We’ll setup mothereduc and fathereduc as instruments and conduct a Two-Stage Least Squares Regression.

We start setting up the first stage using the structural equation:

We combine the output from the first stage in the term for educ in the second stage and find that the coefficient of educ is smaller and the corresponding standard error is larger as expected. However, the Hausman test tells us that we now have consistent OLS estimators. Hurray! The next step is to repeat the TSLS Regression using both mothereduc and fathereduc.

1. For the first TSLS Regression model (using only mothereduc), which of the following variable or variables are statistically significant in the model of the first stage in our worked example?
   1. exper
   2. sq\_exper
   3. mothereduc
   4. No variables are statistically significant
2. This means that the variable mothereduc is very likely correlated with educ. True/False
3. Now, the increase in wage due to education is \_\_\_\_\_\_. Enter your answer as a percent by taking the appropriate output and multiplying by 100. Round your answer to two decimal places.
4. The standard error for the educ term in this model is \_\_\_\_\_\_. Enter your answer rounded to two decimal places.
5. When we include both mothereduc and fathereduc in a TSLS Regression model they are both statistically significant. True/False
6. With both mothereduc and fathereduc included as instruments, the increase in wage due to education is now \_\_\_\_\_\_. Enter your answer as a percent rounded to two decimal places.
7. The corresponding standard error for the educ term is now \_\_\_\_\_\_. Be sure to round your answer to two decimal places.
8. This means that by adding fathereduc as an instrument, the increase in wage rate due to education has increased and the related standard error has decreased (when you consider standard error to three or more decimal places). True/False