**Week 5 Lab Handout- Multicollinearity & Transforming Variables**

**PA 5032 – Applied Regression**

February 19th, 2021

Wenchen Wang

**PART A: Multicollinearity ~10min**

**PART B: Interpreting Dummy Variables~20min**

**PART C: Quadratic Regressions~10min**

**PART D: Logarithmic Regressions~20min**

**PART A: Multicollinearity**

**Data:** WAGE1.dta (On Class Canvas Site)

**Contents:** 526 observations

**Variables:** wage = average hourly earnings

educ = years of education

exper = years of experience

tenure = years with current company

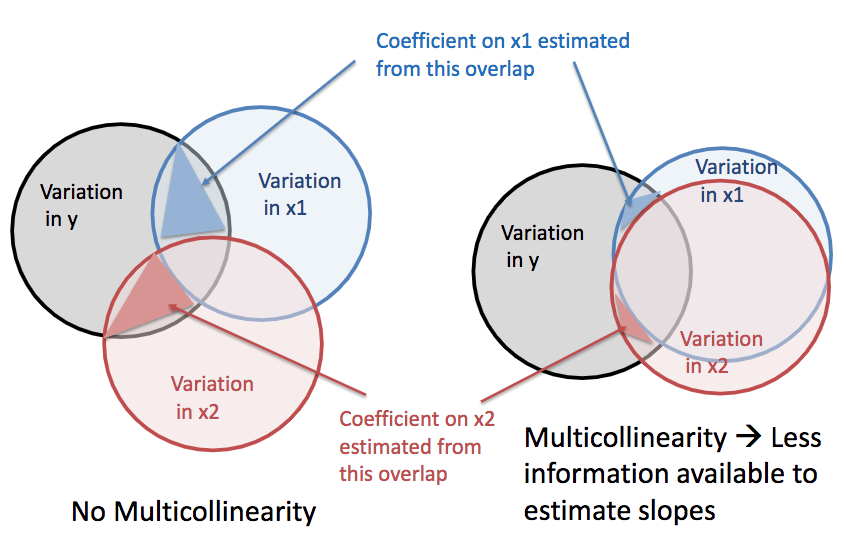
nonwhite = dummy variable where 1 = nonwhite

female = dummy variable where 1 = female

numdep = number of dependents

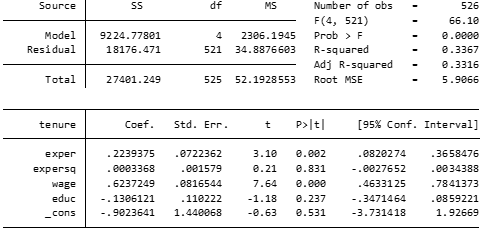
Also includes a number of dummies for region of residence and occupation

Multicollinearity can come in two forms. Perfect collinearity occurs when one of your variables is a function of another one of your variables, such as when you include income in dollars and income in thousands of dollars in your regression. In that case, STATA would drop one of your variables. Imperfect multicollinearity occurs when one or more of your independent variables are highly correlated. This results in there not being enough information to estimate your coefficients.

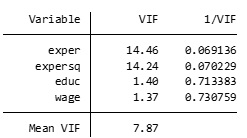


Let's see if there’s any multicollinearity present when we regress tenure on experience, controlling for experience^2, wage, and educ.

*reg tenure exper expersq wage educ*

**

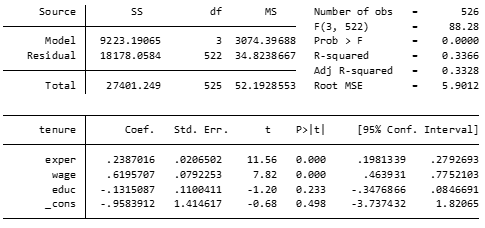
It looks like the standard errors for expersq and educ are pretty high. Let’s do a VIF test to see if multicollinearity may be present in our model.



How can we interpret the outcome of our VIF test? What can we do to alter our model?

*Answer:*

*reg tenure exper wage educ*



**PART B: Interpreting Dummy Variables**

**Data:** eitc.dta (On Class Canvas Site)

**Contents:** 13,746 observations (1991-1996)

**Variables:** state= State of residence

year =Year

urate = State unemployment rate

children = Number of children

nonwhite = Dummy variable where 1= Hispanic/Black

finc = Annual Family Income (97$)

earn= Annual Earnings (97$)

age= Age of Women

ed= Years of Education

work= Dummy =1 if Employed last year

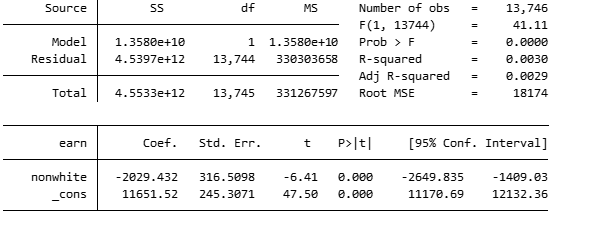
unearn = Unearned Income (97$)

1. **Dummy Independent Variables**

If we are interested in testing the impact of being Hispanic or Black on income, we might consider creating a dummy race variable. I have already done this for you in the variable *nonwhite* (0=white/non-hispanic, 1=black/hispanic).

Now let’s regress income on our race dummy variable and interpret the coefficient on nonwhite.

*reg earn nonwhite*



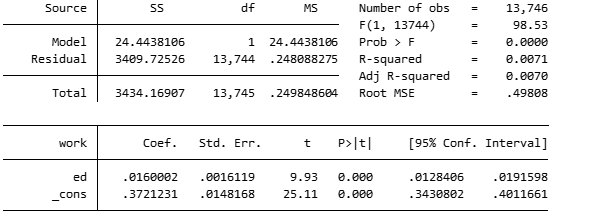
1. **Dummy Dependent Variables**

If our research question is based upon something happening, or not happening, we should generate a dummy dependent variable. Let’s assume we are interested in how an individual’s educational attainment impacts their likelihood of being employed.

For this example, we will use the dummy variable, work, where 1=the individual was employed in the previous year, and 0=not employed.

Let’s regress employment (*work*) on the continuous educational attainment (*ed*) variable and interpret the coefficient on *ed.*

*regress work ed*

**

1. **Categorical Variables converted to Dummy Variables**

Often continuous educational attainment variables are converted into categorical dummy variables. So instead of having each year of completed education coded as a value, different cutoffs are used to create Yes/No dummy variables.

Why might this be done in theory?

Let’s recode our education variable into 4 dummy variables (NoSchool, HSD, MiddleSchoolGrad, MiddleSchool+)

*egen Educ = cut(ed) at(0,1,6,8,20)*

*recode Educ (6=2) (8=3)*

*label define Educlabel 0 "NoSchool" 1 "LessThanMiddleSchool" 2 "MiddleSchoolGrad" 3 "MiddleSchool+"*

*label values Educ Educlabel*

To create the dummy variables, you can either generate 4 new variables and recode them accordingly, or you can use an *index* command to have Stata do this work for you if the variable is already in clearly defined categories.

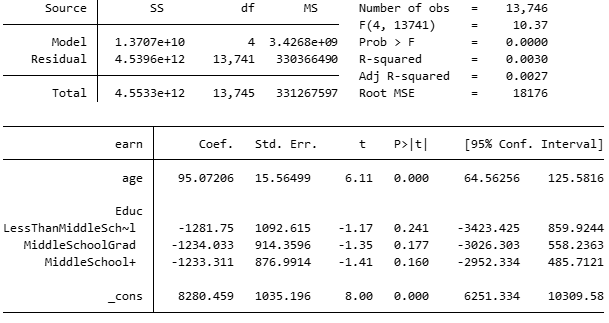
*tab Educ, gen(dummy)*

OR

just include *i.Educ* in the regression model

**Note:** You ***must*** omit one of your categorical dummy variables to avoid perfect collinearity. If you do not select one to be your reference group,then Stata will automatically omit one of them.

*reg earn age i.Educ*



Interpret the coefficient on LessThanMiddleSchool and MiddleSchool+.

Interpret the constant term.

**PART C: Quadratic Regressions**

Sometimes the relationship between the variables in your model will be nonlinear. One way to check the nature of the relationship between your X and Y is to make aspecialized scatterplot to see the graphical representation of your model.

*lowess earn age*

The true relationship between income and age does not appear to be linear. Let’s create a squared term and see what changes in our regression model.

earn = 𝛽0 + 𝛽1age + 𝛽2age^2

*gen age2= age^2*

*reg earn age*

*reg earn age age2*

|  |  |  |
| --- | --- | --- |
| VARIABLES | Earnings | Earnings |
| Age | 95.89\*\*\* | -1,230\*\*\* |
|  | (15.26) | (121.5) |
| Age2 |  | 18.24\*\*\* |
|  |  | (1.659) |
| Constant | 7,056\*\*\* | 29,232\*\*\* |
|  | (559.3) | (2,092) |
| Observations | 13,746 | 13,746 |
| R-squared | 0.003 | 0.012 |

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

What happens to the coefficient on age? Is the second model better at explaining the variation in income?

**PART D: Logarithmic Regressions**

**Reference Table for Interpreting Logged Variables**

|  |  |  |
| --- | --- | --- |
|  | **Non-Logged Dependent** | **Logged Dependent** |
| **Non-Logged Independent** | A \_\_\_ unit increase in X is associated with a \_\_\_\_ unit increase in Y. | A \_\_\_ unit increase in X is associated with a \_\_\_% increase in Y. |
| **Logged Independent** | A 100% increase in X is associated with a \_\_\_ unit increase in Y. | A 100% increase in X is associated with a \_\_\_% increase in Y. |

1. **Logged Dependent Variables**

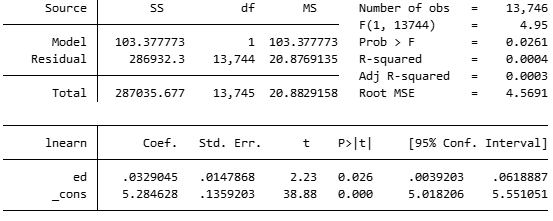
While variables like age are often squared, other variables tend to have logarithmic trend patterns. There are three main types of log models, log dependent, log independent, and log-log models. In a log dependent model, only the Y variable is logged. Let’s continue to explore the impact of education on earnings, but first we need to generate a logged income variable. Then we can run our model.

**\*NOTE: ln 0 = −∞ and ln 1 = 0 So, convert all values of x<1 to 0 . If you don’t, Stata will set them to missing.**

*gen lnearn=ln(earn)*

*replace lnearn=0 if earn<1*

*reg lnearn ed*

**

The trickiest part of any logged model is interpreting the output. In a logged-dependent model, each 1 unit change in Y is now a % change. So, if your X goes up by 1 unit, then Y goes up/down by \_\_%.

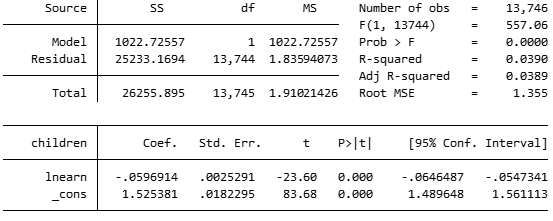
%∆𝑦=100\*𝛽1 ∗ ∆X

Practice interpreting the coefficient on ed.

1. **Logged Independent Variables**

The second log model is when you log an independent variable. Let’s use our lnearn variable to try and understand the impact of the mother’s logged earnings on the total number of children they have in their family.

*regress children lnearn*



In logged-Independent models, each 1 unit change in X is now a 100% change in X. So, if your X goes up by 100%, then Y increases/decreases by B1 units of Y.

∆𝑦 = 𝛽1 ∗ %∆X

Practice interpreting the coefficient on *lnearn*.

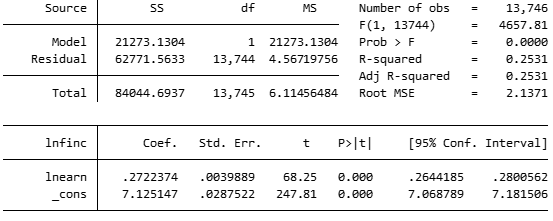
1. **Log-Log Models**

And finally, we have the beloved log-log model. Although this dataset would not likely call for a log-log, let’s go ahead and run a regression to see how a female’s logged earnings impacts the log of total family income.

*gen lnfinc=ln(finc)*

*replace lnfinc=0 if finc<1*

*reg lnfinc lnearn*



To interpret a log-log mode, BOTH X and Y are thought of in %. Thus, if your X goes up by 1%, then your Y increases/decreases by \_\_\_\_ %.

%∆𝑦=𝛽1∗ %∆X

Interpret the coefficient on lnearn. Is it significant?