Further Issues and Topics in Multiple Variable Regression
Econometric Methods, ECON 370
Department of Economics
Saint Francis Xavier University

January 7, 2008

1 Qualitative Explanatory Variables

It is common and very useful to examine how qualitative variable might explain the variation in the the dependent variable, such as the earnings of an individual dependent on whether the individual has had a bachelors or masters degree, or the gender and race of the individuals. The question then is how we can incorporate these qualitative information. We can do so by creating/generating *dummy variables* (*qualitative* or *categorical* variables). We call these dummy variables because the are essentially indicator variables that take on a value of 1 if the observation belongs in that category, and 0 otherwise. This dummy variables or categories must be mutually exclusive, and exhaustive, i.e. it must be possible to assign each observation a single value.

1.1 Definition and Interpretation of Dummy Variables

Dummy variables allow the intercept of the regression line to vary for different groups in the population. To see this consider a simple regression where we wish to examine the differential in years of education by gender;

$$\text{Educ}_i = \beta_0 + \beta_1 \text{Gender}_i + \epsilon_i$$

where

$$\text{Gender}=\begin{cases} 1 & \text{if observation is a woman} \\ 0 & \text{otherwise} \end{cases}$$

This regression examines how the years of education might differ between men and women. Next note that;

$$E\left(\text{Educ}|\text{Gender}=1\right) = \beta_0 + \beta_1$$

While

$$E\left(\text{Educ}|\text{Gender=0}\right) = \beta_0$$

This then says that the difference between male and female educational attainment is nothing but

$$E\left(\text{Educ}|\text{Gender=1}\right) - E\left(\text{Educ}_i|\text{Gender=0}\right) = \beta_1$$

so that if β_1 is statistically significant, then there are differential in attainment by gender, but if statistically insignificant, then there is no difference (in this simple model, we have ignored other intervening concerns which we will elaborate on later in the course, but note the following; did we consider the socioeconomic groups of the observations, the educational attainment of the parents etc). It is an easy task to include other continuous regressors since they will cancel out in the solution for β_1. To see this suppose we included family income as a covariate such that

$$\text{Educ}_i = \beta_0 + \beta_1\text{Gender}_i + \beta_2\text{FamIncome}_i + \epsilon_i$$

Then

$$\begin{aligned} E\left(\text{Educ}|\text{Gender=1}\right) &= \beta_0 + \beta_1 + \beta_2\text{FamIncome} \\ E\left(\text{Educ}|\text{Gender=0}\right) &= \beta_0 + \beta_2\text{FamIncome} \\ \Rightarrow \beta_1 &= E\left(\text{Educ}|\text{Gender=1}\right) - E\left(\text{Educ}|\text{Gender=0}\right) \end{aligned}$$

noting that FamIncome is not an observation's value, but some constant, though the key point is that it will cancel out.

1.2 Interaction Variables

The problem with the above approach is that it is not sufficiently general enough so that the slope coefficients are not allowed to vary by gender. This can easily be solved by multiplying the dummy variables to the continuous variables, a practice commonly referred to as interacting variables. Using the same example thus far, this can be represented as

$$\text{Educ}_i = \beta_0 + \beta_1\text{Gender}_i + \beta_2\text{FamIncome}_i + \beta_3\left(\text{FamIcome}_i \times \text{Gender}_i\right) + \epsilon_i$$

Then in addition to the difference in intercept, the slope coefficients representing the effect of family are now also different by gender. To see this, first note that the intercept and slope coefficient for males are

$$\begin{aligned} E\left(\text{Educ}|\text{Gender=0}\right) &= \beta_0 + \beta_2\text{FamIncome} \\ \frac{\partial E\left(\text{Educ}|\text{Gender=0}\right)}{\partial\text{FamIncome}} &= \beta_2 \end{aligned}$$

While that for females is

$$\begin{aligned} E\left(\text{Educ}|\text{Gender=1}\right) &= \beta_0 + \beta_1 + \left(\beta_2 + \beta_3\right)\text{FamIncome} \\ \frac{\partial E\left(\text{Educ}|\text{Gender=1}\right)}{\partial\text{FamIncome}} &= \beta_2 + \beta_3 \end{aligned}$$

which means that if β_3 is statistically significant, the effect of parental income would have a different effect on women as compared to men.

1.3 Experimental versus Observational Studies

Besides the common qualitative variables noted above, we can similarly use dummy variables in studies using experimental information. It is an oft mentioned critique that data from survey and other aggregated sources typically do not control for intervening effects of outside influences, such as in the study of the effectiveness of a educational policy, we do not control for changes in parental income, and perhaps even changes in an observations socioeconomic status. We can bypass these critique by conducting experiments that control fully for outside influences, within the confines of the experiment venue say. To correctly assess the outcome of an experiment, we would need to compare subjects of the experiment to a group that is otherwise similar in nature, but is not subject to the experiment. We typically call the former the *treatment group*, and the latter the *control group*, since the former was subject to the experiment, while the latter is a subject to some placebo treatment. This then mean that to study the effects of the experiment, we could split the sample up using a dummy variable. All other technical aspect of the regression equation would nonetheless remain the same as before.

1.4 When there are more than Two Groups

It is quite straight forward to include more than two categories. Consider the example above, but extended to parental education. Suppose we are believe the parent with the highest level of education exerts the greatest influence, then we could have the following categories for the variable;

$$\text{Max (PrtEduc)} = \begin{cases} \text{Graduate School} \\ \text{College} \\ \text{High School} \\ \text{Less than High School} \end{cases}$$

Suppose the largest group is that of High School graduate parents, then we can create three dummy variables;

$$\text{Grad} = \begin{cases} 1 & \text{if graduate school} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Coll} = \begin{cases} 1 & \text{if college} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{LessHigh} = \begin{cases} 1 & \text{if less than High School} \\ 0 & \text{otherwise} \end{cases}$$

A simple regression encompassing this information (for simplicity, we now ignore the the gender and family income variables) would be

$$\text{Educ}_i = \beta_0 + \beta_1 \text{Grad}_i + \beta_2 \text{Coll}_i + \beta_3 \text{LessHigh}_i + \epsilon_i$$

Then

$$\begin{aligned} E\left(\text{Educ}|\text{Grad=1}\right) &= \beta_0 + \beta_1 \\ E\left(\text{Educ}|\text{Coll=1}\right) &= \beta_0 + \beta_2 \\ E\left(\text{Educ}|\text{LessHigh=1}\right) &= \beta_0 + \beta_3 \\ E\left(\text{Educ}|\text{Grad=0,Coll=0,LessHigh=0}\right) &= \beta_0 \end{aligned}$$

Then the effect of each type of parent can be found by subtracting the expected effect by $E\left(\text{Educ}|\text{Grad=0,Coll=0,LessHigh=0}\right)$. A similar procedure applies for interaction effects. There are some important notes to keep in mind;

1. Note that in creating the new variables, we have not include one group. This is because like in any simultaneous equation system, if we have 4 dummy variables instead of three, then we would have 5 unknowns, and given that there are only 4 groups, we will have only 4 equations, which means that one of the parameters is not identified. *What happens where we use only 2 categories as opposed to three?*

2. Second, we have excluded the category with the greatest density which we have assumed to be High School attendance. This is because we wish to pin down the variation of all categories with accuracy, i.e. compared to the largest group. The intuition is as follows, in arriving at the estimates for β_1, β_2, and β_3, we are essentially comparing the variation in educational attainment of the observations by the parental education category. The more observations there are in the base group, the more accurate is our prediction of the difference between the base group, and the category in question. If the base group is a small group, we stand to estimate the base group's intercept with poor accuracy, which will affect the veracity of our inference of the other groups.

3. All other issues on inference, and hypothesis testing are similar to prior discussions.

4. Applying interaction effects with continuous variables are as before.

5. Note that it is perfectly possible to interact two dummy variables, for example, the interaction of gender dummy variables with race dummy variables. *How would you interpret the coefficient then? Is it a intercept term, or a slope term?*

2 Effects of Data Scaling on OLS Statistics

In all your assignments and tests and discussion, you should noted that in changing the scale of variables either through log transformations, or other forms of monotonic transformations, that the inferences that you can make in the original form stands. What this hints at is that as long as the transformations

are monotonically increasing in nature (i.e. strictly increasing in nature), the order of the original variable does not vary. To be precise, when variables are rescaled, the coefficients, standard errors, confidence intervals, t statistics, and F statistics, change in ways that preserve all measured effects, and testing outcomes.

2.1 Beta (β_i) Coefficients

It is quite common to have variables used in economic studies that are difficult to interpret. Some examples are IQ scores, or test scores. To eliminate this problem, economists sometimes rescale the variables so that all variables, including explanatory and dependent variables are of the same scale, and the most common of which is to perform standardization (as in what we do to change a random variable into a standard normal random variable) by subtracting the mean, and dividing by the variance of the variable. The interpretation then of the coefficients are than in standard deviations. The benefit of doing so is that it allows us to then examine the importance each variable has in explaining variation in the dependent variable.

Consider a multiple variable regression;

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \widehat{\beta}_2 x_{2,i} + ... + \widehat{\beta}_k x_{k,i} + e_i$$

Subtracting the regression by the mean we get

$$y_i - \overline{y} = \widehat{\beta}_1 \left(x_{1,i} - \overline{x}_1\right) + \widehat{\beta}_2 \left(x_{2,i} - \overline{x}_2\right) + ... + \widehat{\beta}_k \left(x_{k,i} - \overline{x}_k\right) + e_i$$

Lastly dividing all the variables by their standard deviations yield

$$\begin{aligned} \frac{(y_i - \overline{y})}{\widehat{\sigma}_y} &= \widehat{\beta}_1 \frac{\widehat{\sigma}_1}{\widehat{\sigma}_y} \frac{(x_{1,i} - \overline{x}_1)}{\sigma_1} + \widehat{\beta}_2 \frac{\widehat{\sigma}_2}{\widehat{\sigma}_y} \frac{(x_{2,i} - \overline{x}_2)}{\widehat{\sigma}_2} + ... + \widehat{\beta}_k \frac{\widehat{\sigma}_k}{\widehat{\sigma}_y} \frac{(x_{k,i} - \overline{x}_k)}{\widehat{\sigma}_k} + \frac{e_i}{\widehat{\sigma}_y} \\ &\Rightarrow \quad z_y = \widehat{b}_1 z_1 + \widehat{b}_2 z_2 + ... + \widehat{b}_k z_k + error \end{aligned}$$

where

$$\widehat{b}_j = \left(\frac{\widehat{\sigma}_j}{\widehat{\sigma}_y}\right) \widehat{\beta}_j, \forall j = 1, 2, ..., k$$

where $\widehat{b}_j$ is what is typically referred to as the beta coefficient or the standardized coefficients. The interpretation of the coefficient is that if of when x_j increases by one standard deviation, $\widehat{y}$ increases by $\widehat{b}_j$ standard deviations. All variables here are on equal footing, and can easily be compared for importance as noted before. This advantage also extends to situations when the variables has been converted to logs, since even when considering elasticity, we would still have to contend with what is occuring on the average, that is 10% is large for one variable starting from low base but small for another with the average observation at a high level. To create this variable, we could either do so variable by variable, or if our statistical package permits, we could do so directly using a command based operation.

3 More on Functional Forms

3.1 Logarithmic Functional Forms

Consider the log-log regression which we know to be of the following form,

$$\log(y_i) = \beta_0 + \beta_1 \log(x_{1,i}) + \beta_2 x_{2,i} + \epsilon_i$$

We know that in this setup β_1 can be interpreted as percentage change in y for a 1% change in x_1. Whereas for a one unit change in x_2, y changes by β_2%. However, lets examine how we came up with the last statement,

$$\begin{aligned} \frac{\partial y}{\partial x_2}\frac{1}{y} &= \beta_2 \\ \Rightarrow \frac{\Delta \log y}{\Delta y}\frac{\Delta y}{\Delta x_2} &= \beta_2 \\ \Rightarrow \frac{\Delta \log y}{\Delta y}\frac{\Delta y}{\Delta x_2} &= \beta_2 \end{aligned}$$

But note that the interpretation of the above as a percentage change is only an approximation, that is

$$\%\Delta y \approx 100.\Delta \log(y)$$

and this approximation worsens as $\log(y)$ increases. However, this can be corrected by first noting the following (holding x_1 constant),

$$\begin{aligned} \Delta\widehat{\log(y)} &= \widehat{\beta}_2 \Delta x_2 \\ \Rightarrow \widehat{\log y_2} - \widehat{\log y_1} - &= \widehat{\beta}_2 \Delta x_2 \\ \Rightarrow \log\left(\frac{\widehat{y_2}}{\widehat{y_1}}\right) = \log\left(\frac{(\widehat{y_1}+h)-\widehat{y_1}}{\widehat{y_1}}+1\right) &= \widehat{\beta}_2 \Delta x_2 \\ \Rightarrow \frac{(\widehat{y_1}+h)-\widehat{y_1}}{\widehat{y_1}} &= \exp\left(\widehat{\beta}_2 \Delta x_2\right) - 1 \\ \Rightarrow \%\widehat{\Delta y} &= 100.\left[\exp\left(\widehat{\beta}_2 \Delta x_2\right) - 1\right] \end{aligned}$$

This is only a small difficulty when we consider the benefits of the transformation or running this type of function form in our regressions,

1. When $y > 0$, models using $\log(y)$ as the dependent variable often satisfy assumption more closely than models using the level of y. Positive variables typically have conditional distributions that are heteroskedastic, and taking logs can often solve the problem.

2. Taking logs narrows the range of variables, which means that the regressions are less sensitive to the values of the outlying observations.

3. Easily interpreted, since we can be ignorant of the units of measure of the variables since the slope coefficients are invariant to rescaling.

There are however several rules of thumb to abide by;

1. When the underlying variables takes on strictly positive values, and can be easily thought of as continuous, it would mean that it would be advisable to use logs. Examples being income, wages, population numbers etc.

2. Variables that are measured in units such as years should be transformed with the log transformation.

3. It cannot be used when the values take on 0 values. To transform this sorts of variables, we typically add 1 to all the realizations. But that would depend on how much of a change it makes, and be cognizant of the interpretation.

4. Regressions where the dependent variables are different cannot be compared, that is you cannot compare the fit using the R^2 when for one regression you used y, and the other $\log(y)$.

3.2 Models with Quadratics

Quadratic functions are also typically used when we wish to model increasing or decreasing marginal effects. So for example, when we wish to examine whether parental income has an increasing or decreasing effect on the attainment of children, we may add an additional variable that is the squared variable of the parental income variables. Consider a general regression model where we wish to model the marginal effect of x_2, we can then perform the following regression;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + e$$

However, running the regression changes the partial effect of x_2;

$$\frac{\partial y}{\partial x_2} = \beta_2 + 2\beta_3 x_2$$

or more precisely

$$\frac{\Delta y}{\Delta x_2} \approx \beta_2 + 2\beta_3 x_2$$

Further note that the maxima/minima is achieved when

$$\begin{aligned} \beta_2 + 2\beta_3 x_2 &= 0 \\ \Rightarrow x_2 &= \left| \frac{\beta_2}{2\beta_3} \right| \end{aligned}$$

When the dependent variable is $\log(y)$, and the explanatory variable takes on a quadratic form, we have to be careful in interpretation. *See example 6.2 of your text.* What happens when both the first and second order derivatives are of the same sign?

We can also combine the quadratic functional together with the logarithmic. That is

$$\log y = \beta_0 + \beta_1 \log(x) + \beta_2 [\log(x)]^2 + e$$

This then means that the partial effects are

$$\begin{aligned} \frac{\partial y}{\partial x}\frac{1}{y} &= \beta_1 \frac{1}{x} + 2\beta_2 \frac{\log(x)}{x} = (\beta_1 + 2\beta_2 \log(x))\frac{1}{x} \\ \Rightarrow \quad \Delta y &\approx (\beta_1 + 2\beta_2 \log(x))\,\%\Delta x \end{aligned}$$

The significance of this is that the elasticity of y with respect to x, is dependent on $\log(x)$, and consequently we have a nonconstant elasticity model. It is in your interest to understand how under different variations of the Classic Linear Model, the coefficients are interpreted.

Finally. note that it is not uncommon to see polynomial terms, the real functional form of which is dependent on the a priori expectations of the researcher. This means that it is possible for you to have a cubic, or quartic term.

3.3 Models with Interactions

We have dealt with interaction of qualitative variable with continuous variables, and qualitative variables. It is also possible to interact continuous variables with each other. This essentially creates cross partial relationships which researchers sometimes might a priori expect and wish to verify the degree and significance of. Consider the following model;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2) + e$$

Note then that

$$\begin{aligned} \frac{\Delta y}{\Delta x_1} &= \beta_1 + \beta_3 x_2 \\ \frac{\Delta y}{\Delta x_2} &= \beta_2 + \beta_3 x_1 \end{aligned}$$

However, cross partial effects aside, econometrics is about giving meaning to observational values. Consider the partial effects of x_1 alone, which based on the above is when $x_2 = 0$, which might not always be reasonable or easily translated (See page 204-205 of your text.). Now consider a new modified model,

$$y = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 ((x_1 - \mu_1) \times (x_2 - \mu_2)) + e$$

Note that we have not changed β_3 since μ_i are nothing but constants. Now note that

$$\begin{gathered} \frac{\Delta y}{\Delta x_1} = \delta_1 + \beta_3 (x_2 - \mu_2) \\ \Rightarrow \beta_1 + \beta_3 x_2 = \delta_1 + \beta_3 (x_2 - \mu_2) \\ \delta_1 = \beta_1 + \beta_3 \mu_2 \end{gathered}$$

which then means that the effect of x_1 on y is nothing but the partial effect of x_1 on y at the mean value of x_1. Further note that the standard errors of δ_1 is the standard error for deviation from the mean without further calculations.

4 Adjusted R^2 and the Selection of Regressors

Recall that we have stressed that we should not put too much weight on the value of R^2 in judging our models, nor in comparing models. Just because R^2 is small does not mean that the error term is correlated with the independent variables/covariates. R^2 is nothing but a measure of fit. What it does mean is that a small R^2 implies that the error variance is large relative to the variance of y, which means that our estimate of the coefficients of interest may not be accurate. Yet again, as long as we have a large sample, this problem may be circumvented. In addition, note that from your exercises that the relative change in the R^2 when variables are added to a regression equation is very useful due to the F test we had used, testing between restricted and unrestricted models.

4.1 Adjusted R^2, $\overline{R}^2$

You would have noticed after running your regressions that there are two types of goodness of fit measure, R^2 the other called the adjusted R^2. If you have paid attention to some of the applied papers you have read in other economics courses, you might have noticed that it is common to report adjusted R^2 or $\overline{R}^2$. So what is the difference between the two. The formula for the former is

$$R^2 = 1 - \frac{\left(\frac{SSR}{n}\right)}{\left(\frac{SST}{n}\right)} = 1 - \frac{SSR}{SST}$$

Intuitively, if our model perfectly explains the data (i.e. is actually the data generating process (DGP)), the sum of residuals would be zero, and R^2 would be 1, since there would be no residuals to talk about. Further, R^2 is just an estimator for

$$\rho^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}$$

Which means that we are estimating σ_ϵ^2 with $\frac{SSR}{n}$, and recall that in your exercises, you have proved that the estimator is a biased estimator. To obtain the unbiased estimator, we know that all we need to do is to divide SSR by $n-(k+1)$ where k is the number of variables we have used to explain variation in y. Further, we also found that the unbiased estimator for the SST is to divide SST by $n-1$ instead of n. This suggests that a better measure of goodness of fit is actually

$$\overline{R}^2 = 1 - \frac{\frac{SSR}{n-(k+1)}}{\frac{SST}{n-1}}$$

and this formula is known as the adjusted R^2. However, it should be noted that $\overline{R}^2$ is not generally known to be a better estimator. The rationale is as follows; just because the estimators used to estimate σ^2 and σ_y^2 are unbiased does not mean that our $\overline{R}^2$ is an unbiased estimator of ρ^2, the population goodness of fit.

What it does well is that it cautions us against adding explanatory variables without consideration of the true model. To see this note that SSR never falls as we add more explanatory variables which might lull us into adding any variable we could get our hands on that we might suspect could explain the variation in the dependent variable. $\overline{R}^2$ instead penalizes us for adding too many explanatory variables each time we add a new variable since the direction of $\Delta \frac{SSR}{n-(k+1)}$ as more variables are added is unknown. In fact $\overline{R}^2$ will increase if and only if the t statistic of the coefficient to the additional variable is greater than 1 in absolute value (since the denominator increase by 1). This means that the conclusion that $\overline{R}^2$ gives us with regards to an adding an additional explanatory variable is different from that derived from the t and F test since under the usual α we would have to drop a variable if t is at all close to 1 (recall that the rule of thumb for the t statstistic is 2). Lastly note that we can derive a formula for $\overline{R}^2$ in terms of R^2

$$\overline{R}^2 = 1 - \frac{SSR}{SST}\frac{n-1}{n-(k+1)} = 1 - \left\{\left(1-R^2\right)\left(\frac{n-1}{n-(k+1)}\right)\right\}$$

4.2 Using $\overline{R}^2$ to Choose between Non-Nested Models

We say that two models are non-nested when neither is a special case of another, and example of which is the following

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon \end{aligned}$$

The F test we have used till now does not allow us to choose between the two since neither is a restricted model of the other. Of course we can run a regression including all variables we suspect to have explanatory power, and test various permutations against this complete model using the F test. However, we may get a scenario where both models are not rejected, or both models are rejected (see your text for an example using the data on MLB). By comparing $\overline{R}^2$ for two non-nested models, we could then decide which one is a better model, but of course if they are extremely close in quantum, this type of comparison may not be advisable since the relationship may change once we include or exclude another variable.

However, this manner of comparison can be useful when we are choosing between two functional forms,

$$
\begin{aligned}
y &= \beta_0 + \beta_1 \log(x_1) + \epsilon \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon
\end{aligned}
$$

since it is obvious that neither model is a subset of the other, but rather each is a different way of modelling concavity in relationship between the dependent and independent variable.

A final note of caution, you cannot compare non-nested models when the difference between them is due to the functional form of the dependent variable. For example, we know that when we take $\log s$, we effectively reduce variation of a variable, this would mean that the explanatory variables would be able to account for greater amounts of variation and consequently have a larger R^2 and $\overline{R}^2$.

4.3 Controlling for too many factors in Regression Analysis

A principle fallacy in practise when we perform multiple variable regression is to control for too many variable. Ultimately, what model we use is dependent on the question on hand. Each time we add a variable, by the ceteris paribus effect, we are in effect cutting off a possible direction of effect between a independent variable in question in relation to the dependent variable (Read your book for good examples). We very often fall into the trap of over controlling due to feared biases through exclusion of important variables. Nonetheless, the judgement required is not clear cut, and there lies the art in our social science.

4.4 Adding Regressors to Reduce the Error Variance

It is important to realize when choosing a particular model, which model we actually choose ultimately depends on the question on hand, i.e. the context. However, where there is no ambiguity is that we must always include a independent variable when that variable is totally uncorrelated with the other independent variables for the reason that its inclusion does not affect how we interpret the other covariates (since there is no problem of multicollinearity), and more importantly, it reduces the error variance, i.e. SSR or σ_ϵ^2. If you recall from your formulas for you standard error of the estimators, in large sample, the inclusion of such a variable will reduce the standard error of all OLS estimators. The caveat (as there always will be in social sciences) is that such variables are rare, but if it exists, they must be included.

You should read the section of your text from page 214 to 221 on Prediction and Residual Analysis.