PPOL 501-02 & -06, Fall 2014
Course Notes #18: Introduction to Multivariate Regression

I. INTRODUCTION

- Up until now, we've removed ourselves from reality a little bit to imagine that only *one X* variable might explain the variation in *Y* (e.g., we are better able to predict body fat percentage if we know a person's weight; or we are better able to predict wages if we know a person's IQ score; but we haven't gone beyond those variables).

- Even though we know that other factors may affect Y we said that we were not worried about these things (with regard to biasing the coefficient estimate on the independent variable) *IF* SLR1 through SLR4 were met.

- In particular, we were worried about SLR4: that the omitted factors were *not correlated* with the *X* in the model (i.e., the assumption of $E(u|X)=0$)

- How good is this assumption?

- Let's go back to the wages model. In the example we worked with in course notes 17, we said that wages were a function of IQ scores: Wages = *f* (IQ), and we estimated the following regression line: WAGESHAT = 278.023 + 3.285(IQ). For each additional point increase in IQ score, we predicted that monthly wages would increase by 3.29 cents, on average.

- A bunch of other factors might influence wages (we listed a number of factors on the board in class). One of these factors may be education. When we say this, we're saying that we think that *education* (an independent variable) and *wages* (the dependent variable) are related in some way.

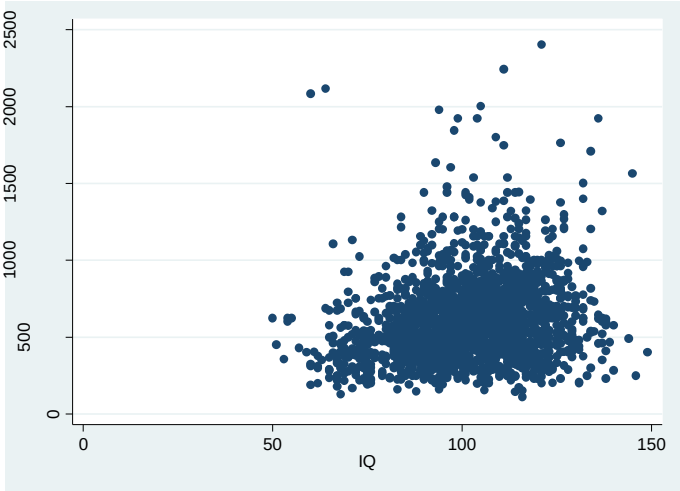
  [And now for a (seemingly) unrelated question: Do you think that there is an association between *IQ* and *education* (two possible independent variables)?

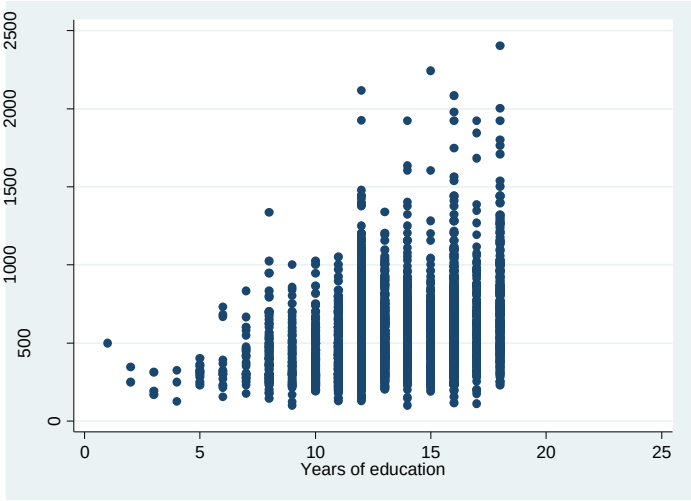
- Let's use STATA to plot the relationships between (1) Wage and IQ, (2) Wage and Education, and (3) IQ and Education:

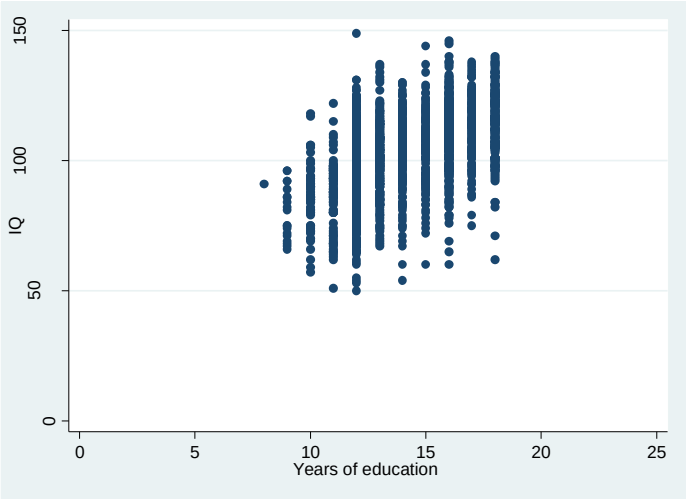
STATA CODE FOR GRAPHS BELOW:

```
use card

graph twoway scatter wage iq, ylabel(0(500)2500) ytitle("Wage in cents per
hour") xlabel(0(50)150) xtitle("IQ")

graph twoway scatter wage educ, ylabel(0(500)2500) ytitle("Wage in cents per
hour") xlabel(0(5)25) xtitle("Years of education")
```

```
graph twoway scatter iq educ, ylabel(0(50)150) ytitle("IQ") xlabel(0(5)25)
xtitle("Years of education ")
```






IQ
0
50
100
150
0
5
10
15
20
25
Years of education

```
. summ wage IQ educ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |      3010    577.2824    262.9583        100       2404
          IQ |      2061    102.4498    15.42376         50        149
        educ |      3010    13.26346    2.676913          1         18


. pwcorr wage IQ educ, obs sig

             |     wage       IQ     educ
-------------+---------------------------
        wage |   1.0000
             |
             |     3010
             |
          IQ |   0.1920   1.0000
             |   0.0000
             |     2061     2061
             |
        educ |   0.3019   0.5103   1.0000
             |   0.0000   0.0000
             |     3010     2061     3010
```

- So how can we account for both IQ and Education in our estimated model?
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ WAGE $= \beta_0 + \beta_1 IQ + \beta_2 EDUC + u$

- To include both IQ and Education in the OLS regression model as explanatory variables, just add EDUC to the regress statement in Stata:

```
. regress wage IQ educ

      Source |       SS       df       MS              Number of obs =    2061
-------------+------------------------------           F(  2,  2058) =   66.24
       Model |  8675400.98     2  4337700.49           Prob > F      =  0.0000
    Residual |   134772670  2058  65487.2062           R-squared     =  0.0605
-------------+------------------------------           Adj R-squared =  0.0596
       Total |   143448071  2060  69634.9861           Root MSE      =   255.9

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          IQ |   1.725291   .4250631     4.06   0.000     .8916926     2.55889
        educ |   20.73191   2.882899     7.19   0.000      15.0782    26.38561
       _cons |   149.1893   41.89648     3.56   0.000     67.02539    231.3532
------------------------------------------------------------------------------
```

- How do you interpret the intercept, $\hat{\beta}_0$?

- For the slope coefficient: First, test for stat sig of $\hat{\beta}_1$. p<0.001 so it is highly stat sig at conventional levels.
  → In Quant 2, you will delve more deeply into how the standard error is calculated for a multiple regression coefficient:

[$t = \dfrac{\hat{\beta}_j - \beta_j}{se_{\hat{\beta}_j}}$ where $s.e._{\hat{\beta}_j}^2 = Var(\hat{\beta}_j) = \dfrac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$

and where $\hat{\sigma}^2 = \dfrac{\sum \hat{u}_i^2}{n-k-1}$ $SST_j = \sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$

R_j^2 = the R-squared from the regression of X_j—where *j* indicates the *X* variable of interest—on **all the *other Xs*** in the original model of interest]

- In this class for now, we'll simply check for stat sig and proceed with interpretation as appropriate.

- The coefficient estimate for IQ in a multivariate regression is interpreted slightly differently than before:

  "*Holding education constant*, each additional IQ point increase is associated with a 1.73 cent per hour increase in wages, on average."

  Thus, the coefficient on IQ gives the *partial effect* of IQ on wages, holding education constant.

  You might say this as, "*ceteris paribus*, each additional IQ point increase…."

  where *ceteris paribus* means "other (relevant) factors being equal"

  -- here, we are explicitly controlling for education in the equation; and anything else that might affect wages we are assuming is random (i.e., not systematically related to the variables in the model)

- Now interpret the coefficient on EDUC….

- What does this mean? Wooldridge (p. 77) provides a really helpful way to think about this. We'll use his words to explain this regression that we just calculated:

  "The power of multiple regression analysis is that it provides this ceteris paribus interpretation even though the data have *not* been collected in a ceteris paribus fashion. In giving the coefficient on [IQ] a partial effect interpretation, it may seem that we actually

went out and sampled people with the same [EDUCATION] but possibly with different [IQ] scores. This is not the case. The data are a random sample [of the U.S. population]: there were no restrictions placed on the sample values of [IQ] or [EDUCATION] in obtaining the data. Rarely do we have the luxury of holding certain variables fixed in obtaining our sample. *If* we could collect a sample of individuals with the same [EDUCATION], then we could perform a simple regression analysis relating [WAGES] to [IQ]. Multiple regression effectively allows us to mimic this situation without restricting the values of any independent variables."

"The power of multiple regression analysis is that it allows us to do in nonexperimental environment what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed."

- The key is that these other factors must be *either* (1) explicitly measured and included in the regression, *or* (2) assumed (or known) to be uncorrelated to either the dependent variable, or the independent variables.

II. MECHANICS AND ASSUMPTIONS OF MULTIVARIATE OLS

- Just as in the simple case, the coefficients in multivariate OLS estimation are selected to *minimize the sum of squared residuals,* i.e.

$$\min \sum_i e_i^2 = \min \sum_i (Y_i - \hat{Y}_i)^2 = \min \sum_i [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_2 + ...\hat{\beta}_k X_k)]^2$$

- Just as we went through the assumptions of a simple regression model, we can state assumptions of the multiple linear regression model (remember, these are *assumptions!* You need to think about whether or not they hold):

MLR Assumption 1: (*linear in the parameters*): The dependent variable, *Y* is a function of a set of independent variables, ***X*.** The coefficient β that corresponds to each of the *X*s is a constant (unknown). The unmeasurable, remaining error (*u*) is assumed to be random. Think of this as the population regression model. We'll also call it the "true" model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k + u$$

MLR Assumption 2: A random sample of *n* observations is drawn from the population of interest.

MLR Assumption 3: The number of observations in the sample is greater than the number of independent variables **X** in the regression model; none of the independent

variables is constant; and there is no exact linear relationship among the independent variables (i.e., no *perfect* collinearity among the *X* variables)

MLR Assumption 4: The expected value of the error term, *u*, is zero conditional on the Xs in the model:

$$E(u|X_1, X_2, X_3, \ldots X_k) = 0$$

As before, we know this assumption is violated when elements of the error term, *u*, are correlated with the *X*s that are included in the model.

Another way of saying this is: If factors *are* associated with *Y* but they *are not* specifically measured in the model (in other words, they are in the *u* term), AND if these factors *are* associated with the Xs that *are* in the model, THEN the assumption is violated.

This is the idea behind understanding **omitted variable bias** in multiple regression coefficients, which we'll talk about in Quant2. It is a key concept in multiple regression.

MLR Assumption 5: The error terms all have the same variance; that is, the error variance is not a function of the *X*s in the model (this is the assumption of *homoskedasticity*):

$$Var(u| X_1, X_2, X_3, \ldots X_k) = \sigma^2$$

$\Rightarrow$ Under Assumptions MLR1 through MLR4, $E(\hat{\beta}_j) = \beta_j$ for $j = 0, 1, 2, \ldots k$.

In other words, when these assumptions hold, $\hat{\beta}_j$ is an *unbiased estimator* of β_j, where β_j is the "true" population parameter.

$\Rightarrow$ Under Assumptions MLR1 through MLR5, no other linear unbiased estimator of the β coefficients can have smaller sampling variances than those of the least-squares estimator, i.e., they are BLUE (**B**est **L**inear **U**nbiased **E**stimators). This is known as the ***Gauss-Markov Theorem*** (and thus MLR1 through MLR5 are often referred to as the Gauss-Markov assumptions) (see Kennedy, pp. 42-46 for additional perspective)

Thus, the assumptions and the G-M theorem provide a baseline – i.e., "use OLS" – that we know we can depend on unless something goes tragically wrong. That is, OLS is BLUE unless the assumptions are violated, in which case we need to worry about what to do.

Many of the topics we cover our Quant sequence are concerned with how to estimate models when these assumptions are violated.