Claire Leavitt
PO 841: Quantitative Methods
Final Exam Review
10 December 2014

I. OLS ASSUMPTIONS

1. Suppose you have a data set that, when you plot your independent variable against your dependent variable, suggests a linear relationship between those variables. What are four assumptions necessary to perform a linear regression?

2. What is OLS regression? Without access to any kind of software, how would you go about performing this type of linear regression?

3. Describe the Gauss-Markov theorem and the assumptions necessary for OLS to constitute the Best Linear Unbiased Estimator (BLUE). How do we define which model is "best"?

4. Say you've collected data on the entire BU student population, and you're trying to figure out what accounts for the variability in students' GPAs. You decide to regress three independent variables: X_1 = each students' incoming/high school GPA; X_2 = a dummy variable that represents whether or not a student is in a fraternity/sorority; and X_3 = each student's response to how much they care about their schoolwork (on a scale of 1-10). You realize too late that you forgot to collect data on how many hours per week each student spent studying, which you think will significantly affect GPA, but you perform a multivariate OLS regression anyway.

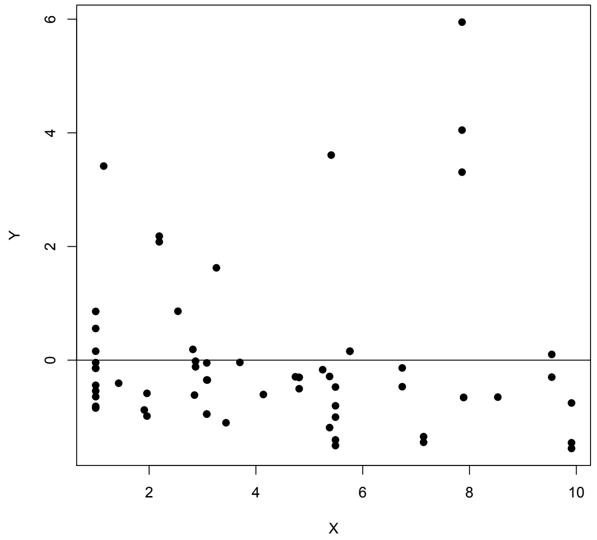
- What are the consequences of omitting the hours-studied variable?

- Let's say you discover that your X_4 variable—how many hours per week each student spends studying—is highly correlated with X_3, the variable that measures how much each student cares about his/her schoolwork. What are the consequences of including X_3 and X_4 as separate predictor variables? What steps could you, as a researcher, take to remedy this problem?

4. When you show your supervisor a residuals plot of your data, he says he think the data exhibit heteroskedasticity. How can he tell, what does this mean, and is there anything you can do to eliminate the problem?

6. Look at the following residuals plot of error terms against an independent X variable. What assumptions appear to be violated? Why?

Claire Leavitt
PO 841: Quantitative Methods
Final Exam Review
10 December 2014



II. OLS REGRESSION

1. The following output shows a bivariate OLS regression of heart rate against body mass (weight); researchers want to detect whether or not there is a significant relationship between weight and heart rate.

```
Call:
lm(formula = bodymass$Rate ~ bodymass$Mass)

Residuals:
    Min      1Q  Median      3Q     Max 
-155.74  -89.16  -16.56   21.36  358.84 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    113.165    179.587   0.630    0.537    
bodymass$Mass   26.879      3.786   7.099 1.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.1 on 17 degrees of freedom
Multiple R-squared:  0.7478,  Adjusted R-squared:  0.7329 
F-statistic:  50.4 on 1 and 17 DF,  p-value: 1.784e-06
```

a. What is the equation of the OLS regression line?

b. What does the intercept value ($\hat{\beta}_0$) of 113.165 mean, substantively?

c. What does the slope coefficient ($\hat{\beta}_1$) of 26.879 mean, substantively?

d. What is the 95% confidence interval for $\hat{\beta}_1$ (construct the interval and then interpret it)?

e. What does the p-value associated with the F-statistic mean?

f. What does the value for multiple r-squared mean?

2. Let's assume you want to test whether a country's degree of urbanity and cultural openness (as measured by Norris and Inglehart's Cosmopolitanism index) has an effect on the percentage of people who consider themselves highly religious. You run a multivariate regression of religiosity against several independent variables (including cosmopolitanism, development as measured by the Human Development Index, national trade flows, level of international tourism and whether or not the majority of a country's population is Muslim) and receive the following output in R:

```
Call:
lm(formula = democracy$fhrelfr ~ democracy$CBINDEX + 
democracy$HDI2005 + 
    democracy$Flows + democracy$Tourism + democracy$Muslim)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.30478 -0.58668  0.05881  0.37937  2.03439 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)        1.209652   3.743582   0.323   0.7494  
democracy$CBINDEX -0.452061   0.183351  -2.466   0.0212 *
democracy$HDI2005  2.373178   4.833418   0.491   0.6279  
democracy$Flows    0.002705   0.013231   0.204   0.8397  
democracy$Tourism  0.123137   0.299881   0.411   0.6850  
democracy$Muslim   1.778347   0.679511   2.617   0.0151 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8511 on 24 degrees of freedom
  (161 observations deleted due to missingness)
Multiple R-squared:  0.6995,	Adjusted R-squared:  0.6369 
F-statistic: 11.18 on 5 and 24 DF,  p-value: 1.206e-05
```

a. Substantively interpret the slope coefficients in this model, keeping in mind that all variables are continuous except “Muslim,” which is a dummy variable where 1 represents a Muslim-majority nation and 0 is a non-Muslim-majority nation.

3. Now assume you want to see whether or not the relationships between cosmopolitanism and religiosity and between development and religiosity are different for Muslim countries versus non-Muslim countries. You add two interaction terms to your model and receive the following output:

```
Call:
lm(formula = democracy$fhrelfr ~ democracy$CBINDEX + democracy$HDI2005 +
    democracy$Flows + democracy$Tourism + democracy$Muslim +
    democracy$CBINDEX * democracy$Muslim + democracy$HDI2005 *
    democracy$Muslim)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4237 -0.3876  0.0000  0.2755  2.0155

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -0.686675   4.706115  -0.146    0.885
democracy$CBINDEX                   -0.512979   0.206410  -2.485    0.021 *
democracy$HDI2005                    4.548222   5.864561   0.776    0.446
democracy$Flows                      0.004984   0.013947   0.357    0.724
democracy$Tourism                    0.089780   0.311774   0.288    0.776
democracy$Muslim                    11.618070  12.818165   0.906    0.375
democracy$CBINDEX:democracy$Muslim   0.736262   1.228475   0.599    0.555
democracy$HDI2005:democracy$Muslim -12.713460  16.367198  -0.777    0.446
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8755 on 22 degrees of freedom
  (161 observations deleted due to missingness)
Multiple R-squared:  0.7086,  Adjusted R-squared:  0.6159
F-statistic: 7.643 on 7 and 22 DF,  p-value: 0.0001031
```

a. Interpret the coefficients of both interactions and the significance of the interaction terms. What does this tell us about how the aforementioned relationship may differ between Muslim and non-Muslim countries?

III. MATRIX ALGEBRA

You’re interested in figuring out whether the number of hours parents read to their children affect how many years of higher education children pursue. You collect the following data for 6 respondents, where Y = years of education (the range is 1-23, including kindergarten and people who can’t finish their PhD dissertations); and X = the number of hours per week parents read to each subject as a child. You estimate your OLS regression model as y = Xβ+ ε.

Years of Education of Respondent	**Number of Hours/Week Parents Read to Respondent**
12	4
20	14
9	1
16	7
12	0
15	5

a. Using matrix algebra, find $\hat{\boldsymbol{\beta}}$.

b. Find $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{\epsilon}}$.

c. Find $\widehat{\boldsymbol{var}}(\hat{\boldsymbol{\beta}})$.