USP 634 Spring 2017
Assignment 4: Association; Linear Regression
Due **Friday, June 9th, 2017**

The objectives of the assignment are to

- Calculate and interpret measures of association and
- Estimate and interpret linear regression model

1. In question 3 from Assignment 3 you tested a hypothesis involving a bivariate relationship where both variables were categorical. Calculate two measures of association for your variables, interpret their meaning, and show how the values are derived by presenting the correct computational formula (15 points).

2. Using the data set **world00.RData**, test the hypothesis of international aid agencies that urban areas are the engines to national economic development. Test this hypothesis by examining the relationship between urban (% of population living in cities) and gdp_cap (Gross domestic product per capita). Calculate the correlation between the two variables, present a scatter plot with linear regression line, and write out the best fitting bivariate linear equation which expresses gdp_cap as a function of urban population %, and present statistics on goodness-of-fit and test significance. (20 points)

3. Consider the following regression output from R. The data consists of individuals drawn from the Panel Study of Income Dynamics (PSID).

```
> summary(lm(income92~age+educ, data=psid92, subset=subset))

Residuals:
   Min     1Q Median     3Q    Max
-24744  -9826   1324   1971 476389

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1162.21     601.42  -1.932  0.05340 .
age           -53.93      19.52  -2.763  0.00576 **
educ         1590.49      66.53  23.905  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17550 on 2858 degrees of freedom
  (66 observations deleted due to missingness)
Multiple R-squared: 0.2118,     Adjusted R-squared: 0.2113
F-statistic: 384.1 on 2 and 2858 DF,  p-value: < 2.2e-16
```

The dependent variable is total labor income ($) in 1992. The independent variables are age and years of schooling (educ).

- In these data, what is the predicted increase in earnings from a one-year increase in years of schooling (holding age constant)? (5 points)
- What is the estimated standard error of the coefficient on years of schooling? What is the likelihood of observing an estimated effect of education of this magnitude if education is not truly associated with earnings? (5 points)
- What proportion of the variation in earnings can be explained by variation in education and age? (5 points)
- Your neighbor is 20 years old and has 10 years of formal schooling. Suppose you wished to use the regression results above to predict how much someone like your neighbor would have earned in 1992. How would you do it? You do not need to perform the calculation, just make it clear which numbers would get added to, subtracted from, multiplied by or divided by which other numbers (5 points).

4. Present a multiple regression analysis (with at least two independent, or explanatory, variables) that addresses some question of significance with your own dataset.

1. In no more than two sentences (use a diagram if appropriate), postulate and present a research question and corresponding hypothesis (hypotheses).
2. Next, present and interpret the correlation matrix.
3. Briefly discuss how variables were entered into your model.
4. Check for homoskedasticity and normality of the residuals with scatterplots.
5. Present and interpret the *key* multiple regression results.
6. Finally, in a brief paragraph, what do the results tell you about your study question?

In responding to these questions, be concise! Also, note that your grade on this question will depend far less on the size of your R^2 than the quality and conciseness of your presentation and interpretations. (35 points)

A good answer will include the following:

- A theory as to which dependent and independent variables may have a causal relationship
- Alternative specifications of models, rather than just one model
- Scatterplots showing relationships between variables
- Transformations of variables to maintain linearity, as appropriate
- Explanations of results that are counterintuitive
- A discussion of coefficients (and significance of results) as well as a discussion of R-squared

5. For the above, re-run your final model using beta weights. This means converting your main independent variables to Z-scores and re-running the regression. Which of your variables has the greatest influence, according to this procedure? Show your work or explain in detail what you did. (10 points)