GX5004: HW 3 [Total 20 points]

R and Python have extensive libraries online that can guide you on this assignment. Please feel free to work collectively on the challenging questions. [10 points]

1. Download the Griliches (1976) data, called "griliches.dta", from the course website.[1] These data are formatted as a Stata dataset.
    a. Read this dataset into R or Python.
    b. Generate summary statistics for the following variables in the data:
        - RNS is an indicator for residency in the southern states
        - MRT is an indicator for marital status
        - SMSA is an indicator of residency in urban areas
        - MED is mother's education in years
        - IQ is IQ score
        - KWW is "Knowledge of the World of Work" test score
        - AGE is age of the individual
        - S is completed years of schooling
        - EXPR is work experience in years
        - LW is log wage (cents per hour)
    c. Generate scatter plots of log wages against:
        - RNS
        - MRT
        - SMSA
        - KWW
        - EXPR
    d. Estimate bivariate least squares models that relate log wages to the variables in c. Do your results make intuitive sense to you?
    e. Estimate a bivariate least squares model relating log wages to schooling. Calculate a 95 percent confidence interval using your results.
    f. Estimate a multivariate least squares model relating log wages to the variables in b.[2] Calculate a 95 percent confidence interval for the estimate of the returns to schooling using your results.
    g. Generate a variable that is age raised to the power of two (i.e., is age squared). Now re-estimate f. including age-squared.[3]
    h. Challenging question: Discuss reasons why your estimates of the returns to schooling in e. and in f. differ from each other.

---

[1] This is a subset of the original data used in Griliches (1976) and was obtained from: http://fmwww.bc.edu/ec-p/data/hayashi/hayashi2000.html. If interested, the original Griliches paper can obtained from: http://dept.ku.edu/~empirics/Courses/Econ818/griliches_jpe76.pdf.

[2] It goes without saying: Don't include log wages as a right-hand-side variable.

[3] Recall that the Griliches data are a sample of younger men, so the curvature discussed in class may not yet be observable in the data.

i. **Submit code, results, and discussion.**

2. Assume the following DGP: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for $\epsilon_i \sim N(0,1)$. For ease, set all of the betas equal to one (i.e., $\beta_0 = \beta_1 = \beta_2 = 1$). [5 points]
   a. Suppose $x_{1i} \sim N(0,1)$ and independently $x_{2i} \sim N(0,1)$. Further suppose you estimate using least squares the following model: $y_i = b_0 + b_1 x_{1i} + u_i$. What number do you think your least squares estimate of b_1 should be close to? Using R or Python, simulate this DGP assuming 10,000 observations and estimate the least squares value for b_1.
   b. Challenging question: Suppose instead $x_{1i} = z_i + v_i$ and $x_{2i} = -z_i + \omega_i$ where $z_i \sim N(0,1), v_i \sim N(0,1),$ and $\omega_i \sim N(0,1)$ are independent. Again, you estimate using least squares the following model: $y_i = b_0 + b_1 x_{1i} + u_i$. What can you say about your least squares estimate of b_1? Using R or Python, simulate this DGP assuming 10,000 observations and estimate the least squares value for b_1.[4]
   c. Challenging question: Suppose I say that any statistical estimates you put in front of me I dismiss as saying you haven't included everything in the world that's relevant. How do you respond?
   d. Submit all code, results, and discussion.
3. This problem addresses the machine learning concept of classification. Download the National Longitudinal Survey of Women (NLSW) data presented in class, called "union.dta", from the course website. These data are formatted as a Stata dataset. [5 points]
   a. Read this dataset into R or Python.
   b. Treat the years 70-78 of the NLSW data as a training set. Using R or Python and the training set, estimate the model presented in class both as a linear and a logit classifier using all of the attributes provided in the dataset.[5]
   c. Treat the years 80-88 of the NLSW data as a set of attributes on individuals that you would like to classify as union/non-union. Using a threshold of 0.25, classify these individuals as union/non-union based on their attributes for the years 80-88 using both the linear and the logit classifiers estimated in b.
   d. For both models, summarize the accuracy of your support vector machine (with a threshold of 0.2) in a table by comparing your union prediction to what was actually observed. It might look something like the table below.

SVM	Number of Union Members (Predicted)	Number of Union Members (Actual)
Linear		
Logit		

   e. Submit all code and results.

[4] Under these conditions, we violate A3 in class.

[5] Union is the outcome we want to classify. Year is a trend variable. SMSA is an indicator variable that takes on value 1 when someone lives in an urban location, such as NYC. Grade is highest education achieved. South is an indicator variable that takes on value 1 when someone lives in the southern USA. Black is an indicator variable that takes on value 1 when someone is African American.