POL 212
Winter 2024
Assignment 8

1. Analyze the *Salary* dataset (salary.dta) using a linear model estimated with ordinary least squares. The outcome of interest is the variable **salary**, which is the annual salary of small university professors measured in dollars. The input variables are:

    - **rank** faculty position (full professor, associate professor, assistant professor, or instructor/lecturer)
    - **yrrank** number of years in current rank
    - **termdeg** highest degree earned coded 1 for PhD and 0 for MA
    - **yrdg** number of years since highest degree was earned
    - **female** coded 1 for female

2. Complete the following tasks:

    a. Specify a multiple regression model that uses all five input variables. Conduct a test of the hypothesis that the gender of a professor has no effect on that professor's salary, holding fixed the other variables you used in your model.

    b. Using that model, predict the salary of a female associate professor who has been at her rank for 2 years, and has a Ph.D. that she earned 9 years prior.

    c. Test the hypothesis that the effect of rank is conditional on the gender of the professor (i.e., an interaction effect) and interpret this relationship.

    d. Assess the degree to which multicollinearity, heteroscedasticity, and endogeneity are problematic in your regression model.

    e. Are there are any influential cases/outliers that could potentially be biasing the results? Be sure to explain how you diagnose these potential problems (whether graphically or with a hypothesis test) and report the results of your diagnoses.

CHALLENGE QUESTION: ANSWER C1, C2, or C3
(all relate to the multiple regression model you specified above)

C1.) Estimate and report bootstrapped standard errors for the regression coefficients. How do these compare to the traditional (analytically derived) standard errors?

C2.) Use a Monte Carlo simulation approach to generate a new input variable with the following properties:

(1) It has a moderately strong *linear* relationship with **yrrank**.
(2) It has a moderately strong *nonlinear* relationship with the outcome **salary**
(3) Is itself normally distributed.

Experiment with how inclusion, omission, and transformations of this simulated variable in the regression model alongside the original input variables affects all of the coefficient estimates.

C3.) Use a Monte Carlo simulation approach to generate the following:

(1) An observation that is a large outlier but has low leverage.
(2) An observation that has high leverage but is not an outlier.
(3) An observation that is both a large outlier and has high leverage (i.e., an influential point). Add these simulated observations to your dataset and repeat the steps in part 1(e) to diagnose each.