Multivariate Statistical Techniques
CLPS2908, Spring 2019
Prof. Bertram Malle

Homework 3

For this homework you will analyze a fictitious data set in which pre-thirties' annual income (in dollars) is predicted from education (in years since first grade), job experience (in months), and being a member of an ethnic minority group (0 = no, 1 = yes).

1. First perform EDA to ensure error-free data and describe range and distribution of all variables.

2. Use both numerical correlations and bivariate plots to discuss the relationships among the predictors and their respective relationships with the outcome variable. Even though you cannot with confidence state what the *multivariate* combination of predictors is going to look like, do look for noteworthy relationships and point out things to monitor in the actual regression analysis.

3. Choose appropriate keywords for the subcommands of the REGRESSION syntax to yield useful information about predictor intercorrelations, zero-order and semi-partial correlations, distribution of residuals, partial plots, and anything else you might be interested in. Consult the lecture handouts and the SPSS syntax guide.

```
REGRESSION
  /DESCRIPTIVES
  /STATISTICS
  /DEPENDENT
  /METHOD=ENTER
  /SCATTERPLOT=   .
```

4. Examine and describe (verbally and with the appropriate statistical parameters) the *overall* predictive power of the set of independent variables.

5. Examine the residuals you got from the regression analysis.
   a. Explain what residuals are.
   b. What does the distribution of residuals in this data set tell you?

6. If we had to pick *one* independent variable to best predict the outcome, which one would it be?

7. Provide a verbal description of the size and direction of *each* independent variable's predictive power. In so doing, compare zero-order correlations with semi-partial ("part") correlations.
   a. Describe in your own words what the two types of correlations (zero-order and semi-partial) mean in general.

b. Then discuss each type of correlation for every one of the predictors. Do you see cases of shared predictive variance? Of suppressor effects? Or any other unusual patterns?

c. One way of examining the various relationships is to compare the overall R^2 and the sum of the semi-partial r^2 values for each predictor. If there is a difference, what does that mean?

d. Another way of examining the relationships is to remove one variable at a time (resulting in 3 models with 2 predictors each), and examine what happens to the now-remaining variables' predictive power. Apply this method to the data. What do you learn about the relationship between the predictors and the outcome?

8. If you wanted to define the most parsimonious model of the data, would you remove any of the predictors? If so, which one(s) and why?

9. Write out the complete regression equation (without removing any of the original predictors) and make a prediction for a new case with the following values:

Education = 9 years
Job experience = 7 years
Ethnicity = Native American.

Think carefully about which regression parameters you want to use and how you represent the specific values in the equation.

10. In your one-page summary, describe how the predictors relate to each other and to the outcome. Refer to the relevant statistical results but also explain them using language that describes the findings in real-world terms.