# *SOC 4015/5050: PS-06 - Multivariate Regression*

*Christopher Prener, Ph.D.*

*Fall 2018*

## *Directions*

Please complete all steps below. All work should be uploaded to your GitHub assignment repository by 4:15pm on Monday, December 3rd, 2018. All data can be obtained from the `testDriveR` package's `gss14` data set.

## *Analysis Development*

Using RStudio and your operating system's file manager, create an R Project in the *existing* directory in your assignments repository named `PS-06`. Add a `README.md` file, notebook, and all necessary folders before beginning.[1]

[1] This initial section follows the project workflow that is available in the `lecture-03` repo!

## *Part 1: Data Preparation*

1. Using the data table `gss14` in the `testDriveR` package, create a new data frame that has *only* the following data:

```
> gssClean
# A tibble: 2,538 x 8
      id hrsWork white black otherRace female fullTime incomeCat
   <int>   <int> <lgl> <lgl> <lgl>     <lgl>  <lgl>        <dbl>
 1     1      60 TRUE  FALSE FALSE     FALSE  TRUE            21
 2     2      40 TRUE  FALSE FALSE     TRUE   TRUE            25
 3     3      NA TRUE  FALSE FALSE     FALSE  FALSE           18
 4     4      20 TRUE  FALSE FALSE     TRUE   FALSE           25
 5     5      NA TRUE  FALSE FALSE     TRUE   FALSE           NA
 6     6      60 TRUE  FALSE FALSE     TRUE   TRUE            25
 7     7      NA TRUE  FALSE FALSE     FALSE  NA              NA
 8     8      40 TRUE  FALSE FALSE     FALSE  TRUE            21
 9     9      NA TRUE  FALSE FALSE     TRUE   FALSE           11
10    10      55 FALSE FALSE TRUE      TRUE   TRUE            22
# ... with 2,528 more rows
```

   Store these cleaned data in your `data/` sub-directory as a `.csv` file.

## *Part 2: Descriptive Statistics and Assumptions*

Using the GSS data created above in Part 1, answer the following questions.

2. Report the *appropriate* descriptive statistics for *all* of the variables displayed in the output included with Part 1. Also create a formatted descriptive statistics table to include with your assignment submission. Store the output in your `results/` sub-directory.[2]
3. Conduct a full set of normality tests on the variables `hrsWork` and `incomeCat` and report your findings.[3]
4. Create a correlation table to identify any possible issues with regression assumptions.
5. Summarize your assessment of how these data meet the assumptions of linear regression.

[2] This output should be left as a `.html` file - it does not need to be reformatted into Microsoft Word.

[3] For the purposes of this assignment, we are going to treat `incomeCat` as a continuous variable.

## *Part 3: Model*

Using the GSS data created above in Part 1, answer the following questions.

6. Construct a hypothesis and null hypothesis for the relationship between number of hours worked (`hoursWork`) and income (`incomeCat`), accounting for the other factors included in your data set.
7. Construct a dissemination ready plot of the relationship between hours worked (`hoursWork`) and income (`incomeCat`).
8. Construct a regression equation modeling how income, accounting for race, gender, and whether or not someone works full time, affects `hoursWork` using LaTeX syntax.
9. Execute a main effects model (model 1) of the effect of income on hours worked (`hoursWork`) (`incomeCat`).
10. Execute a full model (model 2) with all of your control variables.
11. Provide a written summary of the findings of both of your models, including interpretations of the betas and appropriate measures of model fit.

### *Part 4: Post-Hoc Assumptions Checks*

Using the GSS data created above in Part 1, answer the following questions.

14. Using the skills covered in Lecture 14, *fully* check the assumptions and model fit of your second model.

15. Provide a written summary of the findings of your assumption checks.

### *Part 5: Final Model*

Using the GSS data created above in Part 1, answer the following questions.

16. Fit another model (model 3) that properly accounts for any issues discovered in Part 4.

17. Provide a written summary of how re-fitting the model has changed its conclusions. Is model 2 or model 3 a better model overall?