SOC 4930/5050: PS-10 - Multivariate Regression

Christopher Prener, Ph.D.

November 27th, 2017

Directions

Please complete all steps below. Your well-formatted R Notebook source (the `.Rmd` file) and `html` output as well as your LaTeX formatted regression tables should be uploaded to your GitHub assignment repository by 4:15pm on Monday, December 11th, 2017.

Part 1: Data Preparation

1. Using the data table `gss16` in the `testDriveR` package, create a new data frame that has *only* the following data:[1]

[1] Recall that, in `gss16`, the `RACE` variable's values are `1 = white`, `2 = black`, and `3 = other`; the `SEX` variable's values are `1 = men` and `2 = women`; and the `WRKSTAT` variable's key value is `1 = full time work`.

```
# A tibble: 2,867 x 8
      id hrsWork white black otherRace female fullTime incomeCat
   <int>   <int> <lgl> <lgl>     <lgl>  <lgl>    <lgl>     <int>
 1     1      50  TRUE FALSE     FALSE  FALSE     TRUE        26
 2     2      42  TRUE FALSE     FALSE  FALSE     TRUE        19
 3     3      NA  TRUE FALSE     FALSE  FALSE    FALSE        21
 4     4      30  TRUE FALSE     FALSE   TRUE    FALSE        26
 5     5       5  TRUE FALSE     FALSE   TRUE    FALSE        26
 6     6      NA  TRUE FALSE     FALSE   TRUE    FALSE        20
 7     7      55  TRUE FALSE     FALSE  FALSE     TRUE        26
 8     8      30 FALSE FALSE      TRUE   TRUE    FALSE        16
 9     9      80 FALSE  TRUE     FALSE  FALSE     TRUE        20
10    10      NA  TRUE FALSE     FALSE  FALSE    FALSE        20
# ... with 2,857 more rows
```

Part 2: Descriptive Statistics and Assumptions

Using the GSS data created above in Part 1, answer the following questions. These questions build on PS-09, so you should be able to borrow code from that assignment!

2. Report the *appropriate* descriptive statistics for *all* of the variables displayed in the output included with Part 1. Also create a LaTeX formatted descriptive statistics table to include with your assignment submission.

3. Conduct a full set of normality tests on the variable `hrsWork` and `incomeCat` and report your findings.[2]
4. Create a correlation table to identify any possible issues with regression assumptions. There is no need to create a full LaTeX table, just summarize your findings.
5. Summarize your assessment of how these data meet the assumptions of linear regression.

[2] 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 Week 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?