SOC 4930/5050: PS-09 - Bivariate Regression

Christopher Prener, Ph.D.

November 13th, 2017

Directions

Please complete all steps below. Your well-formatted R Notebook source (the `.Rmd` file) and `html` output should be uploaded to your GitHub assignment repository by 4:15pm on Monday, November 20th, 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]

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

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 the hours worked variable (`hrsWork`) renamed in Part 1.

3. Conduct a full set of normality tests on the variable `hrsWork` and report your findings.

[1] Recall that, in `gss16`, the `RACE` variable's values are `1 = white`, `2 = black`, and `3 = other`.

4. Report the *appropriate* descriptive statistics for the variable `race` renamed in Part 1.
5. Report the *appropriate* descriptive statistics for the variable `white` created in Part 1.
6. Report the *appropriate* descriptive statistics for the variable `black` created in Part 1.
7. Summarize your assessment of how these data meet the assumptions of linear regression.

Part 3: Bivariate Regression

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

8. Construct a hypothesis and null hypothesis for the relationship between number of hours worked (`hoursWork`) and race (`race`).
9. Construct two dissemination ready plots of the relationship between hours worked (`hoursWork`) and race (`race`). One plot should be geared towards communicating with an audience with a degree of statistical literacy, and the other plot should be designed for individuals with more limited analytic knowledge.[2]
10. Construct a regression equation modeling how race, using the binary variables you created and leaving the "other" category as the reference, affects `hoursWork` using LaTeX syntax.
11. Execute a bivariate regression model that shows how race, again using the binary variables you created and leaving the "other" category as the reference, affects `hoursWork`.[3] Fully interpret the results of this model.

[2] *Hint*: Look back at the plots discussed during the difference of means weeks for inspiration!

[3] Check the website for techniques to include both binary variables in your regression model. This was not covered in class.

Rubric

Individual Questions

Part 1		Part 2		Part 3	
Question	Points	Question	Points	Question	Points
1	6	2	2	8	1
		3	2	9	3
		4 through 6	1	10	2
		7	2	11	3
Points Possible	6		9		9

Note: Partial credit possible

Notebook Formatting & RMarkdown

Category	Details	Points
Excellent	Syntax used appropriately & without error	3
Good	Minor concerns with syntax use	2.55
Improvement Needed	Significant concerns with syntax	2.25
Unsatisfactory	No RMarkdown used	0
Points Possible		3

Literate Programming

Category	Details	Points
Excellent	Narrative throughout with great detail	3
Good	Some narrative with inconsistent detail	2.55
Improvement Needed	Limited narrative with little detail	2.25
Unsatisfactory	No narrative included	0
Points Possible		3