Lab 14 - Bivariate Regression & Interpretation

George Rhodes

November 28, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 14 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select the main focal relationship you're interested in exploring for your poster project.

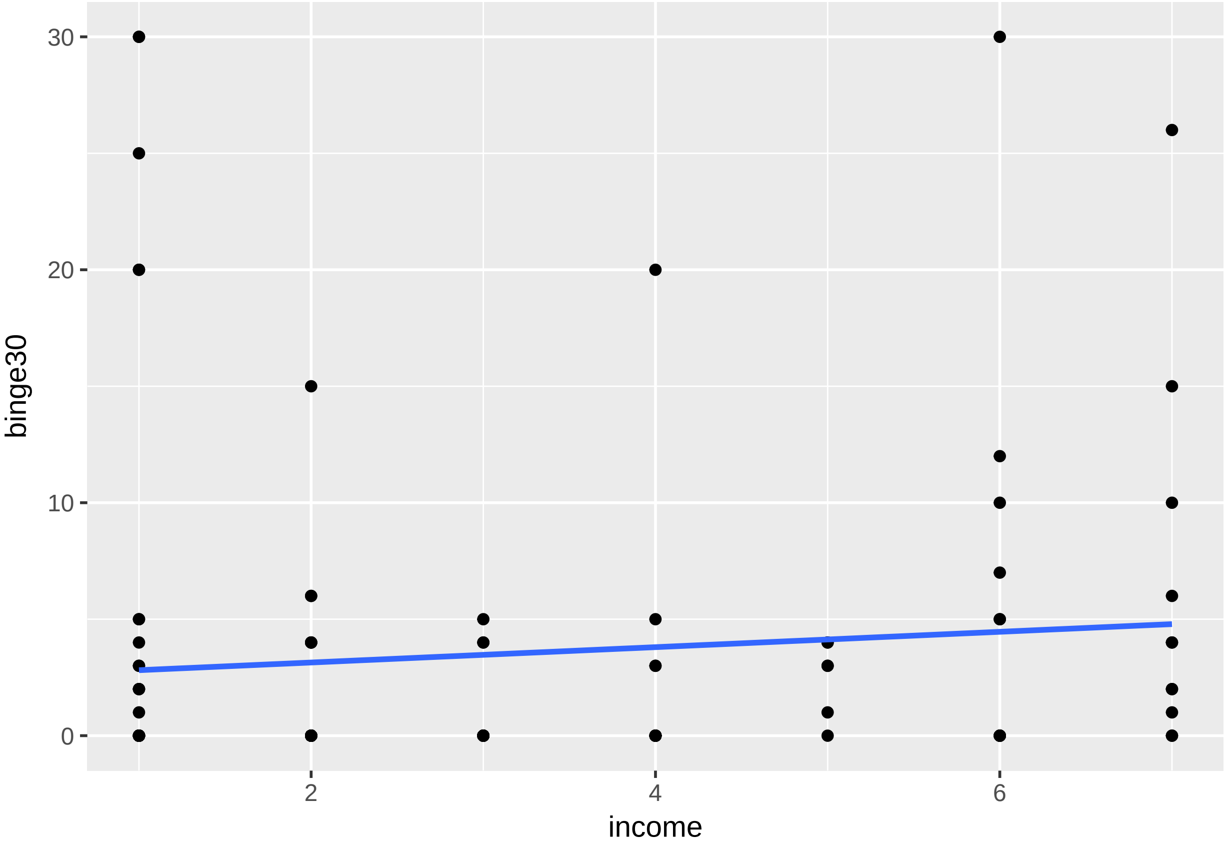
a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

I am mainly interested in how SES impacts treatment outcomes. I can measure SES by Income, Family Income, and Education. I can measure treatment outcomes in a few different ways, but so far I have been focusing on those who have been to treatmetn at some point and haven't not drank for the last year. For this I will operationalize SES as income and treatment outcomes as number of days binge drank (had 4 or more drinks if famale and 5 or more drinks if male) in the past 30 days.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
## Warning in binge30 == 0:30: longer object length is not a multiple of
## shorter object length
```



```
##
## Call:
## lm(formula = binge30 ~ income, data = treated_respondents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7889 -3.4707 -2.8116  0.5322 27.1884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4820     1.1728   2.116   0.0366 *
## income        0.3296     0.3143   1.048   0.2968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.826 on 110 degrees of freedom
## Multiple R-squared:  0.009893,   Adjusted R-squared:  0.0008924
## F-statistic: 1.099 on 1 and 110 DF,  p-value: 0.2968
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.
The bivariate association between income and number of days binge drank in the last 30 is positive, with a coefficient of 0.3296, an intercept of 2.4820, and a p-value of 0.2968.

d. What is the meaning of the model intercept?

It is somewhat meaningless here, as 1 on the x-axis includes incomes from 0-10k. But ithe intercept means that of those who have been to treatment and make $0 annually, they have typically drank more than 4/5 (f/m) drinks 2.4820 times in the past 30 days.

e. How well does the bivariate model fit the data? How is this information calculated?

It's hard to say. It doesn't fit very well. The main problem is that income is catagorical. This information is calculatec by measuring the distance between each point and the linear model, and then determining the error by summing all of those distances.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

I don't think this alone tells me anything. It seems consistent with the theories I've come across recently that suggest that those of higher SES use more alcohol. So far from my models, this holds even post treatment.

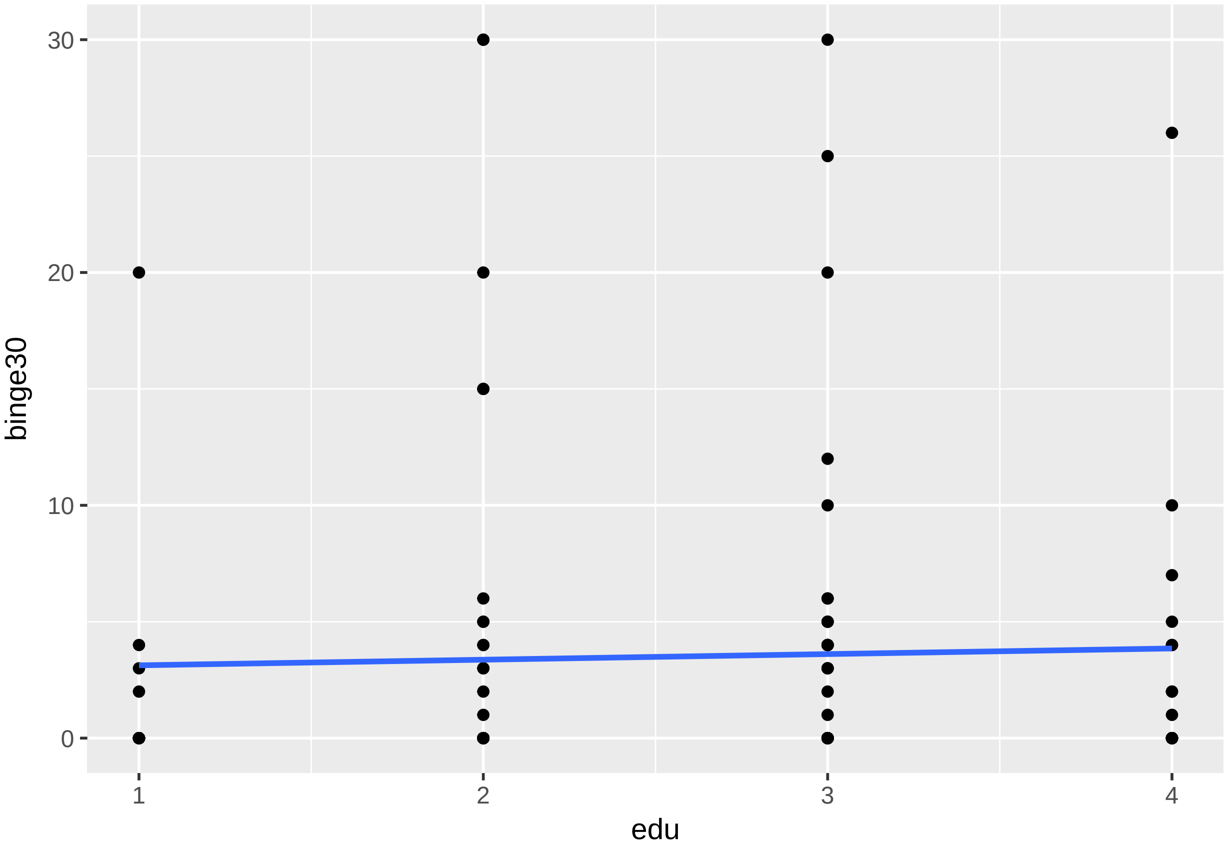
2. Select a different focal relationship related to your project. This could be:

- **A different response and a different explanatory variable**
- **A different response and the same explanatory variable**
- **The same response and a different explanatory variable**

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

I will operationalize SES as education level and treatment outcome as number of days binge drank (4 or more drinks if female and 5 or more drinks if male) in the past 30 days. The

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.



```
##
## Call:
## lm(formula = binge30 ~ edu, data = treated_respondents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8515 -3.6102 -3.1276  0.3898 26.6311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8863     1.8573   1.554    0.123
## edu           0.2413     0.6746   0.358    0.721
##
## Residual standard error: 6.856 on 110 degrees of freedom
## Multiple R-squared:  0.001162,   Adjusted R-squared:  -0.007919
## F-statistic: 0.1279 on 1 and 110 DF,  p-value: 0.7213
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

Again, this is correlation is barely positive with a coefficient of 0.2413 and p-value of 0.7213.

d. What is the meaning of the model intercept?

The intercept doesn’f fit the model as education has been operationalized such that 1 == less than highschool diploma and over 18. There is no concept of 0. If this was operationalized as number of years of schoool completed, than those without any formal schooling would be expected to have binge drank 2.8863 days out of the last 30.

e. How well does the bivariate model fit the data? How is this information calculated?

The large p-value tells me that this model does not fit the data well.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

Again, this is not consistent with my hypothesis, but this simple model does not capture all the aggregate of proceses that determine behavior change.