Lab 15 - Multivariate Regression & Interpretation

George Rhodes

November 30, 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 15 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 a second explanatory variable from your dataset that you think has implications for the theoretical association of your focal relationship.

a. Describe the theoretical reasoning for selecting this variable.

I will use both education and income, as both are measures that may impact number of days binge drank in last 30. I would expect that increased SES would lead to more successful treatment of addiction. Increased income allows consumers more resources and autonomy to pursue their treatment, and more education would be associated with more social capital. HOwever, so far I may be finding that higher SES leads to more use post treatment.

b. What type of relationship do you think this variable has with your focal variables? Given that, what do you expect to happen to your focal relationship when it is added to the model?

I expect the barely positive relationship to be magnified.

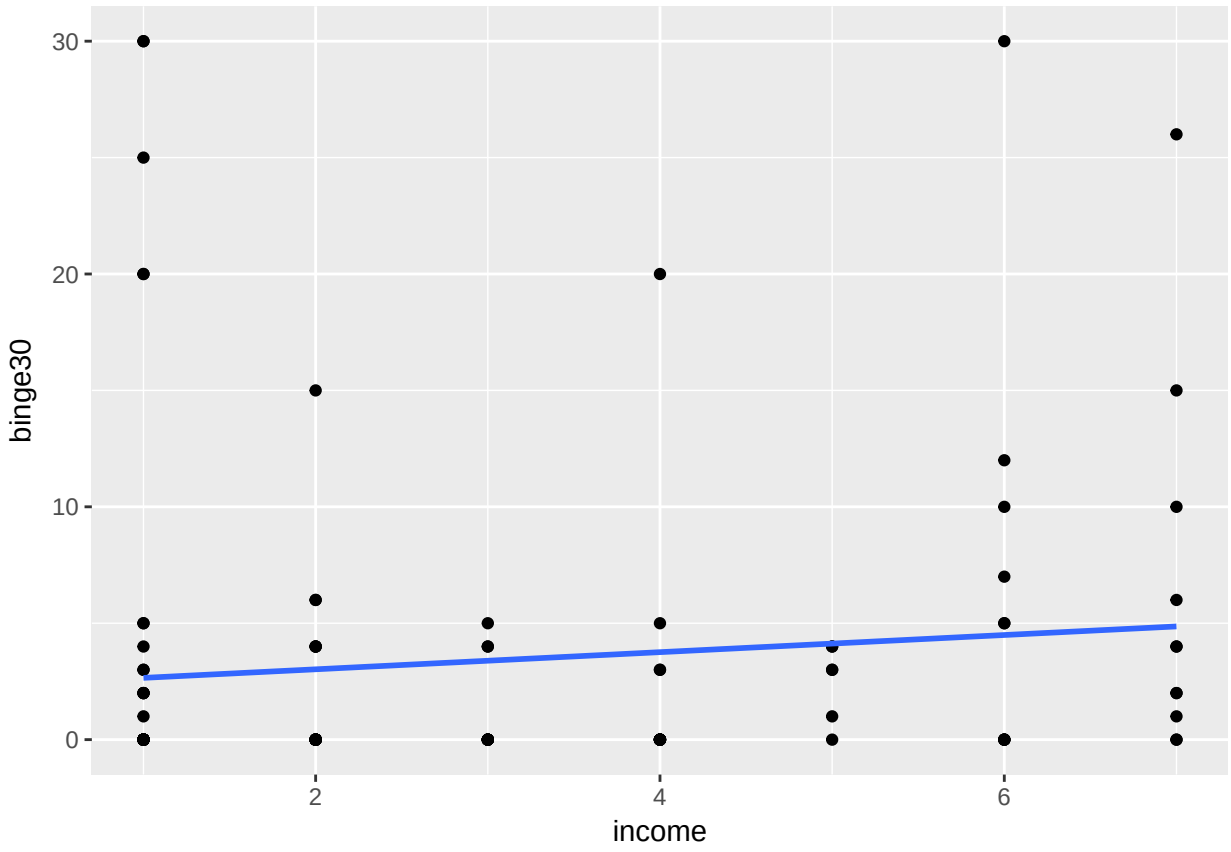
c. Is it a continuous or categorical variable? What implications does this have for a multivariate regression equation?

EDU is another catagorical variable and I'll need to add is as a factor.

d. Conduct a multivariate linear regression with this additional explanatory variable 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 + factor(edu), data = treated_respondents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4298 -3.6959 -1.7984  0.2948 26.3041
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.1532     3.4019   0.045    0.964
## income                        0.3468     0.3384   1.025    0.308
## factor(edu)AA/some college    2.5984     3.6577   0.710    0.479
## factor(edu)college grad       1.2984     3.8480   0.337    0.736
## factor(edu)highschool         3.1959     3.6131   0.885    0.378
## factor(edu)less than HS       0.8795     3.7757   0.233    0.816
##
## Residual standard error: 6.77 on 110 degrees of freedom
## Multiple R-squared:  0.03302,    Adjusted R-squared:  -0.01093
## F-statistic: 0.7513 on 5 and 110 DF,  p-value: 0.5868
```

e. Describe the results of the multivariate analysis, highlighting:

- the apparent association between the control variable and the focal response variable
- how the focal association changed when you incorporatedthe control variable
- the implications of these results for your focal association

The assiation between income and number of days of binge drinking in the past 30 among those who have

attended treatment is not statistically significant with a p-value of 0.5868. Every increase in income (which is a jump to the next income bracket, of 10k) is associated with a 0.3468 increase in days binge drinking when controlling for education level. This is slight increase from when we did not control for edu. This makes sense given that they were both individually positively correlated to # of days binge drinking.

f. How well does this model fit the data? Is it an improvement over the bivariate model? Why or why not?

The p-value is much bigger, so I don't think this is near complete. However, I like the theory that adds these measures.

2. Select any additional variables you want to incorporate into your final model. For each additional variable added to the model answer the following questions:

could add Marital status, age, race, sex, countytype, poverty, and famincome. Will only add first 3 for this model with all the recoding.

a. Describe the theoretical reasoning for selecting this variable.

Race: Racial disparities in health are well recognized and studied, I must control for it. Age: The factors involved with the decision to change behaviors and follow through on treatment change with age. Marital Status: having others to support and that are suppportive of you.

b. What type of relationship do you think this variable has with your focal variables? Given that, what do you expect to happen to your focal relationship when it is added to the model?

I think binge drinking will decrease with age, and marital status, but I have no idea about race. Race is strongly coorellated with SES, which may be reflected in the income and education variables.

c. Is it a continuous or categorical variable? What implications does this have for a multivariate regression equation?

My survey data is all catagorical, so I must use factors.

d. Conduct a multivariate linear regression by adding one explanatory variable at a time and save the model objects. Print out the full results by calling `summary()` on each model object.

```
##
## Call:
## lm(formula = binge30 ~ income + factor(edu) + factor(marital) +
##     factor(race) + factor(age), data = treated_respondents)
##
## Coefficients:
##                (Intercept)                      income
##                     0.8736                      0.3056
## factor(edu)AA/some college    factor(edu)college grad
##                    -3.9243                     -5.8237
##     factor(edu)highschool    factor(edu)less than HS
##                    -4.5474                     -5.1885
##    factor(marital)married       factor(marital)never
##                    -0.1494                      0.7034
##    factor(marital)widowed          factor(race)black
##                    10.5607                      0.7059
##      factor(race)hispanic    factor(race)multiracial
##                    -2.0932                     -2.6073
##     factor(race)native AM     factor(race)Pac Island
##                    -1.1466                     -3.9173
##         factor(race)white           factor(age)18-25
##                    -1.1457                      6.2766
##          factor(age)26-34           factor(age)35-49
```

```
##                     5.8802                     8.5553
##           factor(age)50-64             factor(age)65+
##                     5.7473                         NA
```

```
##
## Call:
## lm(formula = binge30 ~ income + factor(edu) + factor(marital) +
##     factor(race) + factor(age), data = treated_respondents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.129  -3.013  -1.447   1.125  25.779
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.8736     5.6117   0.156  0.87662
## income                        0.3056     0.3896   0.784  0.43467
## factor(edu)AA/some college   -3.9243     5.4459  -0.721  0.47289
## factor(edu)college grad      -5.8237     5.7497  -1.013  0.31365
## factor(edu)highschool        -4.5474     5.4745  -0.831  0.40821
## factor(edu)less than HS      -5.1885     5.3854  -0.963  0.33772
## factor(marital)married       -0.1494     2.0450  -0.073  0.94191
## factor(marital)never          0.7034     1.9420   0.362  0.71801
## factor(marital)widowed       10.5607     3.6002   2.933  0.00418 **
## factor(race)black             0.7059     4.6727   0.151  0.88024
## factor(race)hispanic         -2.0932     4.6002  -0.455  0.65011
## factor(race)multiracial      -2.6073     5.1740  -0.504  0.61545
## factor(race)native AM        -1.1466     5.4479  -0.210  0.83374
## factor(race)Pac Island       -3.9173     7.9825  -0.491  0.62472
## factor(race)white            -1.1457     4.2206  -0.271  0.78663
## factor(age)18-25              6.2766     4.2700   1.470  0.14482
## factor(age)26-34              5.8802     4.3007   1.367  0.17470
## factor(age)35-49              8.5553     4.1638   2.055  0.04260 *
## factor(age)50-64              5.7473     4.4169   1.301  0.19628
## factor(age)65+                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.63 on 97 degrees of freedom
## Multiple R-squared:  0.1822, Adjusted R-squared:  0.03043
## F-statistic: 1.201 on 18 and 97 DF,  p-value: 0.2761
```

e. Describe the results of the multivariate analysis, highlighting:

- the apparent association between each additional control variable and the focal response variable
- how the focal association changed when you incorporated each control variable
- the implications of these results for your focal association

The focal assocation decreased slightly, but the p-value improved dramatically by dropping in half. With all of the catagorical variables, I don't know how to read this information anymore and the apparent association between each additional control variable is lost to me. But the p-value reducing is important to note. I would think that these results are important to the focal association because the model better captures the social reality.

f. How well does the full (all explanatory variables included) model fit? Are any of the other models you ran a better fit? Explain how you came to the conclusion you did.

This has the smallest p-value by at least half, so I would say this fits the best.

g. Select the model that you think best fits the data. Provide a brief synopsis of the analysis of your data using this model and describe the implications for the theoretical arguments you set out to test.

```
lm_multi <- lm(binge30 ~ income + factor(edu) + factor(marital) + factor(race) + factor(age), data = tr
lm_multi
```

```
##
## Call:
## lm(formula = binge30 ~ income + factor(edu) + factor(marital) +
##     factor(race) + factor(age), data = treated_respondents)
##
## Coefficients:
##                 (Intercept)                      income
##                      0.8736                      0.3056
## factor(edu)AA/some college     factor(edu)college grad
##                     -3.9243                     -5.8237
##      factor(edu)highschool     factor(edu)less than HS
##                     -4.5474                     -5.1885
##     factor(marital)married        factor(marital)never
##                     -0.1494                      0.7034
##     factor(marital)widowed           factor(race)black
##                     10.5607                      0.7059
##       factor(race)hispanic     factor(race)multiracial
##                     -2.0932                     -2.6073
##      factor(race)native AM      factor(race)Pac Island
##                     -1.1466                     -3.9173
##          factor(race)white            factor(age)18-25
##                     -1.1457                      6.2766
##           factor(age)26-34            factor(age)35-49
##                      5.8802                      8.5553
##           factor(age)50-64              factor(age)65+
##                      5.7473                          NA
```

```
summary(lm_multi)
```

```
##
## Call:
## lm(formula = binge30 ~ income + factor(edu) + factor(marital) +
##     factor(race) + factor(age), data = treated_respondents)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.129  -3.013  -1.447   1.125  25.779
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.8736     5.6117   0.156  0.87662
## income                         0.3056     0.3896   0.784  0.43467
## factor(edu)AA/some college    -3.9243     5.4459  -0.721  0.47289
## factor(edu)college grad       -5.8237     5.7497  -1.013  0.31365
## factor(edu)highschool         -4.5474     5.4745  -0.831  0.40821
## factor(edu)less than HS       -5.1885     5.3854  -0.963  0.33772
## factor(marital)married        -0.1494     2.0450  -0.073  0.94191
## factor(marital)never           0.7034     1.9420   0.362  0.71801
```

```
## factor(marital)widowed      10.5607     3.6002   2.933  0.00418 **
## factor(race)black            0.7059     4.6727   0.151  0.88024
## factor(race)hispanic        -2.0932     4.6002  -0.455  0.65011
## factor(race)multiracial     -2.6073     5.1740  -0.504  0.61545
## factor(race)native AM       -1.1466     5.4479  -0.210  0.83374
## factor(race)Pac Island      -3.9173     7.9825  -0.491  0.62472
## factor(race)white           -1.1457     4.2206  -0.271  0.78663
## factor(age)18-25             6.2766     4.2700   1.470  0.14482
## factor(age)26-34             5.8802     4.3007   1.367  0.17470
## factor(age)35-49             8.5553     4.1638   2.055  0.04260 *
## factor(age)50-64             5.7473     4.4169   1.301  0.19628
## factor(age)65+                   NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.63 on 97 degrees of freedom
## Multiple R-squared:  0.1822, Adjusted R-squared:  0.03043
## F-statistic: 1.201 on 18 and 97 DF,  p-value: 0.2761
```

This linear model of incomes' association to # of days drinking in the past 30 among those who have attended treatment, controlling for education, marital status, race, and age, is the most theoretically comprehensive model I have made thus far. This is reflected in the smallest p-value compared to other models with fewer control variables.

The intercept tells us that those making 0 dollars would drink 0.8 times. The slope of 0.3056 tells us that for every increase in income bracket (as this is a catagorical variabel representing income brackets 0-10k, 10-20k, 20-30k...75k+), one can expect an increase of 0.3056 days binge drinking in the past month.

This model however, still seems incomplete, as it is completely thrown off by outliers.