Lab 15 - Multivariate Regression & Interpretation

Your name here

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.

The reason why I think the type of the crime also has a theoretical relationship with the total prison length is the amount of stereotyping in policing. Meaning that LE is more likely to police blacks.

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 this variable has a confounding relationship with my race variable is because while this variable may increase minority in the CJS, certain types of crimes may also increase the sentence severity more than others.

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

This is a categorical variable. This affects the multivariate regression by having many different parameters, instead of just one for each variable.

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.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidyve
```

```
## ✓ ggplot2 2.2.1     ✓ purrr   0.2.4
## ✓ tibble  1.3.4     ✓ dplyr   0.7.4
## ✓ tidyr   0.7.2     ✓ stringr 1.2.0
## ✓ readr   1.1.1     ✓ forcats 0.2.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# importing data
data <- read.csv("~/monitoring-federal-criminal-sentences/clean_data/merged_data/96-15.csv")

# only comparing blacks and whites and prison length less than 700 (dp and lwp above that so excluding
data2 <- data %>%
  filter(MONRACE <= 2) %>%
  filter(TOTPRISN <= 700)

race <- lm(TOTPRISN ~ factor(MONRACE), data2)
summary(race)
```

```
##
## Call:
## lm(formula = TOTPRISN ~ factor(MONRACE), data = data2)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -74.34 -36.16 -17.16  15.84 654.84
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      41.15506    0.06697   614.6   <2e-16 ***
## factor(MONRACE)2 33.18971    0.13133   252.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.81 on 1114262 degrees of freedom
## Multiple R-squared:  0.05421,    Adjusted R-squared:  0.05421
## F-statistic: 6.386e+04 on 1 and 1114262 DF,  p-value: < 2.2e-16
```

```
crime_race <- lm(TOTPRISN ~ factor(MONRACE) + factor(TYPE), data2)
summary(crime_race)
```

```
##
## Call:
## lm(formula = TOTPRISN ~ factor(MONRACE) + factor(TYPE), data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96.70 -29.28 -11.96  13.96 672.96
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       23.0370     0.1192  193.34   <2e-16 ***
## factor(MONRACE)2  20.9567     0.1303  160.79   <2e-16 ***
## factor(TYPE)1     42.2451     0.1428  295.77   <2e-16 ***
## factor(TYPE)2     39.9636     0.2051  194.81   <2e-16 ***
## factor(TYPE)3     -3.0225     0.1592  -18.99   <2e-16 ***
## factor(TYPE)4     52.7113     0.3186  165.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.89 on 1114229 degrees of freedom
##   (29 observations deleted due to missingness)
## Multiple R-squared:  0.1721, Adjusted R-squared:  0.1721
## F-statistic: 4.632e+04 on 5 and 1114229 DF,  p-value: < 2.2e-16
```

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 association between race (the main variable) and the tpe of the crime is quite significant. Blacks in general receive harsher sentences (20+ compared to whites), and drug crimes (type 1) receives an additional 41 months worth of imprisonment on average!

When I included the control - the correlation between blacks and sentence length became slightly weeker, and went to drug crimes. Blacks are still very signifciant.

This tells me that crime type is important to account for in future models, and lumping data is not a good idea in case of crime types.

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

This is not a improvement over my bivariate data because it has a weak r-squared, and lots of bad residuals at the tail end.

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:

```
library(tidyverse)

race_type_interaction <- lm(TOTPRISN ~ factor(MONRACE) + factor(TYPE) + XCRHISSR + XFOLSOR, data2)
summary(race_type_interaction)
```

```
##
## Call:
## lm(formula = TOTPRISN ~ factor(MONRACE) + factor(TYPE) + XCRHISSR +
##     XFOLSOR, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -216.07  -18.68   -3.97   11.61  635.99
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -67.783349   0.114664 -591.15   <2e-16 ***
## factor(MONRACE)2   7.767409   0.091235   85.14   <2e-16 ***
## factor(TYPE)1     -5.068605   0.107050  -47.35   <2e-16 ***
## factor(TYPE)2      6.895167   0.146055   47.21   <2e-16 ***
## factor(TYPE)3     -2.769758   0.112885  -24.54   <2e-16 ***
## factor(TYPE)4      6.573262   0.221508   29.68   <2e-16 ***
## XCRHISSR           7.605865   0.023304  326.37   <2e-16 ***
## XFOLSOR            5.156278   0.005112 1008.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.83 on 1106975 degrees of freedom
##   (7281 observations deleted due to missingness)
## Multiple R-squared:  0.6144, Adjusted R-squared:  0.6143
## F-statistic: 2.519e+05 on 7 and 1106975 DF,  p-value: < 2.2e-16
```

a. Describe the theoretical reasoning for selecting this variable.

Selecting both the criminal history and severity of the crime will help us dig deeper into the granularity of the type of the crime and the outcome of the punishment.

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?

These variables have a intervening relationship with my focal variable. I expect it to make the race and type variables weaker.

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

These are continuous variables. This just means that we have really granular data and it only counts for 1 linear paramter!

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.

```r
library(tidyverse)
# only comparing blacks and whites and prison length less than 700 (dp and lwp above that so excluding
data2 <- data %>%
  filter(MONRACE <= 2) %>%
  filter(TOTPRISN <= 700) %>%
  filter(EDUCATN < 3)

race_type_interaction <- lm(TOTPRISN ~ factor(MONRACE):YEAR + factor(MONRACE) + factor(TYPE) + XCRHISSR
summary(race_type_interaction)
```

```
##
## Call:
## lm(formula = TOTPRISN ~ factor(MONRACE):YEAR + factor(MONRACE) +
##     factor(TYPE) + XCRHISSR + XFOLSOR, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -223.90  -18.58   -4.01   12.36  616.56
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             44.920828  17.835170   2.519   0.0118 *
## factor(MONRACE)2       876.459994  33.244600  26.364  < 2e-16 ***
## factor(TYPE)1           -7.710244   0.119498 -64.522  < 2e-16 ***
## factor(TYPE)2            5.239015   0.155241  33.748  < 2e-16 ***
## factor(TYPE)3           -3.131153   0.124942 -25.061  < 2e-16 ***
## factor(TYPE)4            3.485879   0.237234  14.694  < 2e-16 ***
## XCRHISSR                 7.835824   0.024632 318.121  < 2e-16 ***
## XFOLSOR                  5.347210   0.005566 960.751  < 2e-16 ***
## factor(MONRACE)1:YEAR   -0.057448   0.008888  -6.463 1.03e-10 ***
## factor(MONRACE)2:YEAR   -0.490620   0.014051 -34.917  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.09 on 982589 degrees of freedom
##   (4715 observations deleted due to missingness)
## Multiple R-squared:  0.6213, Adjusted R-squared:  0.6213
## F-statistic: 1.791e+05 on 9 and 982589 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)
```

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

This is weird - accounting for XCRHISSR and XFOLSOR - drug crimes actually receive less prison time. The focal (race) did become weaker after including these two variables. This also means that the severity of the crime and the history of the individual plays a larger role in sentencing.

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.

I think no matter what I add - they will be significant and relatively strong. Perhaps I should consider keeping this data, and subsetting data to more disadvantaged individuals and go from there.

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.

```
final <- glm(TOTPRISN + XFOLSOR ~ factor(MONRACE) +
                                  factor(TYPE) +
                                  factor(EDUCATN)+
                                  AGE:factor(MONRACE) +
                                  YEAR +
                                  XCRHISSR,
                                  data = data2,
                                  family = "poisson")
summary(final)
```

```
##
## Call:
## glm(formula = TOTPRISN + XFOLSOR ~ factor(MONRACE) + factor(TYPE) +
##     factor(EDUCATN) + AGE:factor(MONRACE) + YEAR + XCRHISSR,
##     family = "poisson", data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.608   -4.622   -1.837    2.305   55.173
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## (Intercept)           7.292e+00  4.541e-02  160.579  < 2e-16 ***
## factor(MONRACE)2      4.670e-01  9.156e-04  510.116  < 2e-16 ***
## factor(TYPE)1         7.737e-01  3.848e-04 2010.824  < 2e-16 ***
## factor(TYPE)2         5.130e-01  4.843e-04 1059.238  < 2e-16 ***
## factor(TYPE)3        -2.806e-01  5.179e-04 -541.784  < 2e-16 ***
## factor(TYPE)4         7.153e-01  6.360e-04 1124.755  < 2e-16 ***
## factor(EDUCATN)1      7.425e-03  1.321e-03    5.619 1.92e-08 ***
## factor(EDUCATN)2      1.862e-02  1.327e-03   14.027  < 2e-16 ***
## YEAR                 -2.095e-03  2.263e-05  -92.591  < 2e-16 ***
## XCRHISSR              1.535e-01  6.710e-05 2286.978  < 2e-16 ***
## factor(MONRACE)1:AGE  7.606e-03  1.437e-05  529.380  < 2e-16 ***
## factor(MONRACE)2:AGE -2.034e-03  2.212e-05  -91.922  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 54340807  on 981652  degrees of freedom
## Residual deviance: 35072453  on 981641  degrees of freedom
##   (5661 observations deleted due to missingness)
## AIC: 40631498
##
## Number of Fisher Scoring iterations: 5
```

Blacks receive on average of 0.47 additonal months compared to whites. Drug, gun, and violent crime all receive higher sentences (by .5-.7 months) than other and immigration crimes. Education has a weak correlation with the outcome of the sentencing below college graduate level (Beta is small). There is a slight downward trend in crime severity in the recent 2 decades. And white people are more likely than blacks to commit more crime as they get older.