Lab 13 - Chi square, ANOVA, & correlation

Johnathan Hsu

November 21, 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 13 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 two categorical variables from your dataset whose association you're interested in and conduct a chi-square test. *If you only have continuous variables you will need to create categorical versions of these variables to make this work. You can do this using the* **`cut`** *function in mutate to add a new, categorical version of your variable to your dataset.*

Two categorical variables I will use here are education and race.

```
library(gmodels)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.0 --
## √ 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_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

data <- data %>%
  filter(MONRACE <= 2)

tbl = CrossTable(table(as.factor(data$MONRACE), as.factor(data$XCRHISSR)),
                 prop.r = FALSE,
                 prop.c = FALSE,
                 prop.t = FALSE,
                 prop.chisq = FALSE,
                 chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  1114635
##
##
##              |
##            |       1 |       2 |       3 |       4 |       5 |       6 | Row Total |
## -----------|---------|---------|---------|---------|---------|---------|-----------|
```

```
##          1 |   406636 |   105576 |   130800 |    74644 |    43378 |    63728 |   824762 |
## -------------|----------|----------|----------|----------|----------|----------|----------|
##          2 |    96869 |    33236 |    51193 |    34382 |    21666 |    52527 |   289873 |
## -------------|----------|----------|----------|----------|----------|----------|----------|
## Column Total |   503505 |   138812 |   181993 |   109026 |    65044 |   116255 |  1114635 |
## -------------|----------|----------|----------|----------|----------|----------|----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  38467.77     d.f. =  5     p =  0
##
##
##
```

a. Describe any modifications made to your data for the chi-square test and the composition of the variables used in the test (e.g., study time is measured using a three-category ordinal variable with categories indicating infrequent studying, medium studying, and frequent studying).

I did not have to make any modifications.

b. Does there appear to be an association between your two variables? Explain your reasoning.

As the p is 0 (<0.05), we reject the null hypothesis that criminal history is independent of the race of the defendant.

c. What are the degrees of freedom for this test and how is this calculated?

The degree of freedom is 24, calculated by (2-1) * (6 - 1) = 5

d. What if the critical value for the test statistic? What is the obtained value for the test statistic?

The critical value is infinity. THe obtained value is 0. But if I went for 0.05 CI, IT'D BE 11.07, but chi squared is 38467.77

e. How do you interpret the results of this test and the implications for your theoretical arguments about these two variables?

We reject the null hypothesis that criminal history is independent of the race of the defendant.

2. Select one continuous variable and one categorical variable from your dataset whose association you're interested in exploring. *Again, note that you'll need to create a categorical version of your independent variable to make this work.*

```r
prisontime_education <- data %>%
  select(c(TOTPRISN, EDUCATN, YEAR))

prison_edu_anova <- aov(TOTPRISN ~ as.factor(EDUCATN), data = prisontime_education)
summary.aov(prison_edu_anova)
```

```
##                         Df    Sum Sq Mean Sq F value Pr(>F)
## as.factor(EDUCATN)       5 3.666e+07 7332028    1770 <2e-16 ***
## Residuals          1067208 4.420e+09    4142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 47421 observations deleted due to missingness
```

a. Describe any modifications made to your data for the ANOVA test and the composition of the variables used in the test (e.g., college rank is measured using a four-category variable with values indicating freshman, sophomore, junior, and senior class).

The only thing I had to change was to switch education (independent categorical) to factors.

b. What are the degrees of freedom (both types) for this test and how are they calculated?

Degree of freedom for education was 5. This is calculated by 6 education categories - 1 = 5.

c. What is the obtained value of the test statistic?

The test statisitic is 1770.

d. What do the resuts tell you about the association between these two variables? What does this mean for your theoretical arguments about these variables?

We can reject the null hypothesis. There is strong association between these two variables.

3. Select two continuous variables from your dataset whos association you're interested in exploring.

```
# selecting total prison time sentenced and stat. max
continuous <- data %>%
  select(c(MONRACE, EDUCATN, TYPE, XCRHISSR, XFOLSOR, STATMIN, TOTPRISN, STATMAX)) %>%
  filter(STATMAX < 500)



# creating lm for the 2.

prison_statmax <- (lm(TOTPRISN ~ STATMAX, data = continuous))
summary(prison_statmax)
```

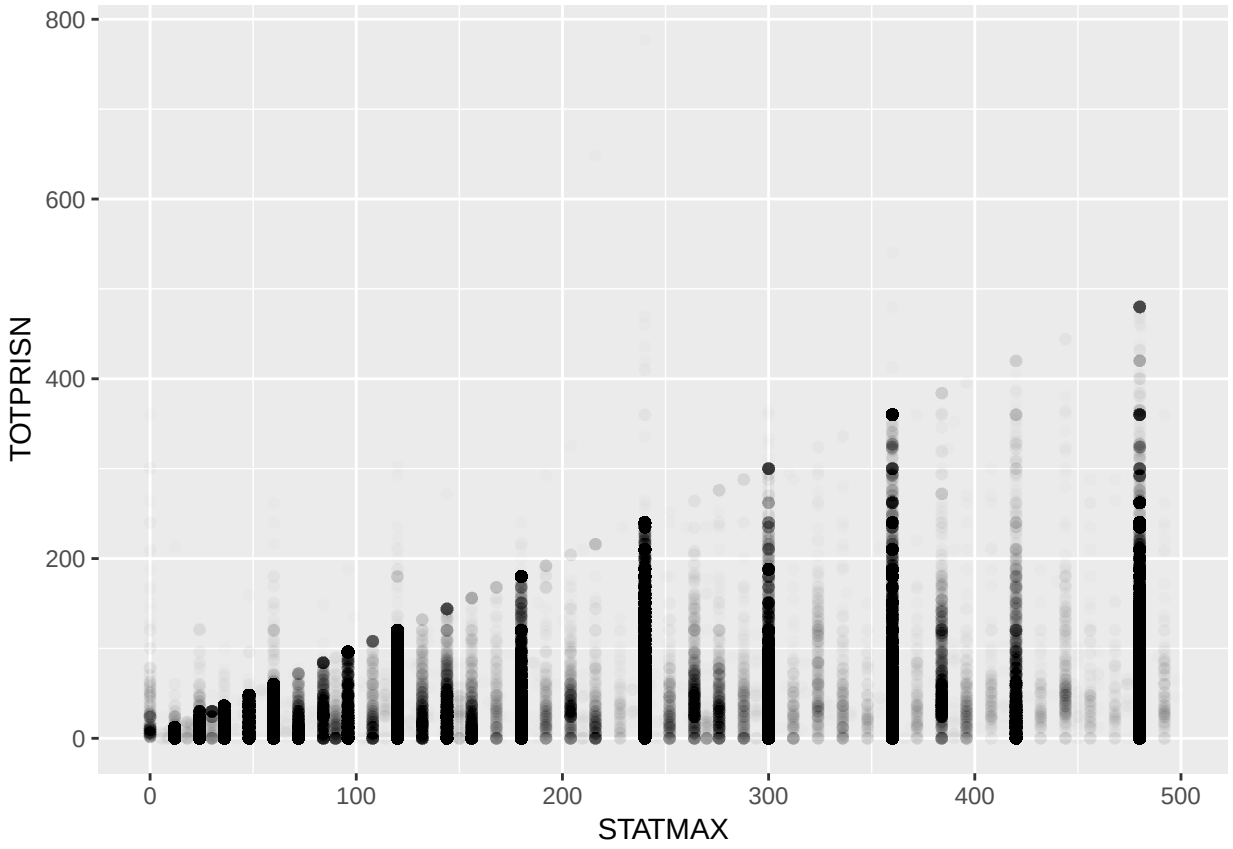
```
##
## Call:
## lm(formula = TOTPRISN ~ STATMAX, data = continuous)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.92 -16.68  -6.78   9.22 740.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.984744   0.061964   161.1   <2e-16 ***
## STATMAX     0.111664   0.000257   434.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.9 on 866671 degrees of freedom
## Multiple R-squared:  0.1788, Adjusted R-squared:  0.1788
## F-statistic: 1.887e+05 on 1 and 866671 DF,  p-value: < 2.2e-16
```

a. What is the correlation between these two variables?

There is a weak but significant correlation between statutory max and total prison sentenced (beta = 0.1, and $p < 0.001$!)

b. Create a scatterplot of the variables you selected. Does the correlation coefficient accurately represent the relationship between these two variables? Why or why not?

```
p <- ggplot(continuous, aes(STATMAX, TOTPRISN))
p + geom_point(alpha = 0.01)
```



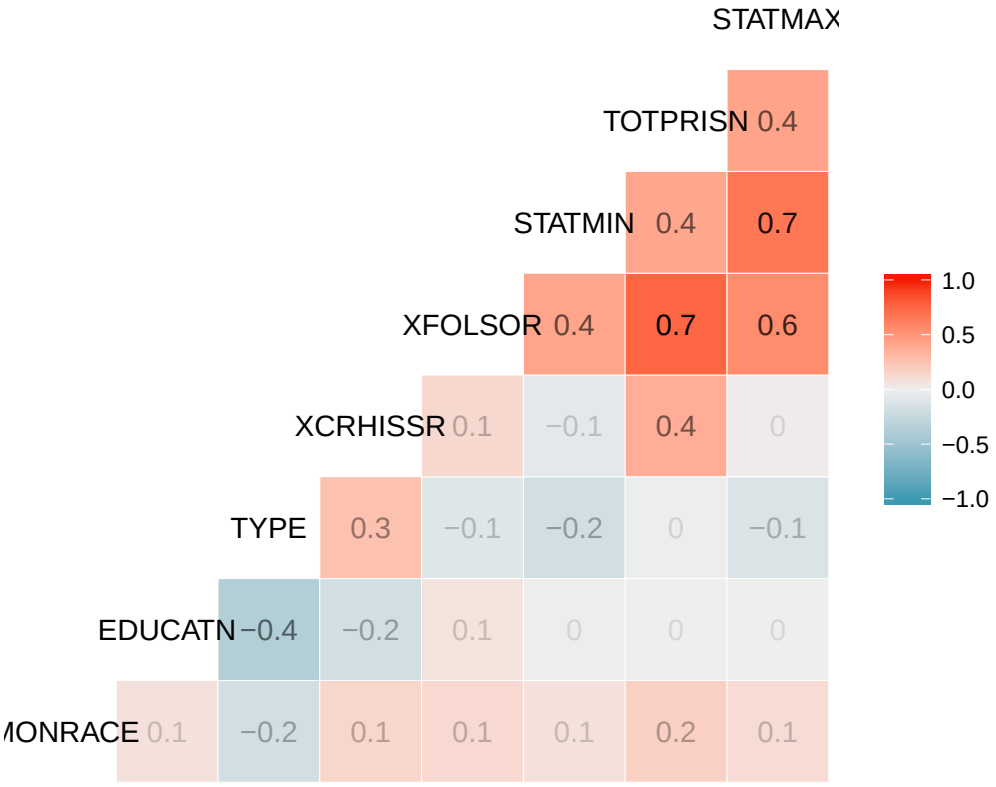
c. Create a correlation matrix of your data using the `ggcorr` function from the `GGally` package. Be sure to label each cell with the correlation coefficient.

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggcorr(continuous,
       label = TRUE,
       label_alpha = TRUE)
```



d. What does this visual representation of correlation coefficients tell you about your data? Are there any relationships (or lack thereof) that are surprising to you? Why or why not?

This representation of correlation tells me that the total calculated offense level and criminal history are strongest indicators of how severe the sentence length is.

e. What are the limitations of correlation coefficients? Can they ever be misleading? If so, in what ways?

They can be misleading - because other factors can act as intervening variables to how sever the outcome is, we see that race has 0.2, which is not trivial at all.