Correlation & Measurement

Note: The dataset used in this tutorial and the R Script are on Moodle:

Loading the 2016 CCES dataset

```
install.packages("foreign", dependencies=TRUE)
library(foreign)

dat <- read.dta(file.choose(), convert.factors=FALSE)
```

Recode Variables Quickly

1. Here, we recode a few variables that we will need later on.

```
table(dat$CC16_340a)
dat$ideology <- recode(dat$CC16_340a, "8=NA")
summary(dat$ideology)

table(dat$birthyr)
dat$age <- 2016 - dat$birthyr
table(dat$age)
summary(dat$age)

table(dat$faminc)
dat$income <- recode(dat$faminc, "31=NA; 97=NA")
table(dat$income)
summary(dat$income)
```

Bivariate Tests

1. When we have a continuous dependent variable, we can test how correlated any independent variable is with the dependent variable. The test can be conducted in an number of ways. For example, let us say that we again wanted to explore the impact of income on ideology.

```
scatterplot(dat$ideology, dat$income)
```

The scatterplot does not really give us a good picture of the relationship in this situation since the variables are both not truly continuous. However, if we were using truly continuous variables, a scatterplot would be the first step in demonstrating a relationship.

Below, we start to test whether the relationship is significant.

```
mat1 <- na.omit(data.frame(
ideology = dat$ideology,
income = dat$income
))

cor(mat1)
```

The result indicates that as income increases, ideology decreases. As you become more wealthy you become more liberal. However, the result does not indicate whether the relationship is statistically significant. Instead, it only indicates the potential strength of the relationship.

2. Or, we could estimate a model that gives us even more information about the relationship.

```
mod <- lm(ideology ~ income, data=dat)

summary(mod)
```

The model results tell us that for every one unit increase in income we decrease .007 in ideology. Since the income variable is measure from 1-15 and the ideology variable is measured from 1-7, the relationship is weak. However, the output indicates that the results are statistically significant. The t statistic is larger smaller than -1.96 and the p ¡ 0.05. Therefore, we have a weak, statistically significant relationship. Finally, the R squared statistic is .002. That means on average our independent variable explains .2% of the variation in the dependent variable.

3. What if we wanted to quickly check the bivariate relationships between more than 2 variables? We could do this by creating a matrix and then testing the bivariate relationships in the matrix.

```
mat2 <- na.omit(data.frame(
ideology = dat$ideology,
income = dat$income,
age = dat$age
))

cor(mat2)
```

We see the same result with income as we did before. In terms of the relationship between age and income and ideology. We see that as you age your income increases and your ideology becomes more conservative.

Measurement

Let's say that we want to get at an underlying latent trait of racism that people might hold. However, we recognize that one measure of racism is not enough to measure such a trait. Therefore, we can take several measures, assess their correlation, and measure a singular latent racism measure. Here, we do that with four questions from the CCES. Note: it is helpful to recode the variables so that they are all in the same theoretical direction.

I am angry that racism exists.

```
table(dat$CC16_422c)

dat$angry <- recode(dat$CC16_422c, "1=-2; 2=-1; 3=0; 4=1; 5=2")
table(dat$angry)
```

White people in the U.S. have certain advantages because of the color of their skin.

```
table(dat$CC16_422d)

dat$advance <- recode(dat$CC16_422d, "1=-2; 2=-1; 3=0; 4=1; 5=2")
table(dat$advance)
```

I often find myself fearful of people of other races.

```
table(dat$CC16_422e)

dat$fear <- recode(dat$CC16_422e, "1=2; 2=1; 3=0; 4=-1; 5=-2")
table(dat$fear)
```

Racial Problems in the U.S. are rare, isolated situations.

```
table(dat$CC16_422f)

dat$prob <- recode(dat$CC16_422f, "1=2; 2=1; 3=0; 4=-1; 5=-2")
table(dat$prob)
```

As before, we can check the correlation between these variables using a correlation matrix.

```
mat3 <- data.frame(
angry = dat$angry,
advance = dat$advance,
fear = dat$fear,
prob = dat$prob
)

cor(na.omit(mat3))
```

Next, we can calculate a statistic that let's us know if these variables can be combined. Generally, anything over .5 indicates a relationship.

```
cronbach(mat3)
```

Next, we can utilize the factor analysis statistical technique in order to estimate a singular latent variable and scores based on the relative weighting based on importance of these four variables.

```
antiimm <- fa(mat3, fm="pa", rotate="promax")
loadings(antiimm)
```

Finally, we can take this latent variable and the scores and add them back to the dataset.

```
dat$antiimm <- scale(antiimm$scores)
```

Lab Activity

In the 2020 Finland European Social Survey dataset, you are to find the variable labels for the four variables provided below assessing the state of trust in governmental institutions in Finland. Then, explore the four variables and recode them so that they are in a usable format. Then, create a correlation matrix with the four variables and explain what the results conveys substantively. Finally, estimate one latent variable that represents trust in the system.

1. Trust in the country's parliament.

2. Trust in the legal system.

3. Trust in politicians.

4. Trust in political parties.