POLSCI 9590: Methods I

Measures of Association for Interval/Ratio Data

Dave Armstrong

Videos

In the videos for today, we learned about:

1. Pearson Correlation Coefficient.
    - Linear relationships
    - Significance Testing
    - Correlation Matrix

Calculating the Correlation Coefficient.

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \times \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Which is

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Varaiance}(x) \times \text{Variance}(y)}}$$

Covariance is an *unbounded* measure of linear association (the scale is based on the values of x and y).

- dividing by the variances of x and y re-scales the values to live in the range $-1 \leq r \leq 1$.

Properties of the Correlation Coefficient

1. Measures **linear** association between variables.
    - This is only one of an infinite set of relationships that could exist, though often it is sufficient to characterize the relationship.
2. Ranges from $-1 \leq r \leq 1$, such that numbers farther from zero indicate stronger relationships (but indifferent directions).
3. The squared correlation coefficient r^2 tells us the proportion of variance in y that is explained by x.

Tests for Statistical Significance.

- Approximate z-statistic with $\mu = 0$ and $\sigma = \frac{1}{\sqrt{n-3}}$

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

- Approximate t-statistic with $n-2$ degrees of freedom

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Permutation test.
  - randomly re-arrange y and calculate $r_{xy}^{(t)}$ for $t = \{1, \ldots, T\}$.
  - $p = \frac{1}{T}\sum_{t=1}^{T} I\left(r_{xy}^{(t)} > r_{xy}\right)$: number of times random r is bigger than original r divided by number of random draws

Correlations in Software

R | Python | Stata

The `cor()` function makes correlations in R.

```
library(rio)
ces <- import("ces19.dta")
cor(ces$leader_con, ces$leader_lib)
```

```
## [1] NA
```

```
cor(ces$leader_con, ces$leader_lib, use="pairwise.complete")
```

```
## [1] -0.3643449
```

Correlation Matrix

R Python Stata

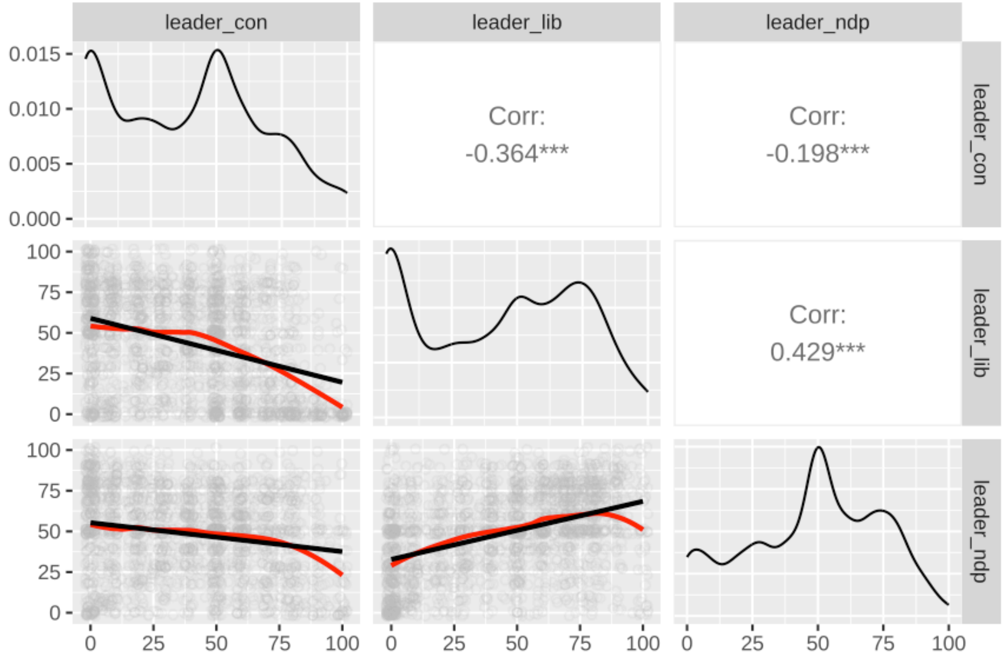
```r
therms = ces %>% select(starts_with("leader")) %>% na.omit()
DAMisc::pwCorrMat(~.,
                  data=therms,
                  method="sim")
```

```
## Pairwise Correlations
##            leader_con leader_lib leader_ndp
## leader_con
## leader_lib -0.364*
## leader_ndp -0.198*    0.429*
```

Is a Linear Relationship Appropriate?

R   Python   Stata

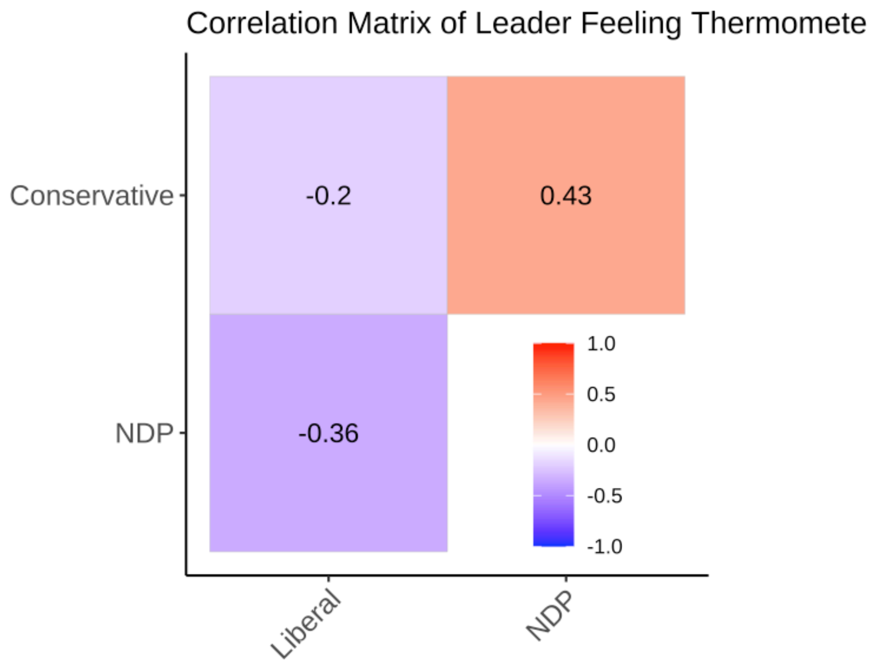
```r
library(GGally)
custom_smooth <- function(data, mapping,
  ..., span=.35, pt.alpha=.25, jitter=TRUE) {
  if(jitter){
    pos <- position_jitter(width=2, height=2)
  }else{
    pos <- position_identity()
  }

  ggplot(data, mapping, ...) +
    geom_point(shape=1, col="gray",
               position=pos, alpha=pt.alpha) +
    geom_smooth(method="loess", span=span,
                family="symmetric",
                se=FALSE, col="red") +
    geom_smooth(method="lm", col="black", se=FALSE)
}
ggpairs(therms,
  lower = list(continuous = wrap(custom_smooth,
               span=.5,
               pt.alpha=.15,
               jitter=TRUE))) +
theme(legend.position = "bottom")
```



Visual Correlation Matrix

R Python Stata

```r
library(ggplot2)
library(ggcorrplot)
r <- cor(therms)
colnames(r) <- rownames(r) <- c("Liberal", "NDP",
                                "Conservative")
ggcorrplot(r,
           ggtheme = theme_classic,
           lab=TRUE,
           type="upper",
           show.diag=FALSE) +
  theme(legend.position = "inside",
        legend.position.inside=c(.75, .25),
        legend.background=element_rect(fill="transparent"),
        legend.title = element_blank()) +
  ggtitle("Correlation Matrix of Leader Feeling Thermometers")
```



Exercises.

Using the `prestige` data from the `carData` package ...

1. Calculate the correlation between `prestige`, `income`, `education` and `women`.
2. Make the correlation plot.
3. Use the `ggpairs()` function to evaluate whether or not the correlation is a good measure of association for these variables.