Correlation and Linear Regression

POSC 3410 – Quantitative Methods in Political Science

Steven V. Miller

Department of Political Science

Goal for Today

Use correlation and linear regression to describe the relationship between two interval-level variables.

Building Toward Normal Political Science

Everything we have done is building toward normal quantitative research.

- We have concepts of interest, operationalized to variables.
- We observe central tendencies and variation in our variables.
- We believe there is cause and effect.
    - Though, importantly, we need to make controlled comparisons.
- We make inference about our claim of cause and effect using the logic of random sampling.

If our sample statistic is more than 1.96 standard errors from a proposed population parameter, we have a lot of confidence (95%) rejecting the proposed population parameter.

What We Will Be Doing Today

We'll go over the following two topics.

1. **Correlation analysis**
2. **Regression analysis**

R Packages We'll Be Using

```r
library(tidyverse) # for all things workflow
library(stevemisc) # for various formatting things
library(stevedata) # for my toy data, including election_turnout
```

Correlation

Question: does a state's voter turnout vary by the state's level of education?

- Education: % of state with high school diploma. (CPS estimates for 2015)
- Turnout: voter turnout for highest office (i.e. president) in 2016 general election.

We get a preliminary judgment using a **scatterplot**.

- But first: let's look at our data a bit.

Students Always Ask These Questions...

Least-educated states in the U.S.

```
election_turnout %>% select(state, perhsed) %>%
  top_n(-5, perhsed) %>% arrange(perhsed)
```

```
## # A tibble: 6 x 2
##   state       perhsed
##   <chr>         <dbl>
## 1 California     81.8
## 2 Texas          81.9
## 3 Mississippi    82.3
## 4 Louisiana      83.4
## 5 Kentucky       84.2
## 6 New Mexico     84.2
```

Be Mindful of Your Education Indicator...

```
election_turnout %>% select(state, percoled) %>%
  top_n(-5, percoled) %>% arrange(percoled)
```

```
## # A tibble: 5 x 2
##   state         percoled
##   <chr>            <dbl>
## 1 West Virginia     19.2
## 2 Mississippi       20.7
## 3 Arkansas          21.1
## 4 Kentucky          22.3
## 5 Louisiana         22.5
```

What About the Most Educated?

```r
election_turnout %>% select(state, perhsed) %>%
  top_n(5, perhsed) %>% arrange(-perhsed)
```

```
## # A tibble: 5 x 2
##   state         perhsed
##   <chr>           <dbl>
## 1 Montana          92.8
## 2 Minnesota        92.4
## 3 New Hampshire    92.3
## 4 Wyoming          92.3
## 5 Alaska           92.1
```

Again, College is Different...

```
election_turnout %>% select(state, percoled) %>%
  top_n(5, percoled) %>% arrange(-percoled)
```

```
## # A tibble: 5 x 2
##   state                percoled
##   <chr>                   <dbl>
## 1 District of Columbia     54.6
## 2 Massachusetts            40.5
## 3 Colorado                 38.1
## 4 Maryland                 37.9
## 5 Connecticut              37.6
```

On Voter Turnout in 2016...

```
election_turnout %>% select(state, turnoutho) %>%
  top_n(5, turnoutho) %>% arrange(-turnoutho)
```

```
## # A tibble: 5 x 2
##   state         turnoutho
##   <chr>             <dbl>
## 1 Minnesota          74.2
## 2 New Hampshire      71.4
## 3 Maine              70.5
## 4 Colorado           70.1
## 5 Wisconsin          69.4
```

Lowest Turnout States

```r
election_turnout %>% select(state, turnoutho) %>%
  top_n(-5, turnoutho) %>% arrange(turnoutho)
```

```
## # A tibble: 5 x 2
##   state         turnoutho
##   <chr>             <dbl>
## 1 Hawaii             42.2
## 2 West Virginia      50.1
## 3 Tennessee          51.2
## 4 Texas              51.6
## 5 Oklahoma           52.4
```

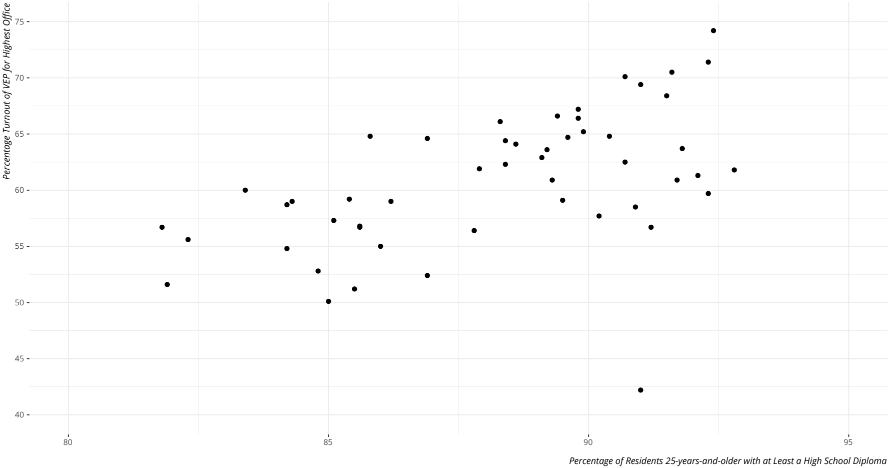
On South Carolina

If you’re curious about South Carolina:

- 12th lowest in % of state with a college diploma (25.8%).
- 14th lowest in % of state with at least a HS diploma (85.6%).
- 12th lowest in voter turnout in 2016 (56.7%)

A Scatterplot of State-Level Education and Voter Turnout in the 2016 General Election

The data are scattered in a formal consistent/positive way. Hawaii was always going to be a clear outlier.



Data: Elections Project, U.S. Census Bureau. Assembled in stevedata package available on Github (svmiller/stevedata).

Correlation

This relationship looks easy enough: positive.

- The relationship is not perfect, but it looks fairly "strong".

How strong? **Pearson's correlation coefficient** (or **Pearson's *r***) will tell us.

Pearson's *r*

$$\sum \frac{\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

...where:

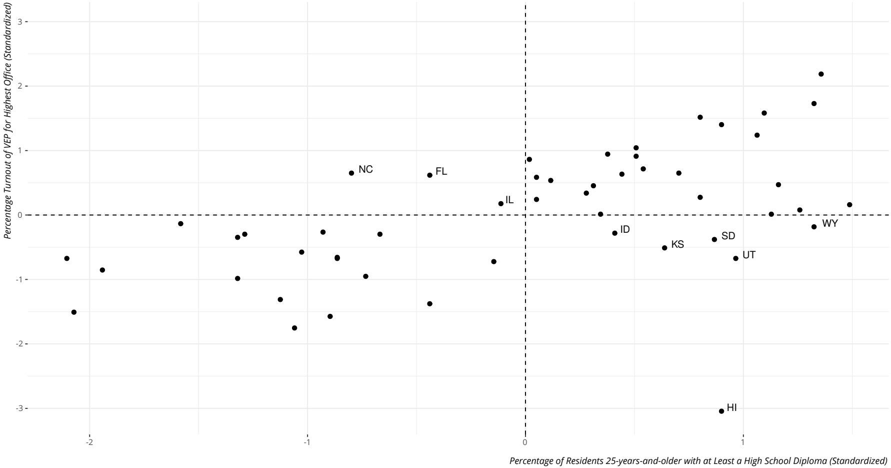
- x_i, y_i = individual observations of *x* or *y*, respectively.
- $\overline{x}$, $\overline{y}$ = sample means of *x* and *y*, respectively.
- s_x, s_y = sample standard deviations of *x* and *y*, respectively.
- *n* = number of observations in the sample.

Properties of Pearsons r

1. Pearson's r is symmetrical.
2. Pearson's r is bound between -1 and 1.
3. Pearson's r is standardized.

A Scatterplot of State-Level Education and Voter Turnout in the 2016 General Election

Observations in the negative correlation quadrants are highlighted for emphasis.



Data: Elections Project, U.S. Census Bureau. Assembled in stevedata package available on Github (svmiller/stevedata).

Education and Turnout (Z Scores)

- Cases in upper-right quadrant are above the mean in both x and y.
- Cases in lower-left quadrant are below the mean in both x and y.
- Upper-left and lower-right quadrants are negative-correlation quadrants.

All told, our Pearson's r is 26.41369/50, or .52.

- We would informally call this a fairly strong positive relationship.

...or in R

```
election_turnout %>%
  mutate(z_perhsed = (perhsed - mean(perhsed))/sd(perhsed),
         z_turnoutho = (turnoutho - mean(turnoutho))/sd(turnoutho)) -> election_turnout

with(election_turnout, sum(z_perhsed*z_turnoutho)/(length(state)-1))
```

```
## [1] 0.5282739
```

```
with(election_turnout, cor(perhsed,turnoutho))
```

```
## [1] 0.5282739
```

If You're Curious about the Hawaii Outlier...

```
election_turnout %>%
  filter(state != "Hawaii") %>%
  summarize(cor = cor(perhsed, turnoutho))
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.654
```

Linear Regression

Correlation has a lot of nice properties.

- It's another "first step" analytical tool.
- Useful for detecting **multicollinearity**.
    - This is when two independent variables correlate so highly that no partial effect for either can be summarized.

However, it's neutral on what is x and what is y.

- It won't communicate cause and effect.

Fortunately, regression does that for us.

Demystifying Regression

Does this look familiar?

$$y = mx + b$$

Demystifying Regression

That was the slope-intercept equation.

- b is the intercept: the observed y when $x = 0$.
- m is the familiar "rise over run", measuring the amount of change in y for a unit change in x.

Demystifying Regression

The slope-intercept equation is, in essence, the representation of a regression line.

- However, statisticians prefer a different rendering of the same concept measuring linear change.

$$y = a + b(x)$$

The *b* is the **regression coefficient** that communicates the change in *y* for each unit change in *x*.

A Simple Example

Suppose I want to explain your test score (*y*) by reference to how many hours you studied for it (*x*).

Table 1: Hours Spent Studying and Exam Score

Hours (x)	*Score (y)*
0	55
1	61
2	67
3	73
4	79
5	85
6	91
7	97

A Simple Example

In this eight-student class, the cherub who studied 0 hours got a 55.

- The cherub who studied 1 hour got a 61.
- The cherub who studied 2 hours got a 67.
- ...and so on...

Each hour studied corresponds with a six-unit change in test score. Alternatively:

$$y = a + b(x) = \text{Test Score} = 55 + 6(x)$$

Notice that our *y*-intercept is meaningful.

A Slightly Less Simple Example

However, real data are never that simple. Let's complicate it a bit.

Table 2: Hours Spent Studying, Exam Score, and Estimated Score

Hours (x)	*Score (y)*	*Estimated Score ($\hat{y}$)*
0	53	55
0	57	
1	59	61
1	63	
2	65	67
2	69	
3	71	73
3	75	
4	77	79
4	81	
5	83	85
5	87	
6	89	91
6	93	
7	95	97
7	99	

A Slightly Less Simple Example

Complicating it a bit doesn't change the regression line.

- Notice that regression averages over differences.
- An additional hour studied, *on average*, corresponds with a six-unit increase in the exam score.
- We have observed data points (*y*) and our estimates ($\hat{y}$, or *y*-hat).

Our Full Regression Line

Thus, we get this form of the regression line.

$$\hat{y} = \hat{a} + \hat{b}(x) + e$$

...where:

- $\hat{y}$, $\hat{a}$ and $\hat{b}$ are estimates of *y*, *a*, and *b* over the data.
- *e* is the error term.
    - It contains random sampling error, prediction error, and predictors not included in the model.

Getting a Regression Coefficient

How do we get a regression coefficient for more complicated data?

- Start with the **prediction error**, formally: $y_i - \hat{y}$.
- Square them. In other words: $(y_i - \hat{y})^2$
  - If you didn't, the sum of prediction errors would equal zero.

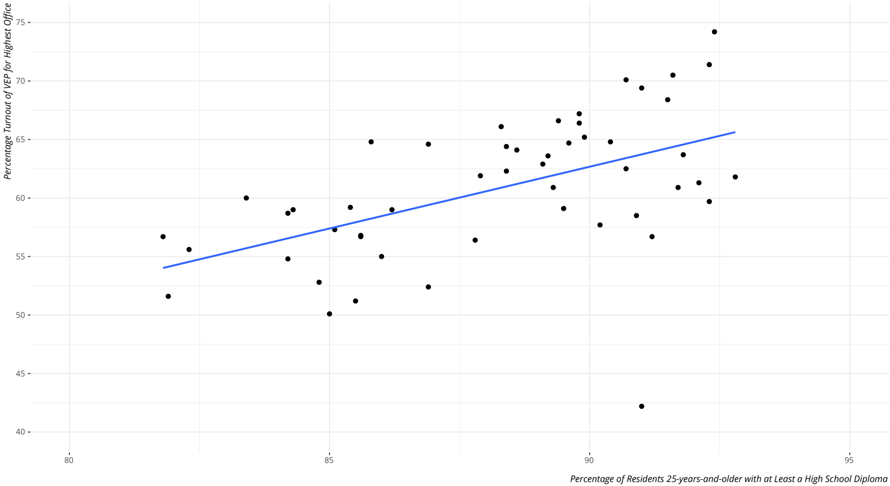
The regression coefficient that emerges minimizes the sum of squared differences ($(y_i - \hat{y})^2$).

- Put another way: "ordinary least squares" (OLS) regression.

The next figure offers a representation of this for our state education and turnout example.

Education and Turnout in the 2016 General Election

The line that minimizes the sum of squared prediction errors is drawn through these points.



Standard Error of Regression Coefficient

Each parameter in the regression model comes with a “standard error.”

- These estimate how precisely the model estimates the coefficient’s unknown value.

This has a convoluted estimation procedure.

- Namely: you need the diagonal of the square root of the variance-covariance matrix.
- This requires matrix algebra, and you probably hate me enough. :P

It’s standard output in a regression formula object in R, though.

If You're Curious...

```r
summary(M1 <- lm(turnoutho ~ perhsed, data=election_turnout))
```

```
##
## Call:
## lm(formula = turnoutho ~ perhsed, data = election_turnout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.529  -3.510   1.176   3.676   8.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.3027    21.3948  -1.510    0.138
## perhsed       1.0553     0.2423   4.355 6.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.247 on 49 degrees of freedom
## Multiple R-squared:  0.2791, Adjusted R-squared:  0.2644
## F-statistic: 18.97 on 1 and 49 DF,  p-value: 6.765e-05
```

If You're Curious...

```
X <- model.matrix(M1) # Intercept + perhsed

# Residual sum of squares
sigma2 <- sum((election_turnout$turnoutho - fitted(M1))^2) / (nrow(X) - ncol(X))

sqrt(sigma2) # residual standard error
```

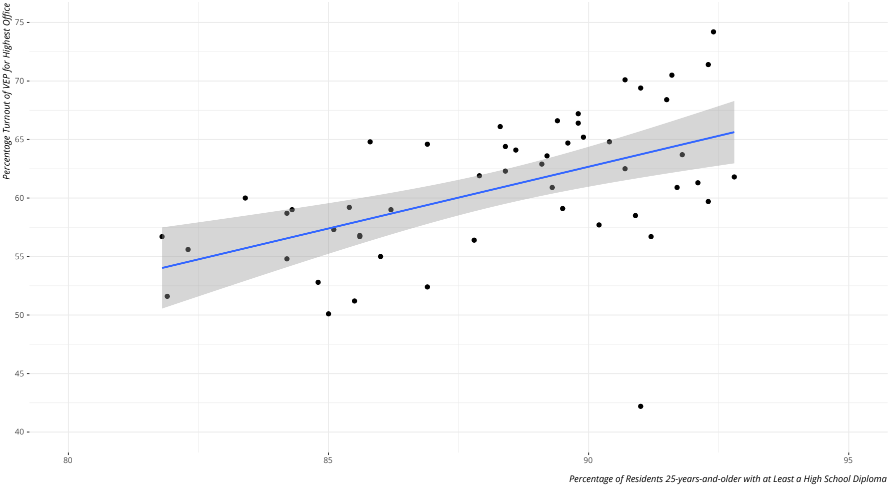
```
## [1] 5.246687
```

```
sqrt(diag(solve(crossprod(X))) * sigma2)
```

```
## (Intercept)     perhsed
##   21.394761    0.242304
```

Education and Turnout in the 2016 General Election

The line that minimizes the sum of squared prediction errors is drawn through these points.



Regression: Education and Turnout

This would be our regression line:

$$\hat{y} = -32.30 + 1.05(x)$$

How to interpret this:

- The state in which no one graduated from high school would have a voter turnout of -32.30%.
    - *Center your variables, people. Seriously...*
- Each unit increase in the percentage of the state's citizens having a high school diploma corresponds with an estimated 1.05% increase in voter turnout.

Inference in Regression

What do we say about that *b*-hat ($\hat{b}$ = 1.05?)

- If we took another "sample", would we observe something drastically different?
- How would we know?

Inference in Regression

You’ve done this before. Remember our last set of lectures? And Z scores?

$$Z = \frac{\overline{x} - \mu}{s.e.}$$

Inference in Regression

We do the same thing, but with a Student's *t*-distribution.

$$t = \frac{\hat{b} - \beta}{s.e.}$$

$\hat{b}$ is our regression coefficient. What is out β?

Inference in Regression

β is actually zero!

- We are testing whether our regression coefficient is an artifact of the "sampling process".
- We're testing a competing hypothesis that there is no relationship between x and y.

Inference in Regression

This makes things a lot simpler.

$$t = \frac{\hat{b}}{s.e.}$$

Inference in Regression

In our state education and turnout example, this turns out nicely.

$$t = \frac{1.05}{.24} = 4.35$$

Our regression coefficient is more than four standard errors from zero .

- The probability of observing it if β were really zero is .000067.
- We judge our regression coefficient to be statistically significant.

Conclusion

Hopefully, this lecture demystified regression.

- It builds on everything discussed to this point.
- The same process of inference from sample to population is used.
- Really nothing to it but to do it, I ’spose.

We’re going to add a fair bit on top of this next.

- If you understand this, everything else to follow is basically window dressing.

Table of Contents