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**

Correlation

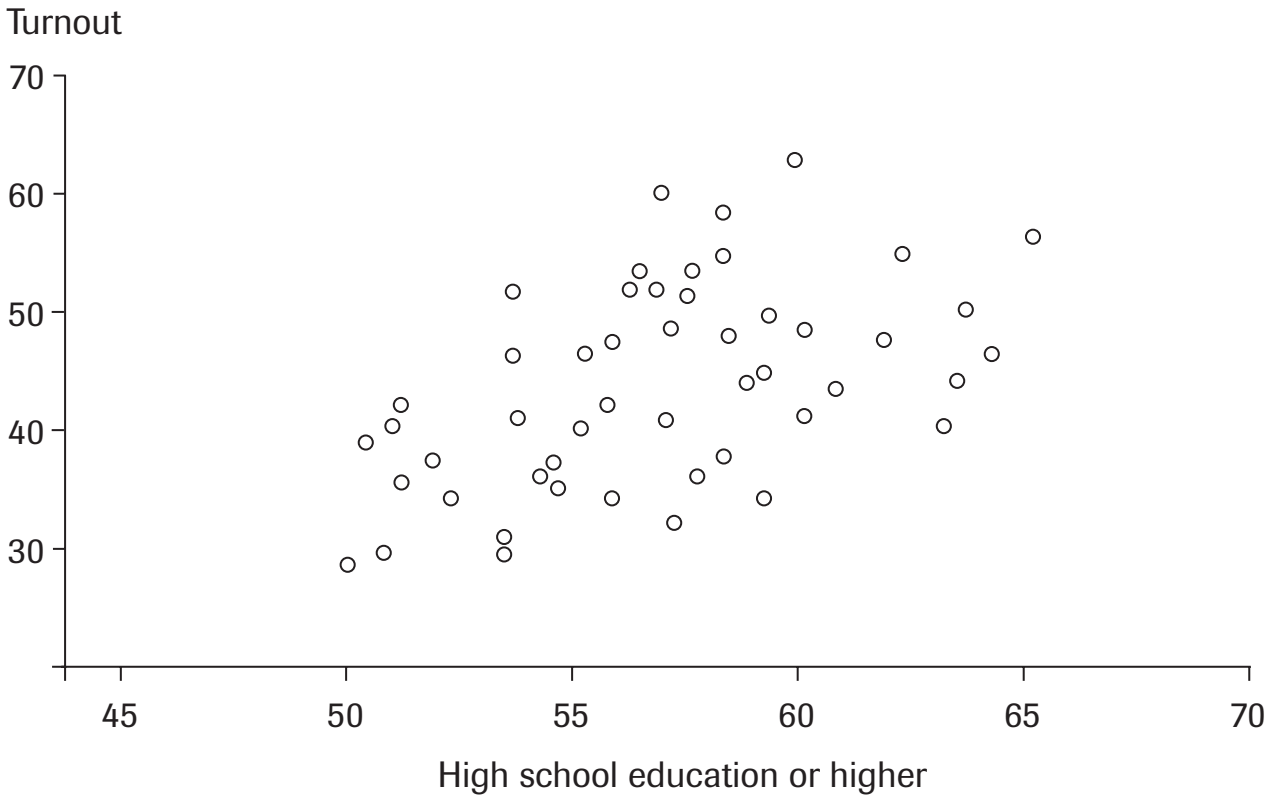
Question: does a state's voter turnout vary by the state's level of education?

- Education: % of state with high school diploma. (Census)
- Turnout: % of voters among voting eligible population in 2006 elections.

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

Education and Turnout

Figure 8–1 Scatterplot: Education and Turnout (in percentages)



Source: Percentage high school or higher calculated from U.S. Census Bureau, www.census.gov/compendia/smadb/TableA-22.pdf. Turnout is percentage of voting eligible population (VEP) in the 2006 elections, calculated by Michael McDonald, Department of Public and International Affairs, George Mason University, Fairfax, Virginia, and made available through his Web site: http://elections.gmu.edu/voter_turnout.htm.

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:

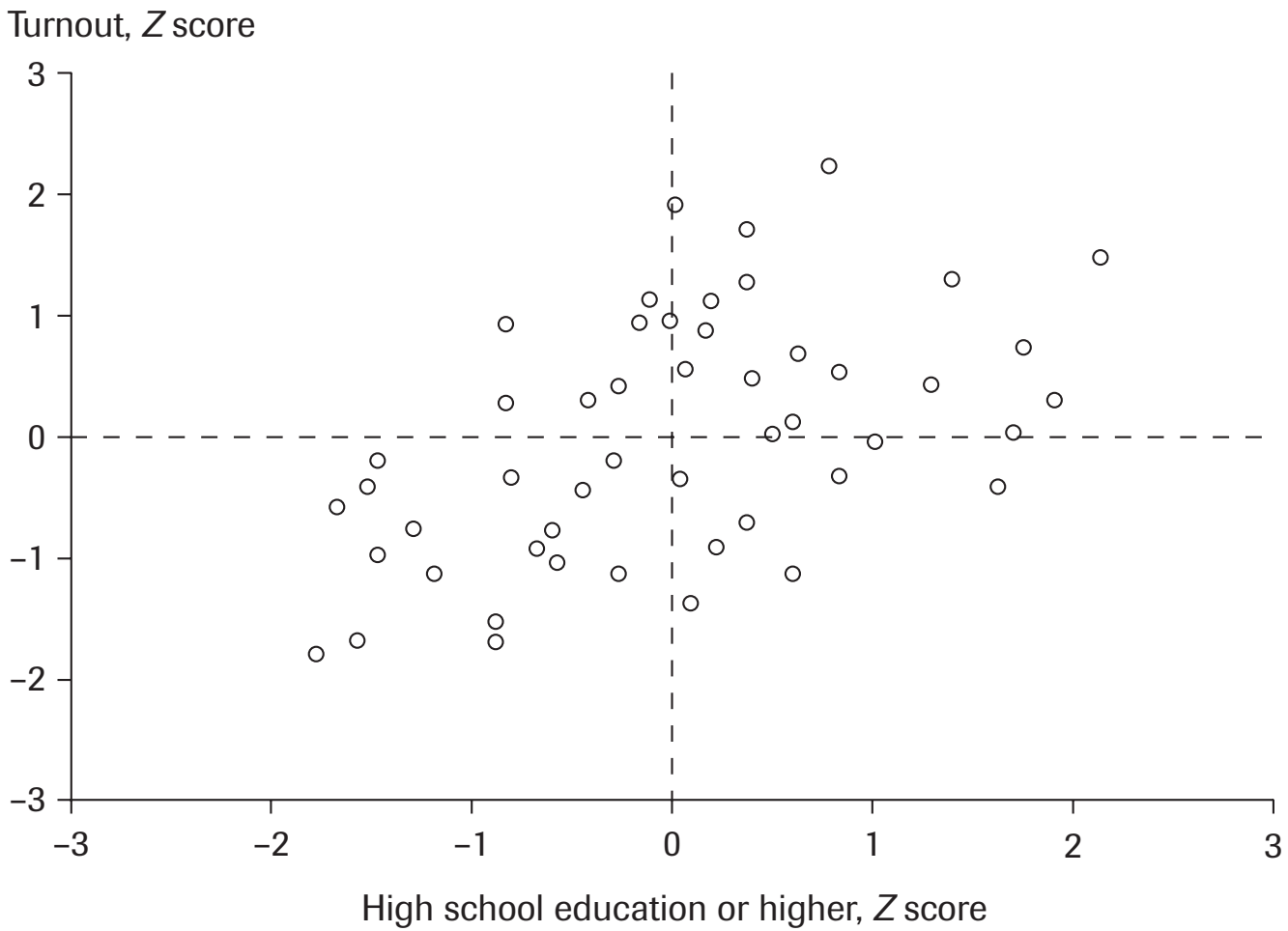
- x_i, y_i = individual observations of x or y, respectively.
- $\overline{x}$, $\overline{y}$ = sample means of x and y, respectively.
- s_x, s_y = sample standard deviations of x and y, respectively.
- n = number of observations in the sample.

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.

Education and Turnout (Z Scores)

Figure 8-2 Scatterplot: Education and Turnout (*Z* scores)



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 $24.94/49$, or .51.

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

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 8-1 Hours Spent Studying (x) and Test Score (y)

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

Note: Hypothetical data.

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 8–2 Hours Spent Studying (x), Test Score (y), and Estimated Score ($\hat{y}$)

Hours (x)	Score (y)	Estimated score ($\hat{y}$) for a given value of x
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	

Note: Hypothetical data.

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.

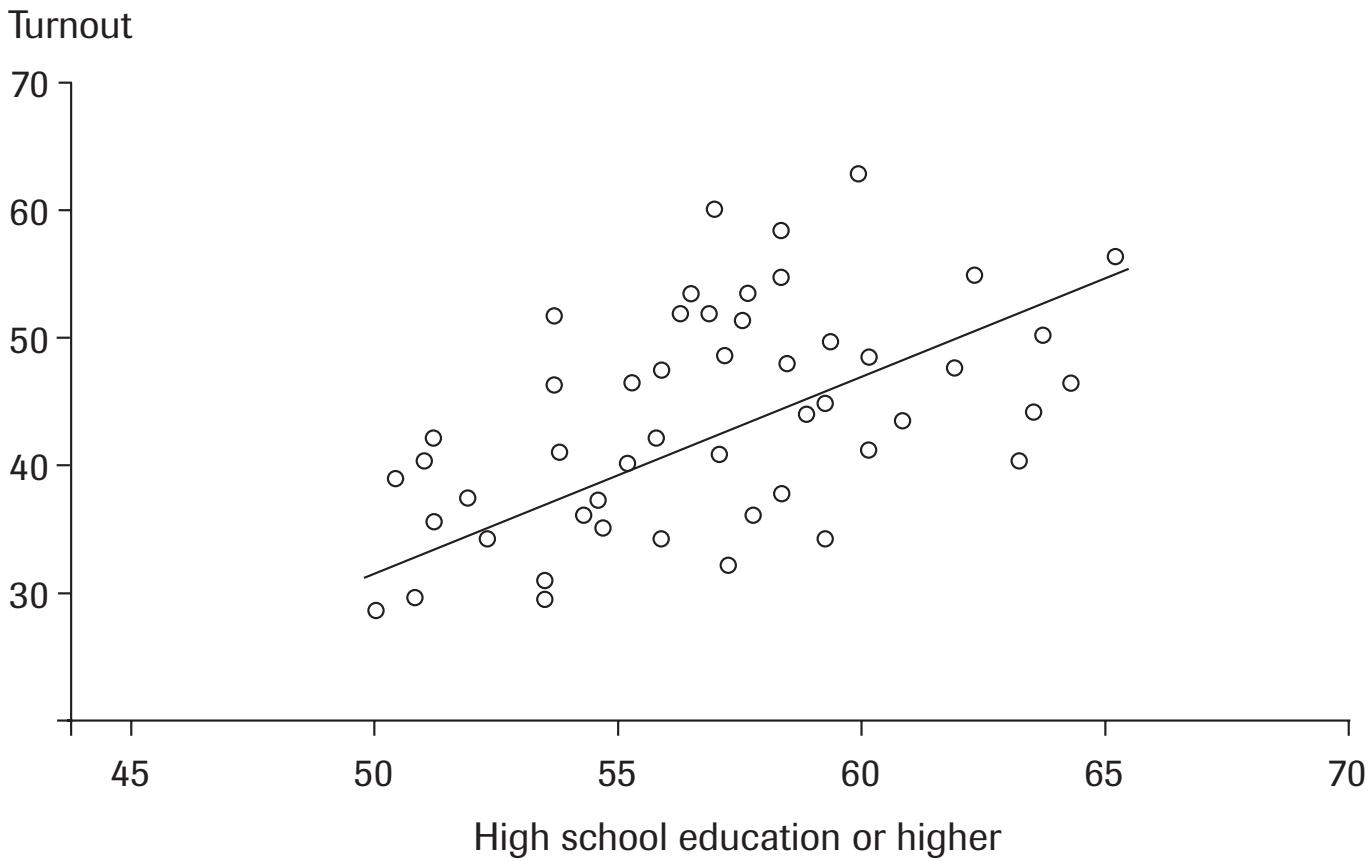
The regression coefficient that emerges minimizes the sum of squared differences ($(y_i - \hat{y})^2$).

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

Figure 8-3 offers a representation of this for our state education and turnout example.

Regression: Education and Turnout

Figure 8-3 Regression: Education and Turnout (in percentages)



Regression: Education and Turnout

This would be our regression line:

$$\hat{y} = -14.53 + 1.02(x)$$

How to interpret this:

- The state in which no one graduated from high school would have a voter turnout of -14.53%.
  - *Center your variables, people. Seriously...*
- Each unit increase in the percentage of the state's citizens having a high school diploma corresponds with a 1.02% increase in voter turnout.

Inference in Regression

What do we say about that b-hat ($\hat{b} = 1.02$?)

- 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 turnous out nicely.

$$t = \frac{1.02}{.25} = 4.08$$

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

- The probability of observing it if β were really zero is .000023.
- 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