Bivariate OLS

EH6105 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Stockholm University

Goal for Today

Use correlation and linear regression to describe the relationship between two continuous 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 learned about random sampling and hypothesis testing.

If our sample statistic is more than 1.96 standard errors from a proposed population parameter, we suggest a population parameter is highly unlikely given what we got.

- This is admittedly an indirect answer to the question you're not asking, but this is what we're doing.

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

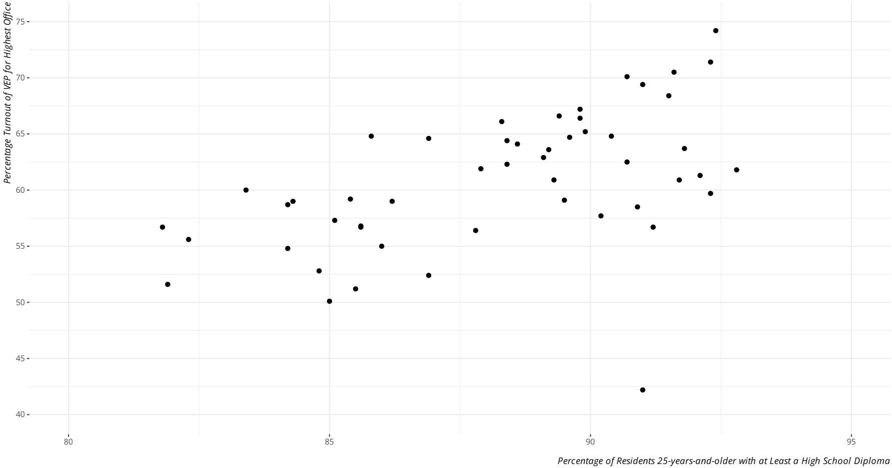
Question: does a state's voter turnout vary by the state's level of education?

- 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**.

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: ?election_turnout in {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*

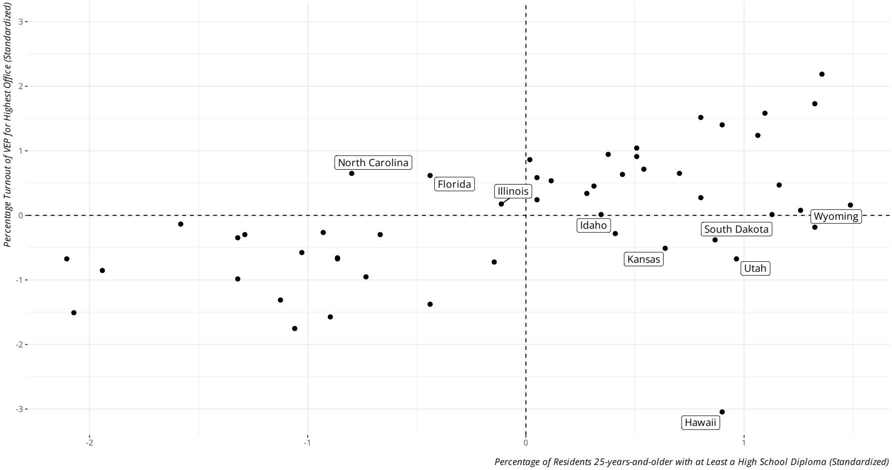
$$\sum \frac{\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

...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.

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: ?election_turnout in {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

election_turnout %>%
  summarize(cor = cor(perhsed, turnoutho)) %>%
  pull()
#> [1] 0.5282739
```

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

```r
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 student who studied 0 hours got a 55.

- The student who studied 1 hour got a 61.
- The student 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.

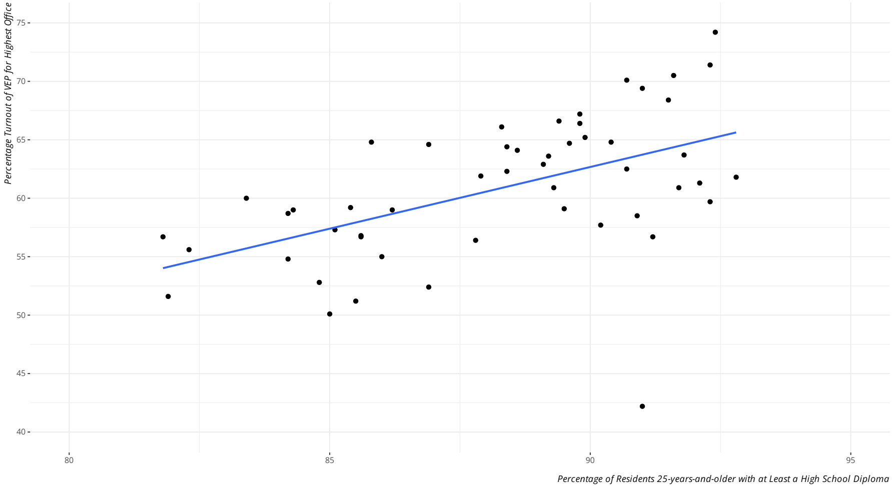
The regression coefficient that emerges minimizes the sum of squared differences ($(y_i - \hat{y})^2$).

- 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.



How You'd Get What You Want in 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
```

On the Output You See

The important stuff:

- "Estimate": y-intercept, and regression coefficients (i.e. "rise over run")
- Standard errors: an estimate of variability around the estimate (coefficient).
- Test statistic stuff (t-statistic, p-value): the stuff you'll use for inference.
- R^2s: measures of how well the model fit the data.

The less important stuff:

- F-statistic: "overall significance" of the model.
- Residual standard error: standard error of the residuals
  - Used for calculating standard errors, in combination with the var-cov matrix (which you don't see).
- Distribution of residuals (at the top): provides a summary of the range of residuals.

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 I hate matrix algebra. :P

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

If You're Curious...

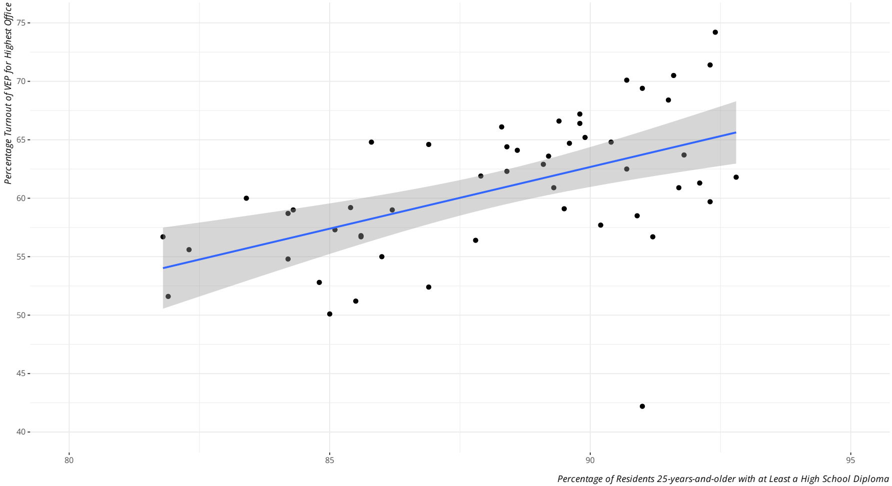
```r
X <- model.matrix(M1) # Intercept + perhsed

# 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%.
    - This is obvious nonsense, which is why you'll want to learn about variable transformations as you progress.
- 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 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 our 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.
  - This is the “null hypothesis” you’ll read about in your travels.

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.”

- This is a fancy (and misleading) way of saying “it’s highly unlikely to be 0.”

Alternatively, in R...

```
# lm() in R is doing this for you, but let's do it ourselves...
# Be mindful there is some rounding for presentation.
broom::tidy(M1)
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)   -32.3     21.4       -1.51 0.138
#> 2 perhsed         1.06     0.242      4.36 0.0000677
broom::tidy(M1) %>% slice(2) -> pershed_info_we_want

# divide the coefficient
pull(pershed_info_we_want[1,2])/
   # over the standard error
  pull(pershed_info_we_want[1,3]) -> t_stat # assign to object

t_stat
#> [1] 4.355235
# two-tail test time
2*pt(t_stat, 49, lower.tail=FALSE) # hi mom!
#> [1] 6.765377e-05
```

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