The Bivariate Linear Model

EH6127 – Quantitative Methods

Steven V. Miller

Department of Economic History and International Relations

Goal for Today

Use correlation and linear regression to describe the relationship between two continuous variables.

Building Toward Normal Social 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**

Correlation

Question: does a country's life expectancy vary by its human capital?

- Human capital (index): how well today can citizen expect to achieve full health and achieve her formal education potential. [0:1]
- Life expectancy: average life expectancy form men and women (in years)
- *Data subset to 2020 for states in the EU (incl. the UK)*

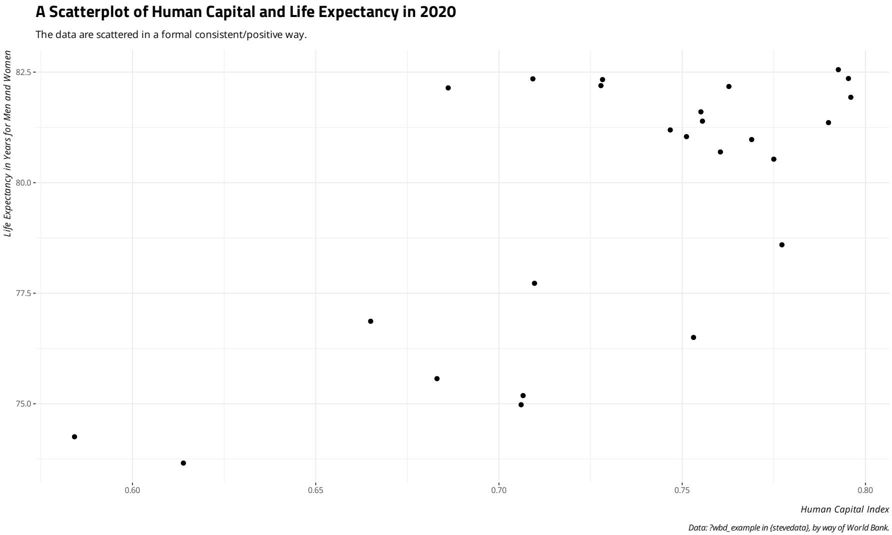
We'll start with a preliminary judgment by way of a scatterplot.

If You'd Like to Follow Along...

```r
# library(stevedata) # has eurostat_codes and wbd_example
# ?eurostat_codes
# ?wbd_example

eurostat_codes %>%
  # filters to any eurostat codes that are/were EU
  filter(cat == "EU" | iso2c == "UK") %>%
  # pulls to a vector, and assigns
  pull(iso2c) -> isocodes_we_want

wbd_example %>%
  na.omit %>% # for ease, given the nature of the data
  # then, filter to just observations in 2020
  filter(year == 2020)  %>%
  # then, filter where isocode matches what we want, and assign to Data
  filter(iso2c %in% isocodes_we_want) -> Data
```


A Scatterplot of Human Capital and Life Expectancy in 2020
The data are scattered in a formal consistent/positive way.
Life Expectancy in Years for Men and Women
82.5
80.0
77.5
75.0
0.60
0.65
0.70
0.75
0.80
Human Capital Index
Data: ?wbd_example in {stevedata}, by way of World Bank.

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

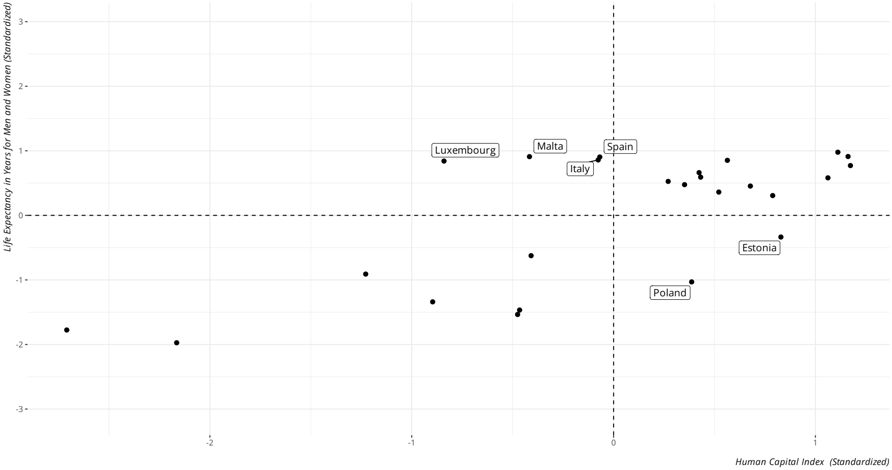
$$\sum \frac{\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

...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 Human Capital and Life Expectancy in 2020

Observations in the negative correlation quadrants are highlighted for emphasis.



Data: ?wbd_example in {stevedata}, by way of World Bank.

Human Capital and Life Expectancy (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 16.91/24, or about .70

- 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 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

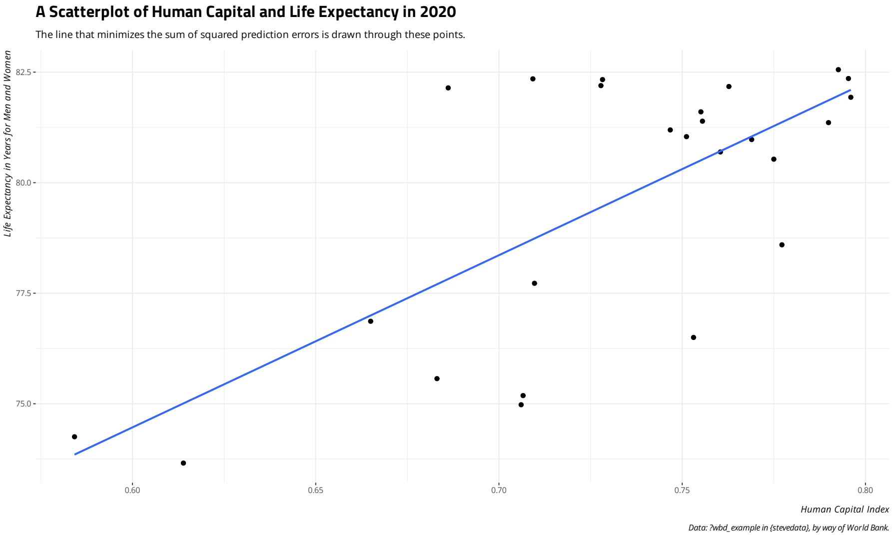
How do we get a regression coefficient for more complicated data?

- 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 example.


A Scatterplot of Human Capital and Life Expectancy in 2020
The line that minimizes the sum of squared prediction errors is drawn through these points.
Life Expectancy in Years for Men and Women
82.5
80.0
77.5
75.0
0.60
0.65
0.70
0.75
0.80
Human Capital Index
Data: ?wbd_example in {stevedata}, by way of World Bank.

How You'd Get What You Want in R

```
summary(M1 <- lm(lifeexp ~ hci, data=Data))
#>
#> Call:
#> lm(formula = lifeexp ~ hci, data = Data)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -3.9279 -1.0137 -0.0183  1.0120  4.3228
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   51.102      6.002   8.515 1.45e-08 ***
#> hci           38.941      8.177   4.762 8.42e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 2.185 on 23 degrees of freedom
#> Multiple R-squared:  0.4965, Adjusted R-squared:  0.4746
#> F-statistic: 22.68 on 1 and 23 DF,  p-value: 8.424e-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.

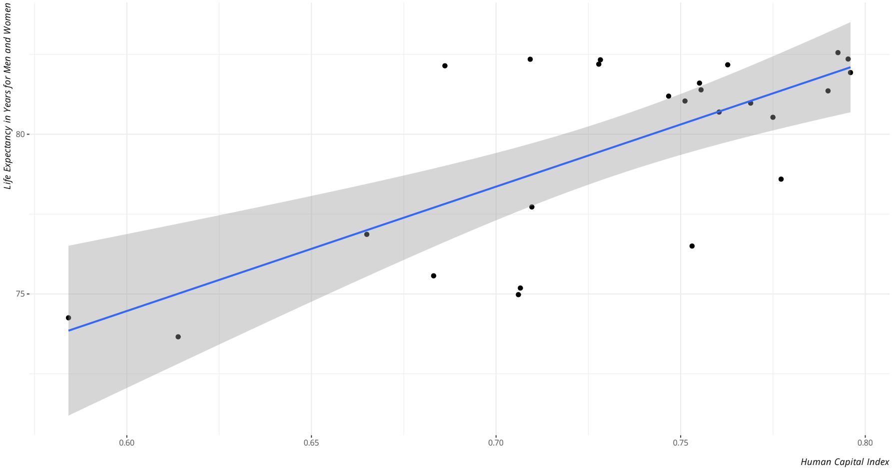
This has a convoluted estimation procedure.

- 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.

A Scatterplot of Human Capital and Life Expectancy in 2020

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



Data: ?wbd_example in {stevedata}, by way of World Bank.

Regression: Human Capital and Life Expectancy

This would be our regression line:

$$\hat{y} = 51.102 + 38.941(x)$$

How to interpret this:

- The country for which human capital is 0 would have an average life expectancy of 51.102.
    - This would never be observed, but it's at least a plausible quantity.
- A one-unit increase in human capital corresponds with an estimate increase in life expectancy of 38.941 years.
    - Given this variable's scale, this is incidentally a min-max effect.

Inference in Regression

What do we say about that *b*-hat ($\hat{b}$ = 38.941?)

- 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 example, this turns out nicely.

$$t = \frac{38.941}{8.177} = 4.762$$

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

- The probability of observing it if β were really zero is 0.0000842

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.”

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

Introduction

Correlation

Linear Regression

- Demystifying Regression
- A Simple Example
- Getting a Regression Coefficient
- Inference in Regression

Conclusion