3. Regression Fundamentals
Part 1: Simple Linear Regression

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Describing data

- Quantitative vs. categorical variables; discrete vs. continuous
- Histograms and densities for continuous variables
- Measures of central tendency (mean, median), location (percentiles), variability (variance, standard deviation)

Inferential statistics

- The importance of *quantifying uncertainty*: confidence intervals and significance testing
- Sampling distributions: what you would expect to see from an estimator (like $\bar{x}$) over *repeated sampling*.
- *Standard error*: a measure of variability in the sampling distribution.

Tonight

Describing the relationship between two variables:

- Scatterplots
- Covariance and correlation
- Linear regression

Tonight's sample datasets

We will refer to two datasets tonight (both found on Github):

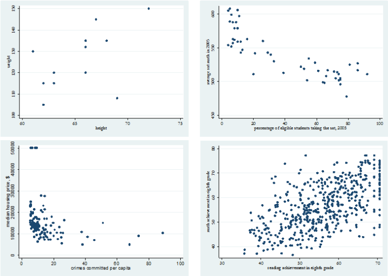
1. `caschool.dta`: data on test performance, school characteristics, and student demographics for California school districts, 1998-99 (N=420). From Stock & Watson text.
2. `TN-lettergrades-2022-23.dta`: letter grades and component scores for Tennessee schools, 2022-23 (N=1,900)

Source of TN data: `https://www.tn.gov/education/districts/federal-programs-and-oversight/data/data-downloads.html`

Scatterplots

Scatterplots

The easiest way to see how two variables are related is a **scatter diagram** or **scatterplot**. In Stata: `scatter` *yvar xvar*



Scatterplots

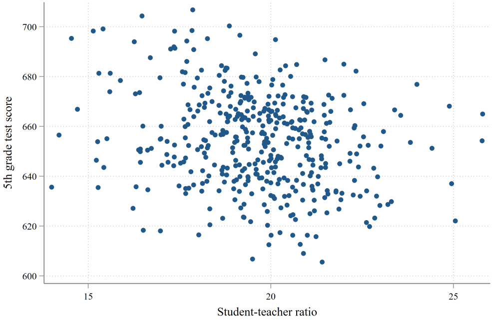
Scatterplots can provide a sense of the *direction* of relationship (if any), *linearity*, and *strength* of association.

Often with scatterplots there are natural **outcome** and **explanatory** variables. We may have in mind a theory in which variation in the outcome is at least in part explained by variation in the explanatory variable.

- Denote the outcome as Y and explanatory variable as X.
- These are also called **dependent** (Y) and **independent** (X) variables, although I avoid these terms, since they have other meanings in statistics.

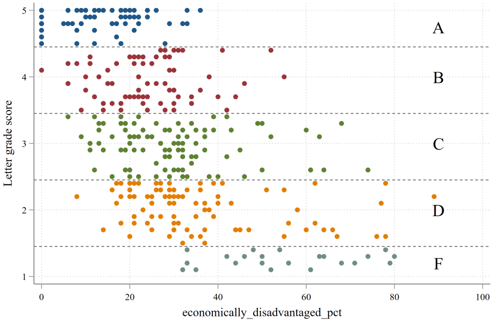
Example 1

From the *caschool* data (California school districts in 1998-99): 5th grade test scores vs. student-teacher ratio



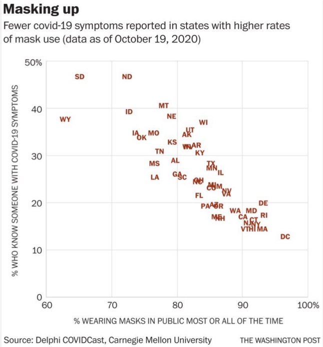
Example 2

From the *TN-lettergrades* data: overall scores vs. percent economically disadvantaged, Tennessee high schools in 2022-23.



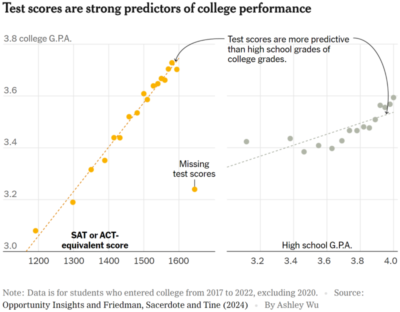
Example 3

COVID symptom reporting and mask-wearing:



Example 4

Predictors of college performance (SAT/ACT vs. high school GPA)



Source: David Leonhardt, "The Misguided War on the SAT," *The New York Times*, January 7, 2024. Note: these are binned, not raw, data. The data come from highly selective universities.

Covariance and correlation

Covariance

A picture can be worth a thousand words, but we might like a summary measure of how two variables are associated.

The **sample covariance** between two variables x and y is:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

The covariance is an average, where—for each observation—we multiply x's deviation from the mean of x by y's deviation from the mean of y.

- If y tends to be higher than average when x is higher than average, these products will tend to be positive (a **positive covariance**).

- If y tends to be *lower* than average when x is higher than average, these products will tend to be negative (a **negative covariance**).

Like the variance, units of covariance are not easily interpreted.

Correlation

The **sample correlation coefficient** is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n-1)} = \frac{s_{xy}}{s_x s_y}$$

Correlation is a standardized or unit-free measure of association:

- It ranges between -1 and +1
  - $r_{xy} = +1$ is a **perfect positive correlation**
  - $r_{xy} = -1$ is a **perfect negative correlation**
  - $r_{xy} = 0$ is **no correlation**
- r is a measure of *linear* association—it is not appropriate for use with non-linear relationships.

Covariance and correlation in Stata

To obtain correlation coefficients in Stata, use `corr` *yvar xvar*. You can include a list of variables in this command.

- Be aware of how Stata handles missing values:
  - **listwise deletion** means observations are not used if *any* of the listed variables in the command are missing.
  - **pairwise deletion** means correlations of pairs of variables are considered in isolation.
- `pwcorr` *yvar xvar* uses pairwise deletion. `corr` uses listwise deletion.

To obtain the covariance in Stata, use `corr` with `cov` option. The result is called a **variance-covariance matrix**. This is used less often.

Exercise: California school district data

Open the *caschool* data and do the following:

1. Create scatterplots between district average 5th grade test scores (*testscr*) and:
   - The percent of low-income students (*meal_pct*)
   - The percent of English language learners (*el_pct*)
   - Expenditures per student (*expn_stu*)
2. Calculate the correlation and covariance between each of the above pairs of variables.

How would you describe the association between these pairs of variables? Positive/negative? Strong/weak? Linear/non-linear?

Strength of correlation

What is a "strong correlation?" It depends on the context. (How strong would you expect the correlation to be? Is there a theoretical reason why the correlation should be particularly strong or weak?)

Rule of thumb ("Cohen's scale") based on the absolute value $|r_{xy}|$:

- $|r_{xy}| < 0.1$: zero to weak correlation
- $0.1 < |r_{xy}| < 0.3$: weak to moderate correlation
- $0.3 < |r_{xy}| < 0.5$: moderately strong correlation
- $|r_{xy}| > 0.5$: strong correlation

Try "guess the correlation:"
`https://istats.shinyapps.io/guesscorr/`

Correlation vs. causation

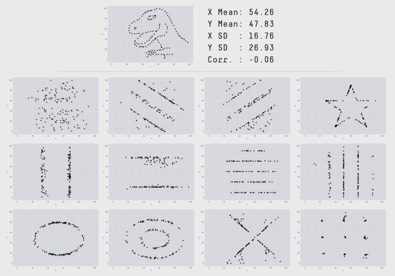
Important: *correlation does not imply causation!*

- *Correlation* means two variables move together.
- *Causation* means that change in one variable is causing change in the other.

For fun, check out this collection of spurious correlations:
`https://tylervigen.com/spurious-correlations`

The importance of visualizing your data

Never trust summary statistics alone! All of the datasets used below have the same $\bar{x}, \bar{y}, s_x, s_y$, and r_{xy}.



Source:
`https://www.autodesk.com/research/publications/same-stats-different-graphs`

Linear regression

Regression

Another way to quantify the relationship between two variables Y and X:

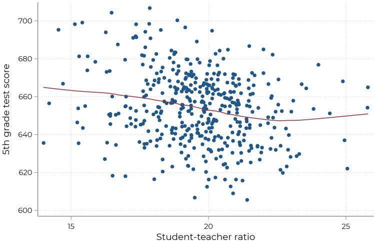
- *How much* does Y change when X changes by one unit?
- How does the *average* level of Y vary with X?

Why might you want to know this?

- **Description**: it's a useful way to describe how variables are related.
- **Prediction**: if you know X, what is your best prediction of Y? Examples: SAT and GPA; class size and test scores.
- **Causal inference**: in some cases, this relationship describes the *causal* effect of X on Y. Example: class size.

Fitting lines using loess

This graph fits a **loess curve** (“locally estimated scatterplot smoothing”) to the *caschool* data. In Stata: `lowess` *yvar xvar*.



This calculates the mean of Y “locally”—at small intervals around each X.

Simple linear regression

This graph is nice, but it’s hard to describe the relationship any other way but visually. What if we could approximate it with a *line*? A line is defined entirely by its *slope* (b) and *intercept* (a):

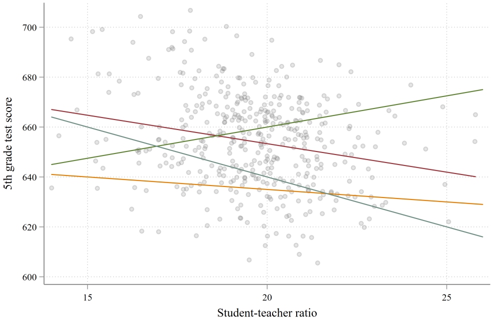
$$y = a + bx$$

The slope of a line tells you how much y changes when x changes by one unit $\Delta Y / \Delta X$.

Simple linear regression finds the **line of best fit**.

Finding the best fit line

What makes a particular line the “best fit”? There are many possibilities, with different values of *a* and *b*. Which is the “best”?



Line of best fit

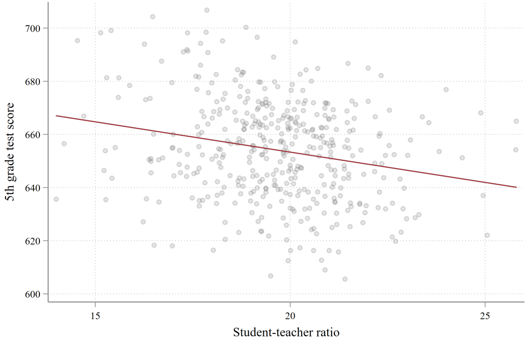
Stata can provide a line of best fit, using two overlaying graphs (`scatter` and `lfit`). Using the *caschool* data:

```
twoway (scatter testscr str) (lfit testscr str)
```

It is conventional to put the outcome variable on the vertical (y) axis, and the explanatory variable on the horizontal (x) axis.

Line of best fit

Overlaid graphs `scatter` and `lfit`:



Line of best fit

In this case, the best fit line has an intercept of 698.93, and a slope of -2.28: $\widehat{\mathbf{y}} = \mathbf{698.93} - \mathbf{2.28x}$

- The best fit line is also called a **prediction equation**
- $\hat{y}$ is the **predicted value** for y, given a value of x.

We can use the prediction equation to predict y for a specific x value (here, the student-teacher ratio):

- Example: suppose $x = 20$ students
- $\hat{y} = 698.93 - (2.28 * 20) = 653.33$
- Predicted 5th grade test score is 653.03 for a class size of 20

Line of best fit

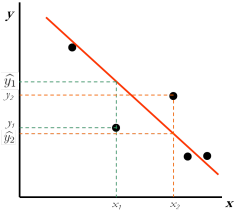
Interpreting the prediction equation: $\widehat{\mathbf{y}} = \mathbf{698.93} - \mathbf{2.28x}$

- 698.93: the predicted 5th grade test score when student-teacher ratio is $x = 0$
- -2.28: the predicted *change* in 5th grade test scores when the student-teacher ratio increases by one year.
- Note *5* additional students per teacher would be predicted to change test scores by: $-2.28 * 5 = -11.4$

Least squares

How does one determine the line of "best fit"? For a given line, we have a set of predictions for y, one for every value of x in the data:

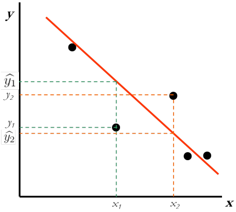
- $\hat{y}_1$ is the predicted value of y when x is x_1
- $\hat{y}_2$ is the predicted value of y when x is x_2
- ...and so on, up to $\hat{y}_n$



Least squares

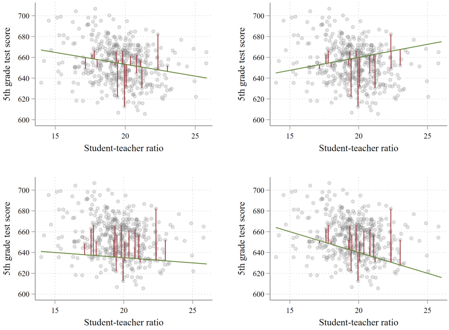
For a given line, we have a **residual** (or **prediction error**) for every value of x in the data: $\hat{u} = y - \hat{y}$:

- $\hat{u}_1$ is the residual when x is x_1
- $\hat{u}_2$ is the residual when x is x_2
- ...and so on, up to $\hat{u}_n$



Least squares

Four candidate lines for the 5th grade test score data:



Least squares

The line that minimizes the *sum of the squared residuals* between the data points y and the line $\hat{y}$ is the **least squares** or **ordinary least squares (OLS)** regression line:

$$\min_{a,b} \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2$$

$$\min_{a,b} \sum_{i=1}^{n} \left(y_i - a - bx\right)^2$$

i.e. choose intercept and slope (a, b) to minimize the sum of the squared residuals $(\hat{u}_i = y_i - \widehat{y_i})$

Least squares

It can be shown that the least squares slope (b) and intercept (a) are as follows:

$$b = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

Least squares

It is easy to show that the slope b can also be written:

$$b = r_{xy}\frac{s_y}{s_x}$$

The slope coefficient has the same sign $(+/-)$ as the correlation coefficient r_{xy} (re: s_y and s_x are both greater than zero)

Regression in practice

To compute the least squares intercept and slope coefficient in Stata use `regress` or `reg` *yvar xvar* (aka "running a regression"). Example using *caschool* data:

```
. reg testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

$$\hat{y} = 698.93 - 2.28x$$

Note: "_cons" refers to the intercept, or **constant term**.

Interpreting coefficients

Steps for interpreting a regression slope coefficient—general guidelines:

1. Identify the **explanatory** variable and its units (e.g., height in inches, students per teacher).
2. Describe a **one-unit increase** in the explanatory variable in everyday language (e.g., one additional student per teacher).
3. Identify the **outcome** variable and its units (e.g., weight in pounds, 5th grade test scores).
4. Describe the coefficient as the **change in the outcome** predicted for a one-unit change in the explanatory variable (e.g., an additional student per teacher is predicted to decrease 5th grade test scores by 2.28 points).

Note: be sure your interpretation reflects the appropriate unit of observation (e.g., individual, school, district).

Note: Adapted from Remler & Van Ryzin (2011), chapter 8.

Predicted values and residuals

It is possible to have Stata compute the **predicted values** and **residuals** (prediction errors) for each observation after `reg`:

- `predict` *yhat*`, xb`
- `predict` *uhat*`, resid`

These are called **postestimation** commands in Stata:

- `xb` refers to the predicted value ($\hat{y}$)
- `resid` refers to the residual ($\hat{u}$, calculated as $y_i - \widehat{y_i}$)

Measuring fit

How well does the regression line fit the data?

- Mechanically, the least squares intercept and slope can be calculated for any set of data points (x, y).
- The line of best fit (OLS) is not necessarily a *good* fit.
- Least squares minimizes the sum of the squared residuals, but performs better with some data than others.

R^2, the **coefficient of determination**, is a measure of the goodness of fit.

R^2

R^2 is the proportion of the total variation in y from its mean that is "explained" (predicted) by x.

The total variation in y around its mean is the **total sum of squares (TSS)**:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Using the predicted y instead of the actual, the **explained sum of squares (ESS)** is:

$$ESS = \sum_{i=1}^{n} (\widehat{y_i} - \bar{y})^2$$

R^2

The R^2 is therefore:

$$R^2 = \frac{ESS}{TSS}$$

R^2 is **always between 0 and 1**

The explained sum of squares (ESS) is sometimes called the "model" sum of squares (see Stata output).

R^2

The "unexplained" variation in y is the **sum of squared residuals (SSR)** (a.k.a. the error sum of squares):

$$SSR = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

It makes sense that the R^2 should be related to the SSR, which we aim to minimize when finding the best fit line. In fact, we can write R^2 as:

$$R^2 = 1 - \frac{SSR}{TSS}$$

R^2 for class size regression

Example using 5th grade test scores and student-teacher ratio:

```
. reg testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

$$R^2 = \frac{ESS}{TSS} = 0.0512$$

Mean squared error

A related measure is the **mean squared error (MSE)**:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n-2} = \frac{\sum_{i=1}^{n} \widehat{u_i}^2}{n-2} = \frac{SSR}{n-2}$$

The MSE is the *average* squared deviation of the predicted y from the actual y (uses $n-2$ in the denominator). Note the numerator is the residual sum of squares (SSR).

Note: least squares minimizes SSR so it also minimizes MSE

Standard error of the regression

The square root of the MSE is the **standard error of the regression (SER)** a.k.a. root mean squared error (RMSE):

$$SER = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\widehat{u}_i^2}{n-2}} = \sqrt{\frac{RSS}{n-2}}$$

Just as the standard deviation can be interpreted (intuitively, not literally) as the average deviation of y from its mean, the SER can be interpreted (intuitively, not literally) as the average deviation of y from its *predicted value*. I.e., "how close," on average, is your prediction?

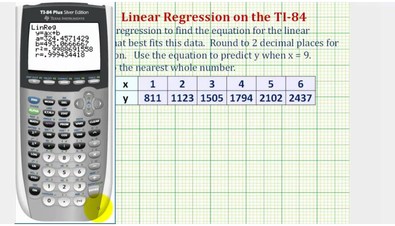
Exercise: Tennessee letter grade data

Open the *TN-lettergrades* data and do the following:

1. Create a scatterplot and best fit line (`lfit`) between the letter grade score (*lg_score*) and the percent economically disadvantaged.
2. Calculate the least squares slope and intercept for the best fit line, and interpret.
3. Interpret the R^2 from the above regression line.
4. Interpret the SER (aka RMSE) from the above regression line.
5. Have Stata save the predicted values and residuals from the above regression. You can view these using`browse`.

Moving beyond best fit lines

Finding “best-fit” lines is easy—you could do this all day. Even your TI-84 calculator can do it!



Can we take things a little further?

Conditional mean interpretation

Suppose we are willing to assume that the relationship between the mean of y in the population is related in a **linear** way to x. That is, the **conditional mean** of y given x is:

$$E(y|x) = \beta_0 + \beta_1 x$$

This is called the **population regression function**. I am switching from a and b to β_0 and β_1 because they are now unknown population parameters to be estimated. One can use a and b as *estimators* of β_0 and β_1.

Note $E()$ refers to the expectation or population mean.

Inferences about β_1

If we are using a sample of y and x to *estimate* the population intercept and slope coefficients in the population, we need to quantify our uncertainty in the same way we did for $\bar{x}$:

- Confidence intervals for β_1
- Hypothesis tests about β_1

To do so we need to know the *sampling distribution* of the slope estimator (b). As with $\bar{x}$, this will require some assumptions (next week).

Next time

Inferences about β_1

- Read: Stock & Watson chapter 4 (sections 4.4-4.5) and 5
- Familiarize yourself with the three sample articles: Magnuson et al. (2004), Gershenson & Holt (2015), and Reber & Smith (2023).