Logistics

↪ PSet 1 – grades released today

- I posted solutions
- Everyone did well
- Some people were not able to correctly answer question 5 – please ask Lily / me if you still have questions

↪ PSet 2 - due next Monday

- Slightly longer than PSet 1
- I encourage you to start early

↪ Topics slightly reorganized – see canvas

Recap

We have learned how to summarize **one variable:**

↪ Typical values: mean, median, mode

↪ Dispersion: variance, standard deviation, IQR

Topic 3: Relationships in Data

Correlation

Now, we are going to look at ways to summarize relationships between multiple variables.

↪ More specifically: relationships between *two* variables.

↪ We'll discuss other multivariate relationships later.

Correlation

Correlation is the term we use to describe the extent to which two featured of the world (variables) occur together.

↪ Two variables are **positively correlated** if they occur together.

↪ Two variables are **negatively correlated** if one occurs when the other does not and vice versa.

↪ Two variables are **uncorrelated** if there is no apparent relationship between when they occur.

Some people like to use the phrase **association** instead of correlation. For our purposes, these are synonyms.

Correlation

Why do we want to know if two variables are correlated?

Three main uses:

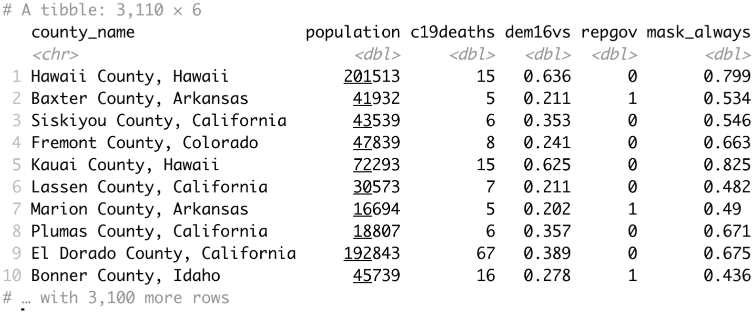
1. **Description**: e.g., is there a racial disparity in arrest rates in the United States?
2. **Forecasting or prediction**: e.g., can we use polling data to predict outcomes in the 2022 midterm elections?
3. **Causal inference**: e.g., are people less likely to contract Covid-19 after a dose of an mRNA vaccine?

There is a lot to say about how correlations help us with these tasks. It's not easy. But before all that, a more basic question: *does a relationship even exist?*

Correlation

To answer this: we'll visualize and measure them.

I will show you using an example dataset from a NY Times survey on mask wearing. Here's a snippet:

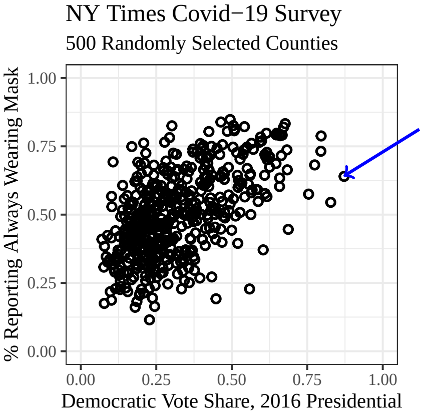


```
# A tibble: 3,110 × 6
   county_name                   population c19deaths dem16vs repgov mask_always
   <chr>                              <dbl>     <dbl>   <dbl>  <dbl>       <dbl>
 1 Hawaii County, Hawaii             201513        15   0.636      0       0.799
 2 Baxter County, Arkansas            41932         5   0.211      1       0.534
 3 Siskiyou County, California        43539         6   0.353      0       0.546
 4 Fremont County, Colorado           47839         8   0.241      0       0.663
 5 Kauai County, Hawaii               72293        15   0.625      0       0.825
 6 Lassen County, California          30573         7   0.211      0       0.482
 7 Marion County, Arkansas            16694         5   0.202      1       0.49
 8 Plumas County, California          18807         6   0.357      0       0.671
 9 El Dorado County, California      192843        67   0.389      0       0.675
10 Bonner County, Idaho               45739        16   0.278      1       0.436
# … with 3,100 more rows
```

Source: https://github.com/nytimes/covid-19-data/tree/master/mask-use

Visualizing Relationships: Graphically

On a **scatter plot**, each point is a unit (row) in the dataset, and its location tells you the values of two variables for that unit.

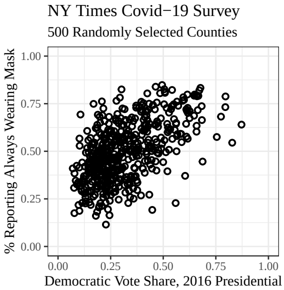


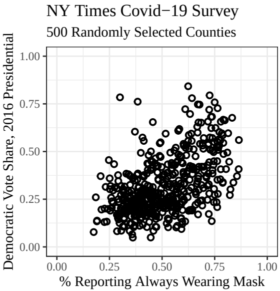
Clayton County, GA
(0.842, 0.624)

Visualizing Relationships: Graphically

Choosing your axes in a scatter plot is important.

↪ What goes on the *x* and *y* axes determines how you interpret.
But: it does not change the distribution or correlation!

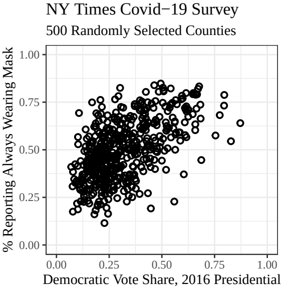


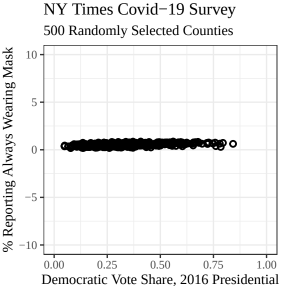


Visualizing Relationships: Graphically

Choosing your axes in a scatter plot is important.

↪ Choosing how you scale the *x* and *y* axis doesn't change the distribution or correlation, but it can mislead people!





Visualizing Relationships: Numerically

A **crosstab** or **contingency table** is a way to represent frequencies across two variables.

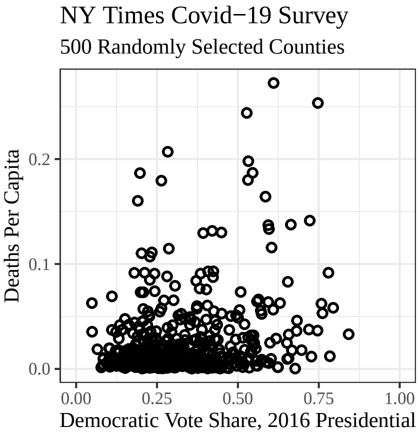
↪ This is a numerical version of a histogram over two variables.

For example: counties by vote choice and deaths/capita.

	Voted Clinton	Voted Trump	Total
1st Quartile Deaths/Capita	70	708	778
2nd Quartile Deaths/Capita	69	708	777
3rd Quartile Deaths/Capita	112	666	777
4th Quartile Deaths/Capita	236	541	778
Total	487	2623	3110

(Is this a good way to visualize this bivariate distribution?)

Visualizing Relationships: Numerically



Visualizing Relationships: Numerically

Consider instead: number of countries by whether they're oil producers and whether they're democracies.

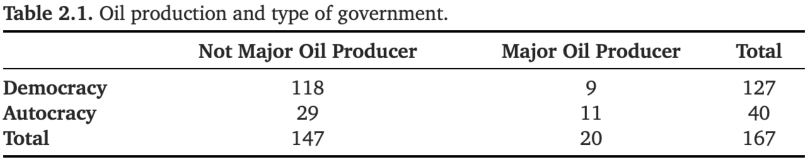


Table 2.1. Oil production and type of government.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

Source: BdM and Fowler (2021), p. 14.

This is a great example of a crosstab because it's so easy to see the relationship in the bivariate distribution.

Covariance and Correlation Coefficients

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

$$\mathrm{Cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

Notice that covariance will be...

↪ ... positive when both variables get bigger at the same time or smaller at the same time.

↪ ... negative when one variable gets bigger while the other gets smaller.

↪ ... closer to zero when x and y are relatively uncorrelated.

↪ ... sensitive to the measurement units of x and y.

Covariance and Correlation Coefficients

Consider a fake dataset. How are these variables correlated?

x	y
-1	8
1	13
2	10
3	8
7	2

Covariance and Correlation Coefficients

Covariance tells us if and how variables are correlated, but you can't really compare covariance across different distributions.

Instead, we **normalize** covariance by dividing by the product of the standard deviations:

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{s_x s_y}$$

This gives us the **correlation coefficient** for x and y.

$\hookrightarrow$ Sometimes people write this as $\mathrm{corr}(x, y)$.

Correlation coefficients are always between –1 and 1. Now we can compare correlations across different distributions!

Covariance and Correlation Coefficients

Consider a fake dataset. How are these variables correlated?

x	y
-1	8
1	13
2	10
3	8
7	2

Conditional Mean and Regression

Both covariance and the correlation coefficient tell us how strong a correlation is, but it does not tell us the relationship in a *substantively* meaningful way.

To do that, we can go back to thinking about means. (Why?)

However, since we're looking at relationships between two variables, we will need to calculate **conditional sample means**.

One way to mathematically denote the **sample mean of y conditional on x** is $\overline{y}(x)$.

Notice this is a function! So, it simply needs to tell us the mean of y subsetting to units taking each value of x.

Conditional Mean and Regression

Recall the data from BdM and Fowler:

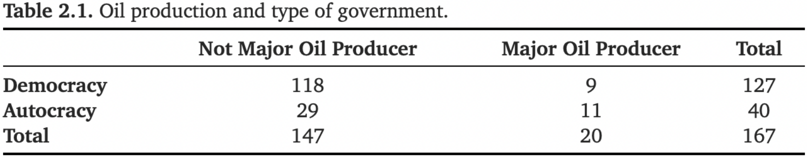


Table 2.1. Oil production and type of government.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

Source: BdM and Fowler (2021), p. 14.

Let's use some math notation: x_d is 1 if a country is a democracy and 0 otherwise, x_o is 1 if a country is a major oil producer and 0 otherwise. Then:

$$\overline{x}_d(x_o) = \begin{cases} 9/20 & \text{if } x_o = 1 \\ 118/147 & \text{if } x_o = 0 \end{cases} \quad \overline{x}_o(x_d) = \begin{cases} 9/27 & \text{if } x_d = 1 \\ 11/40 & \text{if } x_d = 0 \end{cases}$$

Conditional Mean and Regression

Those conditional means were really easy to calculate because the two variables were binary.

What if we have two continuous variables?

↪ Example from Covid-19 data: Democratic vote share (between 0 and 1), and % always wearing mask (between 0 and 1)

In theory, we can still write a function that will give the conditional sample mean, but in practice, it's hard:

↪ There may not be enough data at every possible x to calculate $\overline{y}(x)$, so we have to approximate.

↪ Since we necessarily have to approximate $\overline{y}(x)$, we need to make a "theoretical" choice about *how* to approximate.

Conditional Mean and Regression

The most common way to approximate $\overline{y}(x)$ is to assume a linear relationship between x and y.

Recall from algebra the formula for a line: $y = a + bx$.

So, we "fit" a **regression line** that is defined by two **coefficients**:

↪ a **slope** that captures its "steepness" (also called β)

↪ an **intercept** that captures "how high up" the line is (also called α)

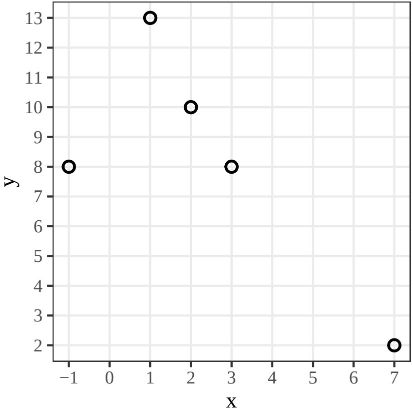
↪ Often written as $\overline{y}(x) = \alpha + \beta x$

How do we do this?

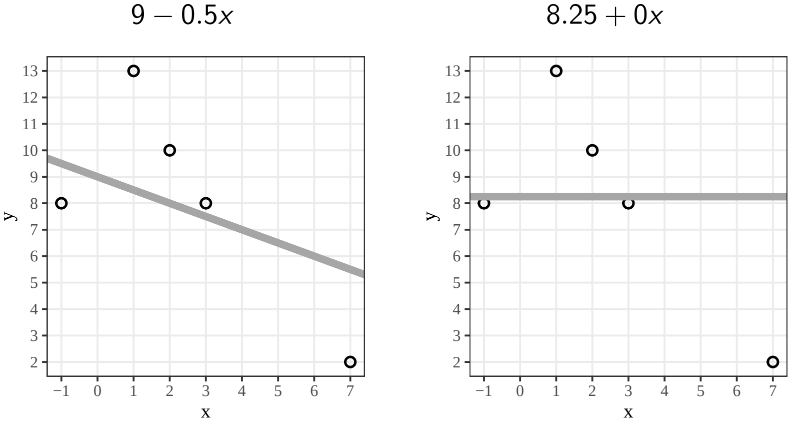
↪ Even though we've committed to a linear relationship, it's still not obvious *which* line to choose!

Conditional Mean and Regression

x	y
−1	8
1	13
2	10
3	8
7	2



Conditional Mean and Regression



$9 - 0.5x$

$8.25 + 0x$

Conditional Mean and Regression

So what do we do?

We pick the line that is the “best” fit, whose coefficients can be calculated this way:

$$b = \frac{r_{xy} s_y}{s_x} = \frac{\mathrm{Cov}(x, y)}{s_x^2} \qquad\qquad a = \overline{y} - b\overline{x}$$

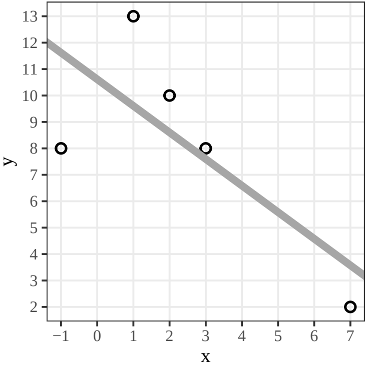
A line calculated this way is often called a **line of best fit** since it its coefficients create a line that is the best fit!

(More on why this is a best fit in a minute.)

We also call this: **OLS regression** or **linear regression**.

Conditional Mean and Regression

$$\overline{y}(x) = 10.61 - 1.01x$$



Conditional Mean and Regression

Recall that we had to be clear about what our "best" guess was when we talked about sample means and medians.

Why is the linear regression line the "best" guess for a linear relationship between x and y?

Consider the errors (or **residuals**) produced by "best" guess:

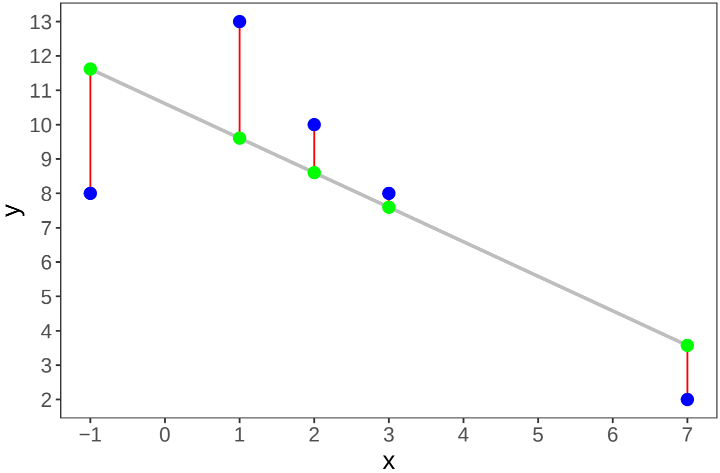
$$e_i = y_i - \overline{y}(x) = y_i - a - bx_i$$

It turns out that when we calculate a and b using the formulas above, it minimizes the mean square error of our best guess!

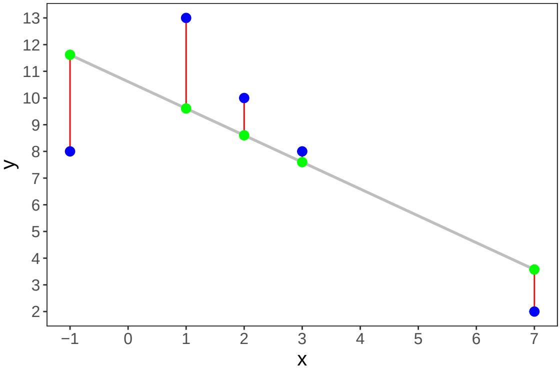
$$MSE = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - (a - bx_i))^2$$

Conditional Mean and Regression

$$\overline{y}(x) = 10.61 - 1.01x$$



Conditional Mean and Regression



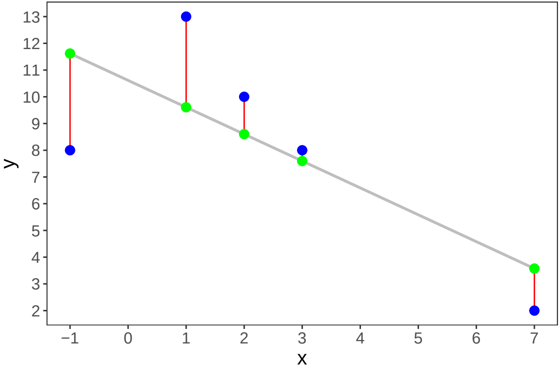
↪ $\overline{y}(x) = 10.61 - 1.01x$

↪ Blue dots: Actual values y_i

↪ Green dots: Predicted values $\hat{y}_i$ (or $\overline{y}(x)$)

↪ Red lines: Residuals $y_i - \hat{y}_i = y_i - (10.61 - 1.01x_i)$

Conditional Mean and Regression



The grey line minimizes the sum of the squared residuals.

↪ I.e. it minimizes the sum of the squared red lines.

↪ I.e. it minimize: $(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_5 - \hat{y}_5)^2$

↪ Which can be written as $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

Conditional Mean and Regression

What does the linear regression line tell us about correlation?

The slope has all the clues:

- ↪ If the slope is positive (upward sloping), then there is a positive correlation.
- ↪ If the slope is negative (downward sloping), then there is a negative correlation.
- ↪ The steepness/flatness of the line is the strength of the correlation (flat means uncorrelated).

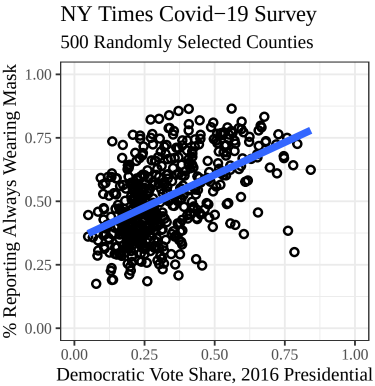
But there's more! The slope also gives us a substantive way to talk about a relationship between the two variables.

$$\bar{y}(x) = 0.35 + 0.51x$$

How to interpret this?

"A one percentage point increase in 2016 Democratic vote share is associated with a 0.51 percentage point increase in share of respondents always wearing masks."

(We rarely interpret a.)

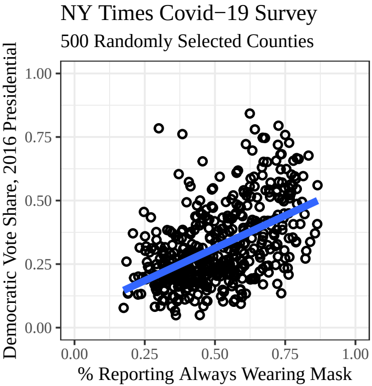


$\bar{y}(x) = 0.06 + 0.51x$

How to interpret this?

"A one percentage point increase in share of residents always masking is associated with a 0.51 percentage point increase in Democratic vote share."

It's a coincidence that b is same!!



Correlation is Linear

When you use one of the three main measures of correlation, you are *assuming* a **linear relationship** between two variables.

↪ This is obvious with linear regression, since it gives us a line.

↪ But it's true for covariance and correlation too since these report one number as a measure of the relationship.

↪ The number gives us two pieces of information—its sign and its magnitude—the ingredients for the slope of a line!

↪ (Technically regression lines give us a little more information: the intercept, but this isn't useful for measuring a correlation.)

↪ And of course, the reason why a linear regression tells us about correlation is that correlation is linear.

Correlation is Linear

A bit more subtly: we can really only talk about correlation between two cardinal variables.

↪ Correlation tells us direction (sign) and strength (magnitude).

↪ Ordinal variables have no magnitude.

↪ Categorical variables have no magnitude or direction.

With some creative transformations, we can look at correlations of non-cardinal variables.

↪ To do this, we rely on the fact that dummy variables convert binary categorical variables into binary cardinal variables.

↪ Example: is a judge a Democratic or Republican appointee?

Correlation is Linear

Civil rights cases assigned to a judge in Eastern District of California, 1995–2016

	Case Settled	Case Didn't Settle	Total
Republican Appointee	2,476	7,843	10,319
Democratic Appointee	1,619	4,152	5,771
Total	4,095	11,995	16,090

Converting to dummies: (1) is the judge a Republican appointee? and (2) did the case settle?

Correlation is Linear

Civil rights cases assigned to a judge in Eastern District of California, 1995–2016

	Case Settled	Case Didn't Settle	Total
Republican Appointee	2,476	7,843	10,319
Democratic Appointee	1,619	4,152	5,771
Total	4,095	11,995	16,090

Converting to dummies: (1) is the judge a Republican appointee? and (2) did the case settle?

↪ Regression line: $\overline{\mathrm{settled}}(\mathrm{republican}) = 0.28 - 0.04 \cdot \mathrm{republican}$.

↪ Interpreting: in this dataset, a one unit increase in $\mathrm{republican}$ is associated with a 0.04 unit decrease in $\mathrm{settled}$.

↪ Or: having a Republican instead of a Democrat is associated with a 0.04% decrease in the probability of settlement.

Correlation is Linear

Civil rights cases assigned to a judge in Eastern District of California, 1995–2016

	Case Settled	Case Didn't Settle	Total
Republican Appointee	2,476	7,843	10,319
Democratic Appointee	1,619	4,152	5,771
Total	4,095	11,995	16,090

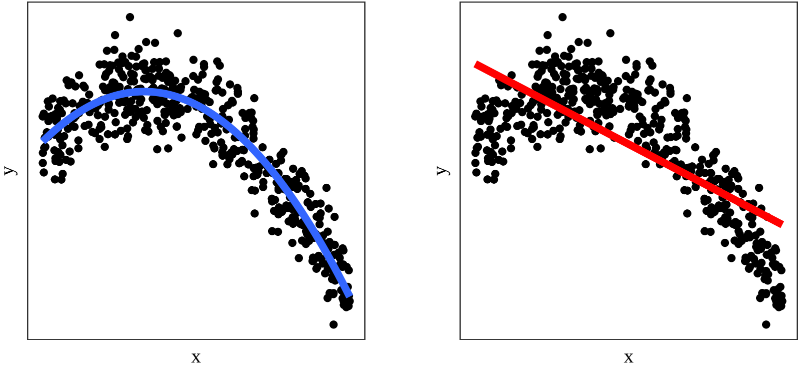
Converting to dummies: (1) is the judge a Republican appointee? and (2) did the case settle?

But: what about categorical variables with more than two categories? Then, things are more complicated. (More below.)

In this example, the settled variable came from another variable with many categories, and I just collapsed those other categories into "not settled" to make it binary.

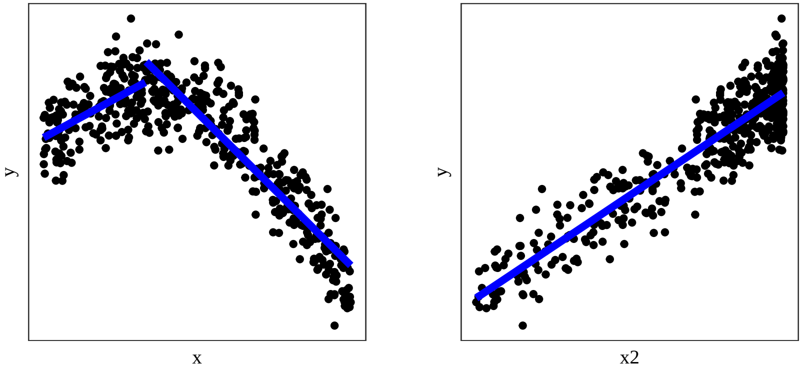
Correlation is Linear

Even when we have two cardinal variables: obviously not all relationships between two variables is linear. Some fake data:



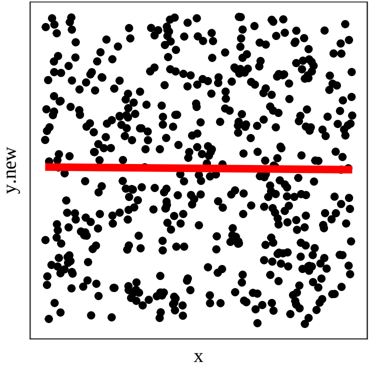
Correlation is Linear

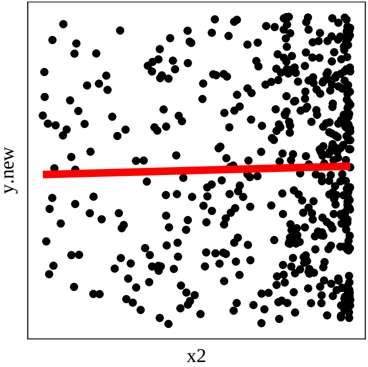
Even when we have two cardinal variables: obviously not all relationships between two variables is linear. Some fake data:



Correlation is Linear

But transforming is not magic! You have to transform correctly.





Correlation is Linear

The lesson: think, examine, then think some more!

↪ Think: Do I have two cardinal variables and do I want to measure a linear relationship between them?

↪ Examine: Plot the data. Is the relationship actually linear?

↪ Examine: If yes, then compute some stuff! If no, then what kind of transformation do I need to do?

Do not start computing correlations between variables without first seeing if it is appropriate to do so.

Interpreting Regression Output

We use the Chetty data again:

↪ $\texttt{Public} = 1$ if the college is public, 0 if private

↪ $\texttt{par\_median}$ is the parental median income in \$1,000s

↪ $\texttt{par\_top1pc}$ is % share of students with parents in the top 1%

↪ $\texttt{sat\_avg\_2013}$ is the average SAT score in 2013

Regressions:

↪ $\texttt{par\_median} = 75.5 + 6.9 \cdot \texttt{par\_top1pc}$

↪ $\texttt{par\_median} = 98.3 - 14.7 \cdot \texttt{Public}$

↪ $\texttt{sat\_avg\_2013} = 1087.7 - 39.4 \cdot \texttt{Public}$

↪ $\texttt{sat\_avg\_2013} = 711.4 + 3.9 \cdot \texttt{par\_median}$

Correlation and Causation

An important point I've made before and will make again:

The fact that two variables are correlated does not imply that there is a causal relationship between them.

But that's not all:

The fact that one variable causes another does not imply that there is the expected correlation between them (or any correlation).

Most people think hospitals cause people's health to improve (otherwise, why do we have them?), but we still often see negative correlations between hospitalization and health outcomes.