HOW TO READ A QUANTITATIVE PAPER

Reading a Quantitative Paper

1. Read the **abstract** and write down/underline the main point.
2. Read the **introduction**.
3. Skim the **data** section to figure out what *variables & units of observation* they care about and why.
4. Examine the **tables** presenting the data analysis.
   a) Focus on the variables the data section and introduction said were important.
   b) Look for stars (or calculate stars if needed) on those variables.
   c) Look at the direction of the effect.
5. Read the **conclusion**.
6. Skim the **results** section (always comes after the data and theory sections).
7. **Read the full paper (ignore anything in the methods section that doesn't make sense).**

UNDERSTANDING A COMMON TABLE: REGRESSIONS

VISUAL INTUITION

Regression function

- At the most basic level

$$y = \beta x + c$$

- c is a constant (the y intercept value)

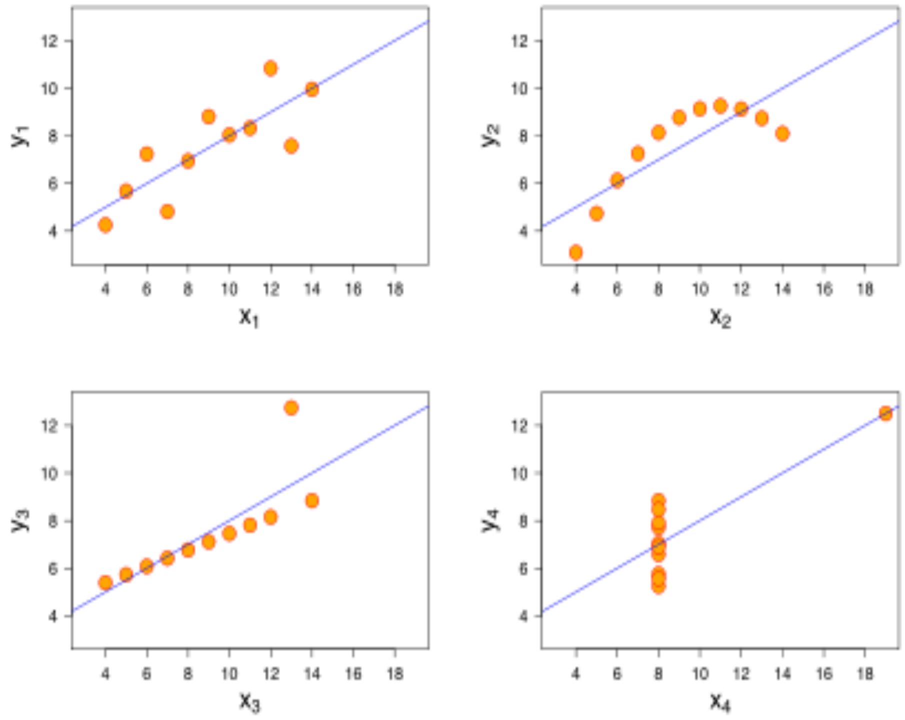
Potential for linear regressions and other techniques to obscure relevant information

- Each graph shows the best fitting straight line for the data

- All relationships were designed so that the standard numerical summaries look exactly identical even though the underlying relationships are clearly different.

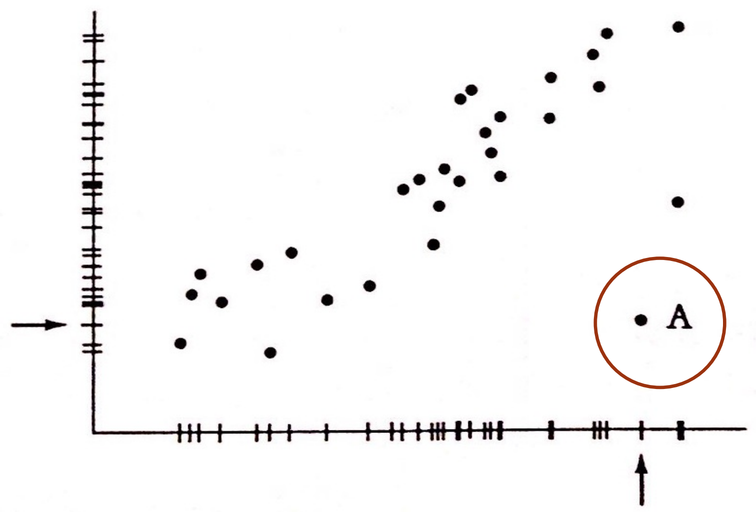
- Quick question:
  - Is beta positive or negative in each of these four graphs?

$$y = \beta x + c$$



More Complex Outliers

- Point A is called an outlier of the data
- It will corrupt your linear regressions
- Scatterplots reveal these distorting points



Regression Tables

Regression Table

Anscombe Results

	y1	y2	y3	y4
x1	estimated slope (std. error)			
x2		estimated slope (std. error)		
x3			estimated slope (std. error)	
x4				estimated slope (std. error)
_cons	estimated constant (std. error)	estimated constant (std. error)	estimated constant (std. error)	estimated constant (std. error)
R^2	corr. squared	corr. squared	corr. squared	corr. squared
N	observations	observations	observations	observations

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

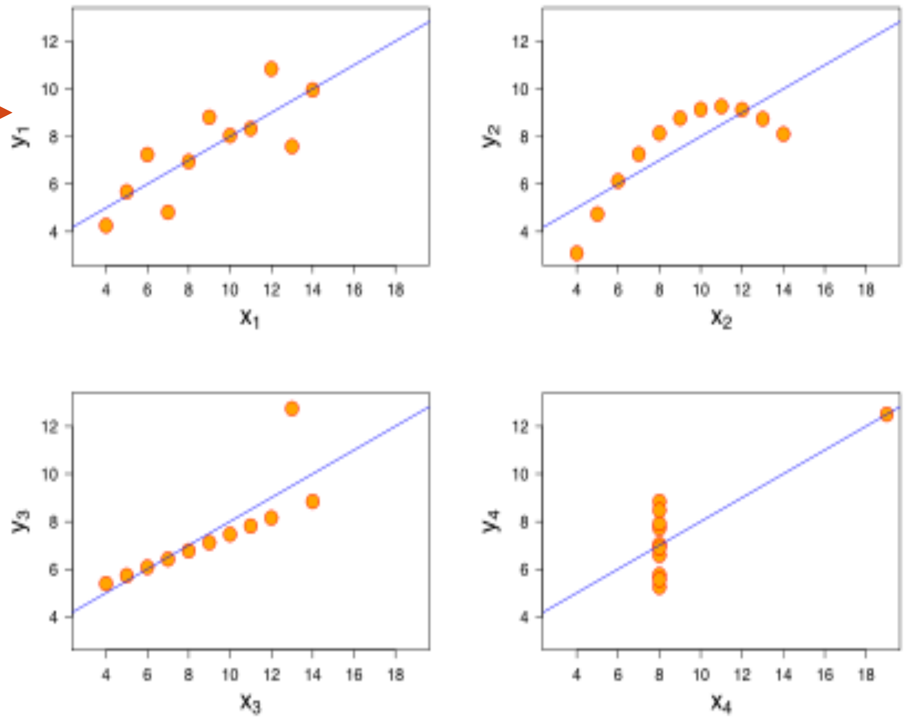
Anscombe Regression Results

Anscombe Results

	y1	y2	y3	y4
x1	0.500**			
	(0.118)			
x2		0.500**		
		(0.118)		
x3			0.500**	
			(0.118)	
x4				0.500**
				(0.118)
_cons	3.000*	3.001*	3.002*	3.002*
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



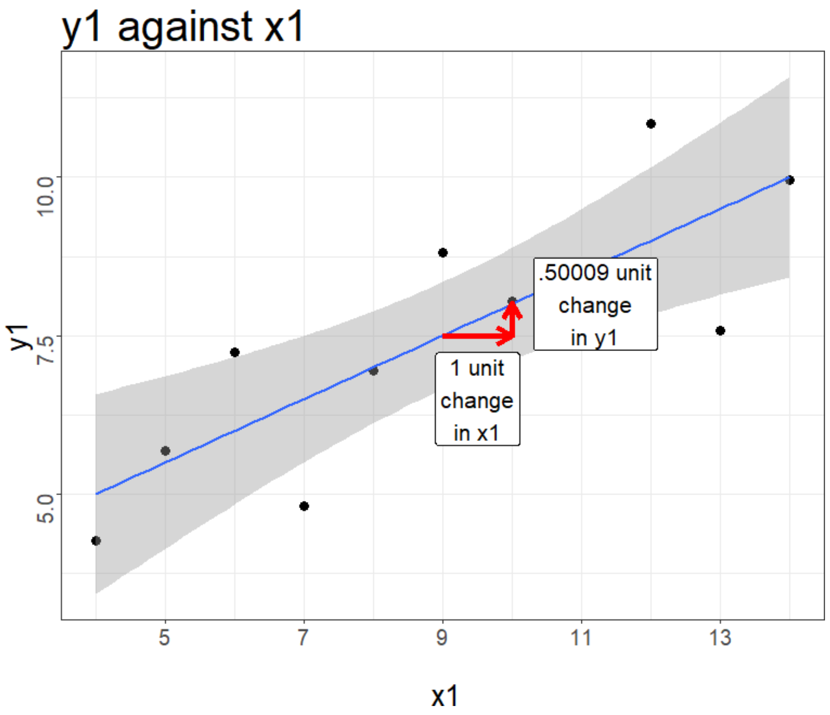
Anscombe Regression Results

Anscombe Results

	y1	y2	y3	y4
x1	0.500^{**}			
	(0.118)			
x2		0.500^{**}		
		(0.118)		
x3			0.500^{**}	
			(0.118)	
x4				0.500^{**}
				(0.118)
_cons	3.000^{*}	3.001^{*}	3.002^{*}	3.002^{*}
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



$$y = \hat{\beta}x + \hat{c}$$

$$y_1 = 0.5000909x_1 + 3.0000909$$

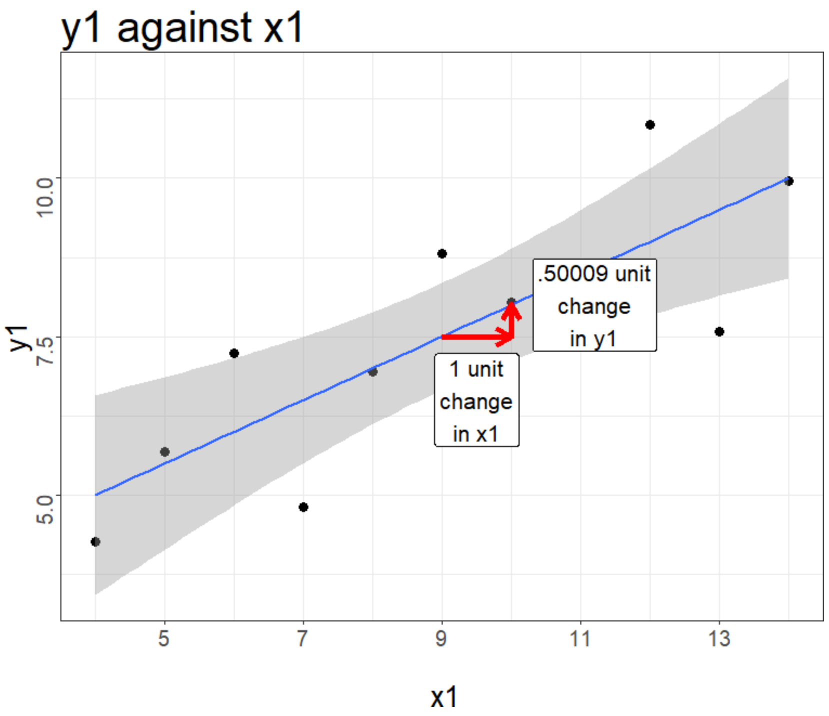
Anscombe Regression Results

Anscombe Results

	y1	y2	y3	y4
x1	0.500^{**}			
	(0.118)			
x2		0.500^{**}		
		(0.118)		
x3			0.500^{**}	
			(0.118)	
x4				0.500^{**}
				(0.118)
_cons	3.000^{*}	3.001^{*}	3.002^{*}	3.002^{*}
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



$$y = \hat{\beta}x + \hat{c}$$

$$y_1 = 0.5000909x_1 + 3.0000909$$

Anscombe Regression Results

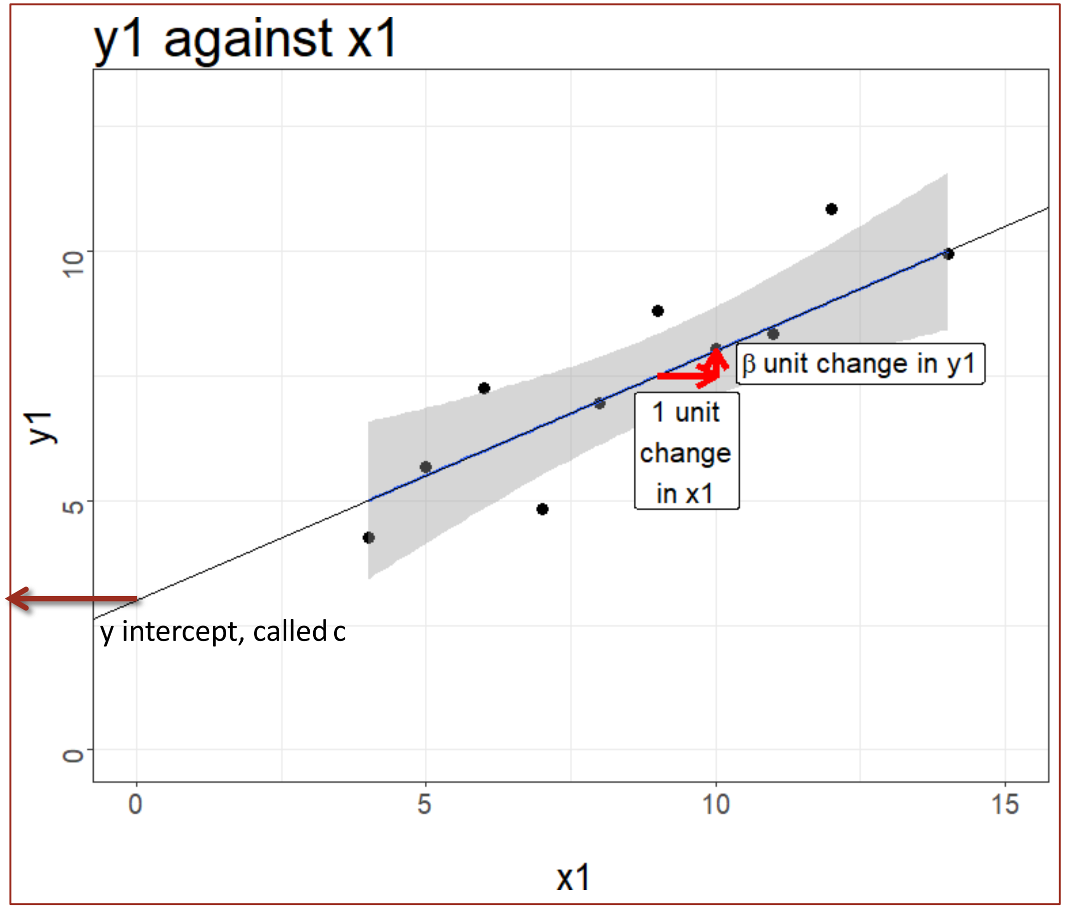
Anscombe Results

	y1	y2	y3	y4
x1	0.500**			
	(0.118)			
x2		0.500**		
		(0.118)		
x3			0.500**	
			(0.118)	
x4				0.500**
				(0.118)
_cons	3.000*	3.001*	3.002*	3.002*
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.0000909



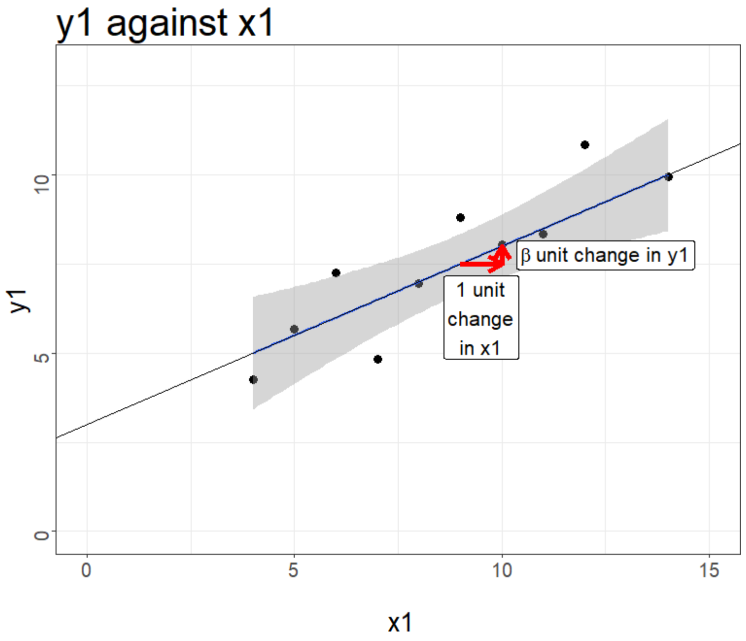
Anscombe: Regressions & Uncertainty

Anscombe Results

	y1	y2	y3	y4
x1	0.500**			
	(0.118)			
x2		0.500**		
		(0.118)		
x3			0.500**	
			(0.118)	
x4				0.500**
				(0.118)
_cons	3.000*	3.001*	3.002*	3.002*
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



The grey area represents the 95% confidence interval for the values of the slope given the variation in the data.

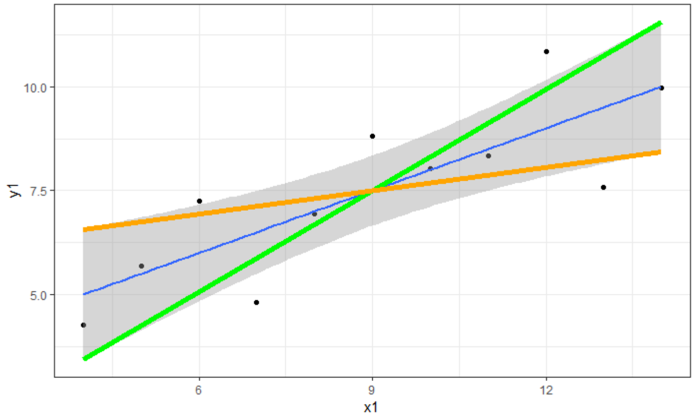
This visualizes the range of confidence we have in the slope. If a flat line fits inside the gray shaded area, then the slope could be 0. (It isn't here)

$$\begin{aligned}\text{estimated slope} \pm 1.96 * \text{standard error} &= \hat{\beta} \pm 1.96 * \sigma \\ &= 0.5 \pm 1.96 * 0.118 \\ &= (0.5 - 1.96 * 0.118, .5 + 1.96 * 0.118) \\ &= (0.269, 0.731)\end{aligned}$$

Anscombe Regression Uncertainty

$$y = \hat{\beta}x + \hat{c}$$

$$y_1 = 0.5000909x_1 + 3.0000909$$



Anscombe Results

	y1	y2	y3	y4
x1	0.500^{**}			
	(0.118)			
x2		0.500^{**}		
		(0.118)		
x3			0.500^{**}	
			(0.118)	
x4				0.500^{**}
				(0.118)
_cons	3.000^{*}	3.001^{*}	3.002^{*}	3.002^{*}
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$$\begin{aligned} \text{estimated slope} \pm 1.96 * \text{standard error} &= \hat{\beta} \pm 1.96 * \sigma \\ &= 0.5 \pm 1.96 * 0.118 \\ &= (0.5 - 1.96 * 0.118, .5 + 1.96 * 0.118) \\ &= (0.269, 0.731) \end{aligned}$$

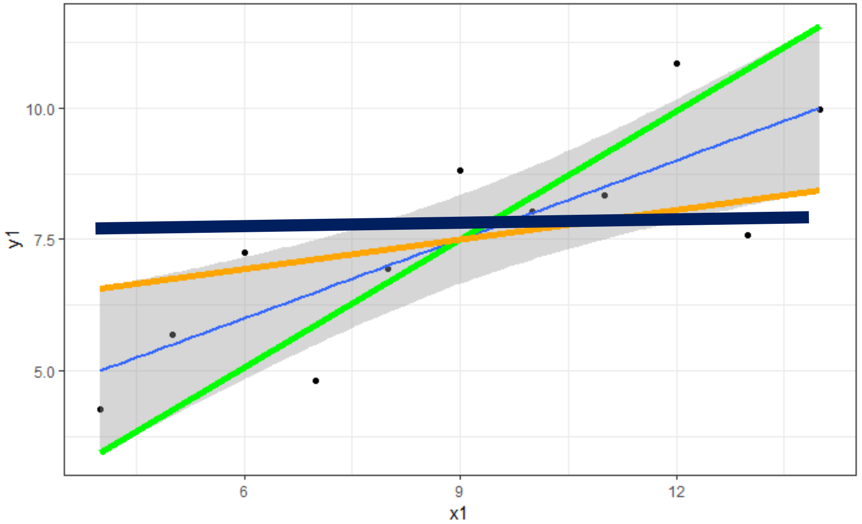
The lowest slope in the 95% confidence interval (orange): 0.269

The steepest slope in the 95% confidence interval (green): 0.731

Anscombe Regression Uncertainty

$$y = \hat{\beta}x + \hat{c}$$

$$y_1 = 0.5000909x_1 + 3.0000909$$



Anscombe Results

	y1	y2	y3	y4
x1	0.500^{**}			
	(0.118)			
x2		0.500^{**}		
		(0.118)		
x3			0.500^{**}	
			(0.118)	
x4				0.500^{**}
				(0.118)
_cons	3.000^{*}	3.001^{*}	3.002^{*}	3.002^{*}
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

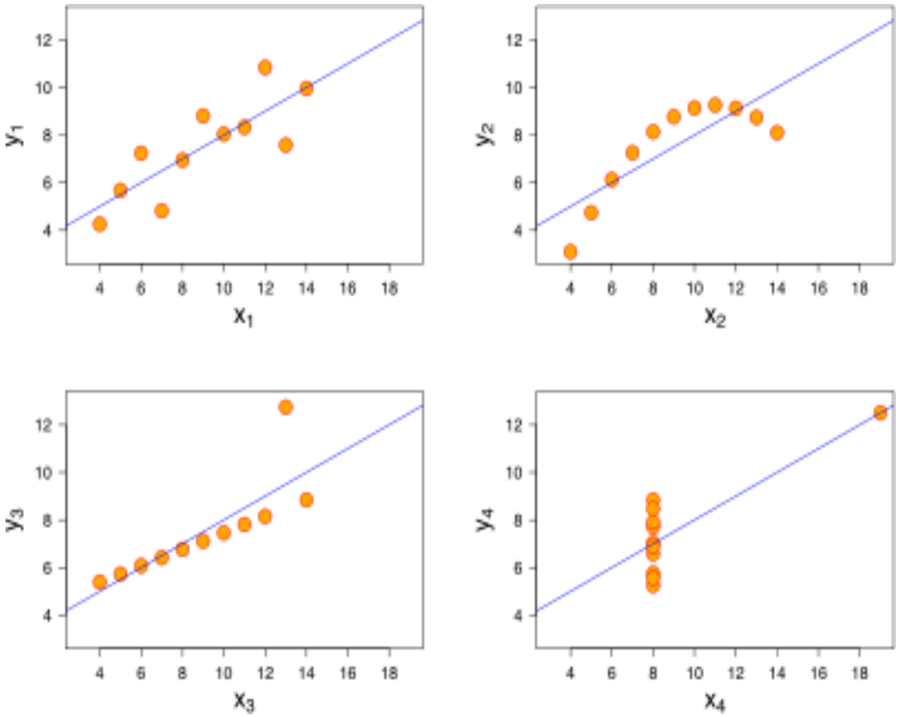
Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Conclusion: No flat line can live in the grey 95% confidence interval; therefore, we are 95% confident that the slope is not flat.

Potential for linear regressions and other techniques to obscure relevant information

- Each graph shows the best fitting straight line for the data
- All relationships were designed so that the standard numerical summaries look exactly identical even though the underlying relationships are clearly different.



Anscombe Results

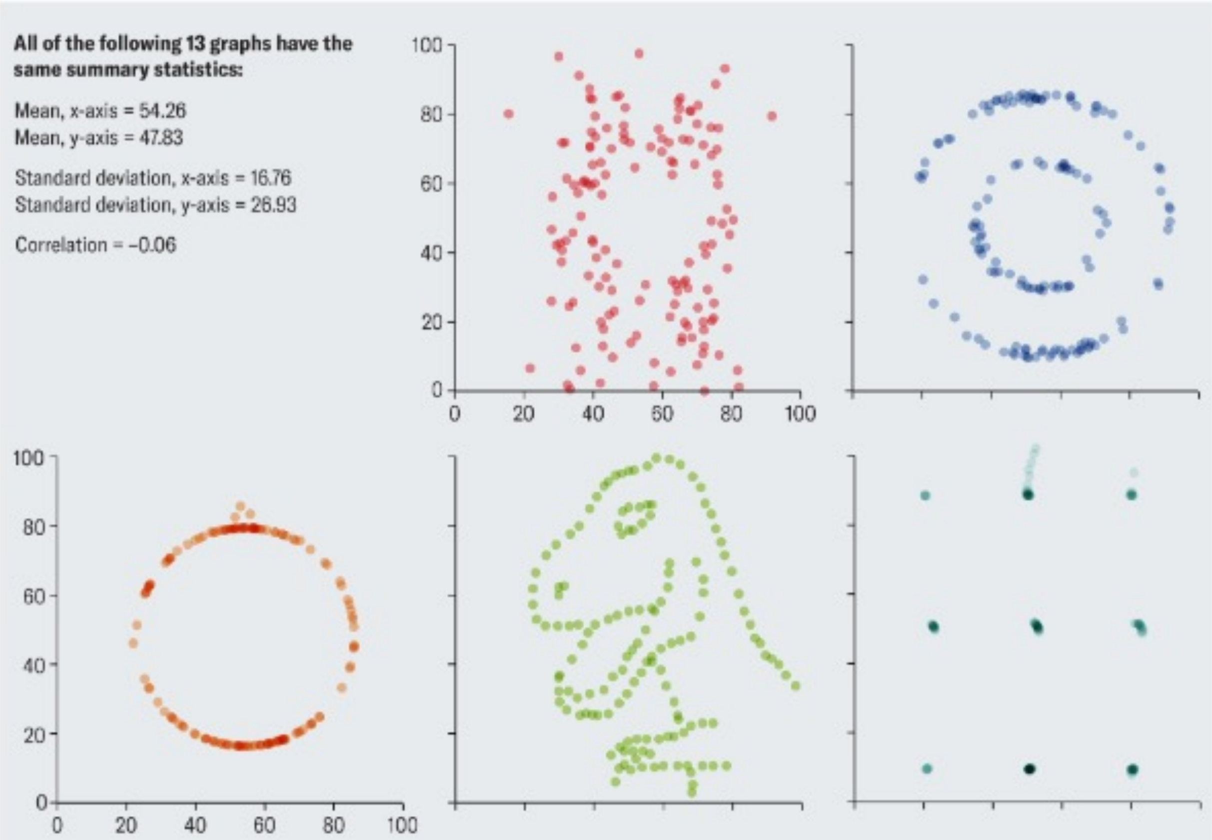
	y1	y2	y3	y4
x1	0.500**			
	(0.118)			
x2		0.500**		
		(0.118)		
x3			0.500**	
			(0.118)	
x4				0.500**
				(0.118)
_cons	3.000*	3.001*	3.002*	3.002*
	(1.125)	(1.125)	(1.124)	(1.124)
R^2	0.667	0.666	0.666	0.667
N	11	11	11	11

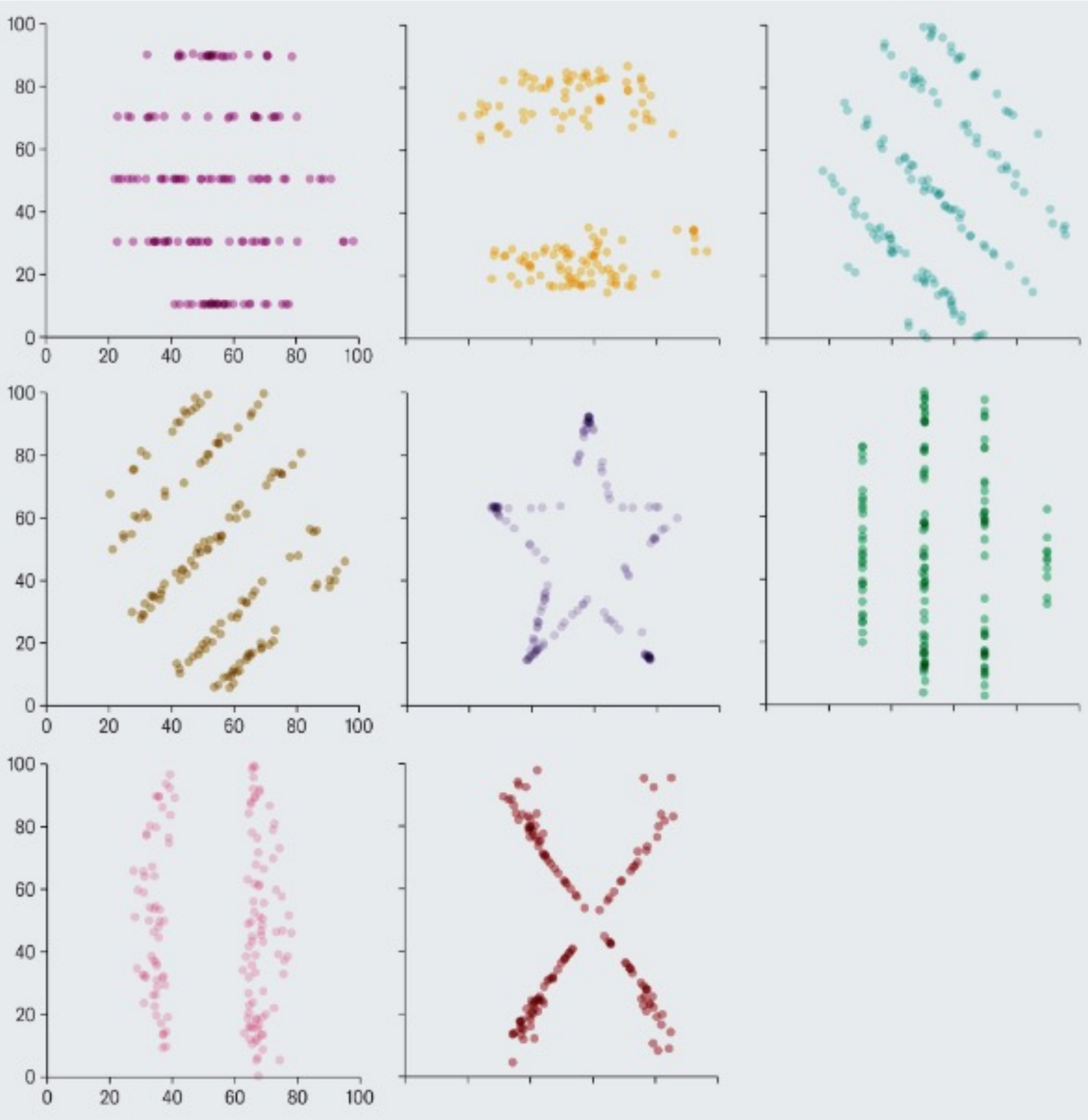
Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For fun:

All of these graphs have the same summary statistics





APPLICATION:
Counties Shifting to Trump

See link

https://swat-ssql.github.io/read-regression-table/

Statistical vs Substantive SIGNIFICANCE

Ted Underwood
@Ted_Underwood

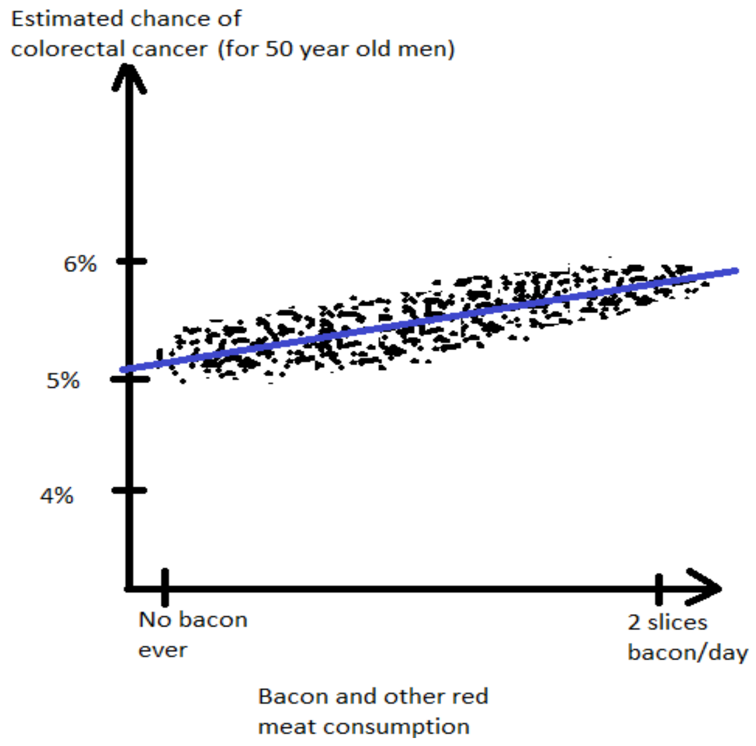
Follow

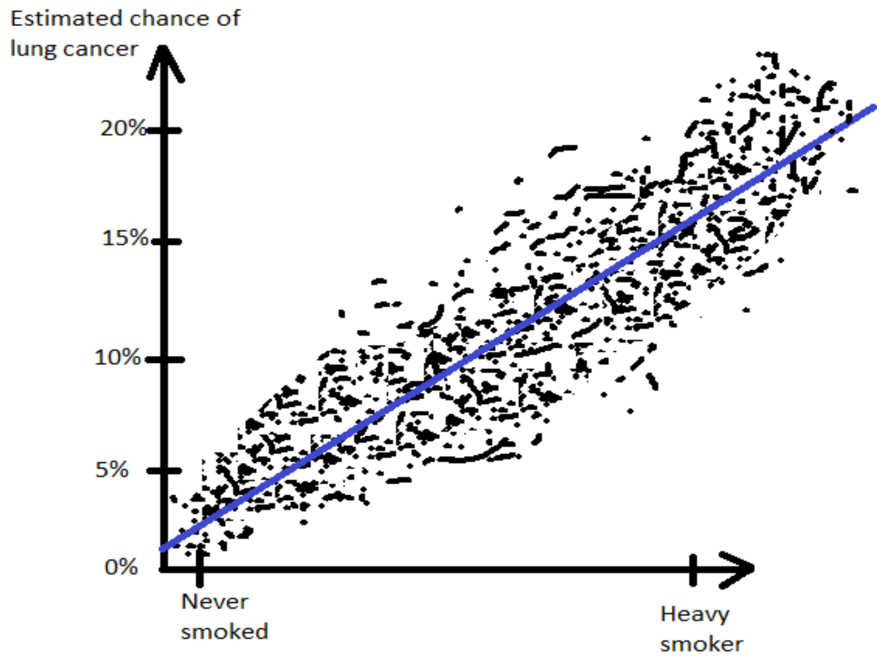
A stubborn love of bacon just taught more Americans the difference between p values and effect size than 100 stats courses could.

5:28 AM - 27 Oct 2015

1,490 Retweets **1,105** Likes

Large effect size vs fuzzy data (not real data, but based on real numbers)





Statistical "significance"/p values are not *substantive significance/effect sizes*

- In layman's English, "significance" means "importance"
- In statistics, it just means how much do we believe the effect is real/not zero
- I believe that eating more bacon causes colorectal cancer. I'm ***not*** *sure whether I think a .05% higher chance of cancer from eating bacon once per month is* ***important***.
- I believe that smoking causes lung cancer. I am sure that a 9.1% higher chance of cancer (0.4% to 9.5%) is ***very important*****.**

You can have a very small effect size with a lot of stars

- This usually happens with a lot of data. The more data you have, the more likely the data reflects reality. Even if the reality is very small in size.

Mathematical definition is complex

NOTE: This is a very surface level description. The stars represent the “p-value”, which indicates how often the sample being used would yield the direction of the coefficient if the ‘true’ effect was 0.

Reading a Quantitative Paper

START BY SKIMMING

Reading a Quantitative Paper

1. Read the **abstract** and write down/underline the main point.
2. Read the **introduction**.
3. Skim the **data** section to figure out what *variables & units of observation* they care about and why.
4. Examine the **tables** presenting the data analysis.
   a) Focus on the variables the data section and introduction said were important.
   b) Look for stars (or calculate stars if needed) on those variables.
   c) Look at the direction of the effect.
5. Read the **conclusion**.
6. Skim the **results** section (always comes after the data and theory sections).
7. **Read the full paper (ignore anything in the methods section that doesn't make sense).**

Reading a Quantitative Paper

Steps 1-5 are the key steps to understand the main point of a paper

- Won’t let you critique the methodology

Step 1-2 are often enough to give you a cursory idea of what the paper says.

Steps 6 and 7 are needed to critique the content of the paper.

seriously, don’t spend too much time trying to understand the methods section

More on Confidence Intervals:

Why We Pay Attention to Coefficients with Stars

Normal Distribution

Most common distribution in statistics and life

Symmetric, unimodal, bell curve

Many distributions of events are effectively normal (height, blood pressure, SAT scores)

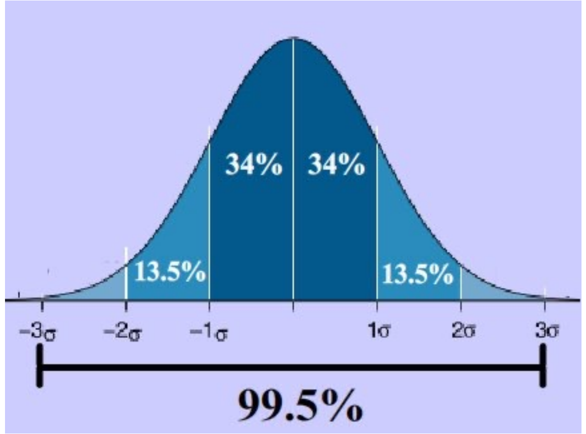
Critically, the distribution of any average value follows a normal distribution

- This also applies to any expected value

No distribution will be perfectly normal, because we live in the real world. But many will be so close that it's the most effective distribution to use in calculations.

95% Confidence Interval

Generic Normal Distribution



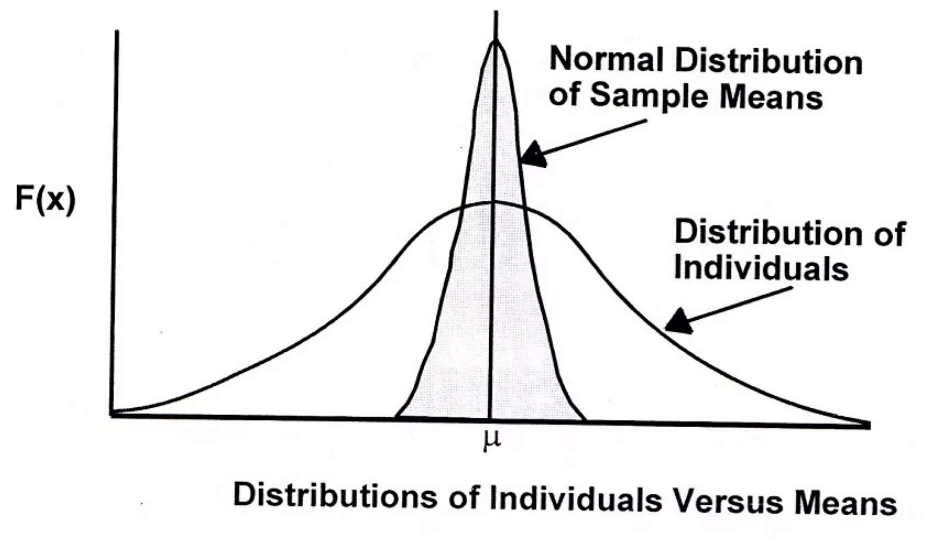
95% of the population falls within 1.96 standard deviations (σ) of the mean value.

99% of the population falls within 2.58 standard deviations (σ) of the mean value.

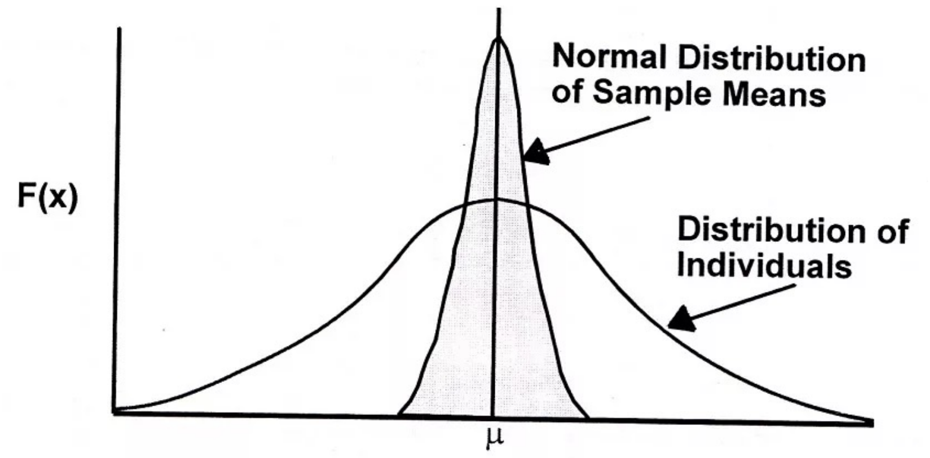
99.9% of the population falls within 3.291 standard deviations (σ) of the mean value.

Difference between distribution of observations and distribution of means

The distribution of the individuals/observations is much wider than the distribution of the means of those observations



95% Confidence Interval



Distributions of Individuals Versus Means

95% of all possible coefficients fall within 1.96 standard errors (σ) of the expected coefficient.

99% of all possible coefficients fall within 2.58 standard errors (σ) of the expected coefficient.

99.9% of all possible coefficients fall within 3.291 standard errors (σ) of the expected coefficient.

More confidence interval examples

Step 0: Know the variables and what they mean

Demographics and budgetary spending

1. Budget for the chief of staff of a member of Congress, measured as a percentage of their total expenditures on staffing salaries.
2. Median income of the district of a member of Congress

Question: Do members of Congress from higher income districts spend more of their budget on their chief of staff?

Step 1 (Republicans): Stars

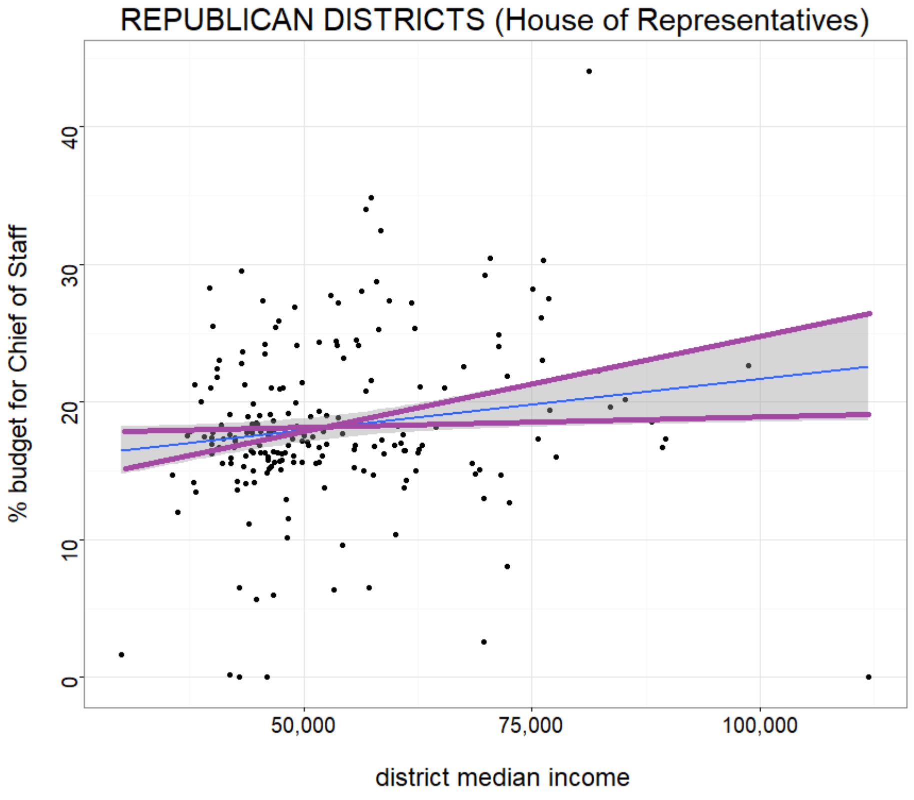
There are stars on the coefficient for median income for Republicans (but not Democrats).

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.058*	−0.030
	(0.027)	(0.028)
Constant	14.969***	17.389***
	(1.705)	(1.643)
Observations	233	201
Note:	*p<0.05; **p<0.01; ***p<0.001	

Step 1 (Republicans) : Stars and the Standard Error

The number in parentheses is the **standard error**. The larger this number is relative to the estimated effect size, the less certain the estimate is.

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.058*	−0.030
	(0.027)	(0.028)
Constant	14.969***	17.389***
	(1.705)	(1.643)
Observations	233	201
Note:	*p<0.05; **p<0.01; ***p<0.001	



Step 1 (Republicans): Calculating the maximum/minimum line in the shaded area

95% Confidence Interval (95% of all expected regression coefficients will be in this range):

$$\text{estimated slope} \pm 1.96 * \text{standard error} =$$

$$\begin{aligned} \hat{\beta} \pm 1.96 * \sigma &= 0.058 \pm 1.96 * 0.027 \\ &= (0.058 - 1.96 * 0.027, 0.058 + 1.96 * 0.027) \\ &= (0.00508, 0.11092) \end{aligned}$$

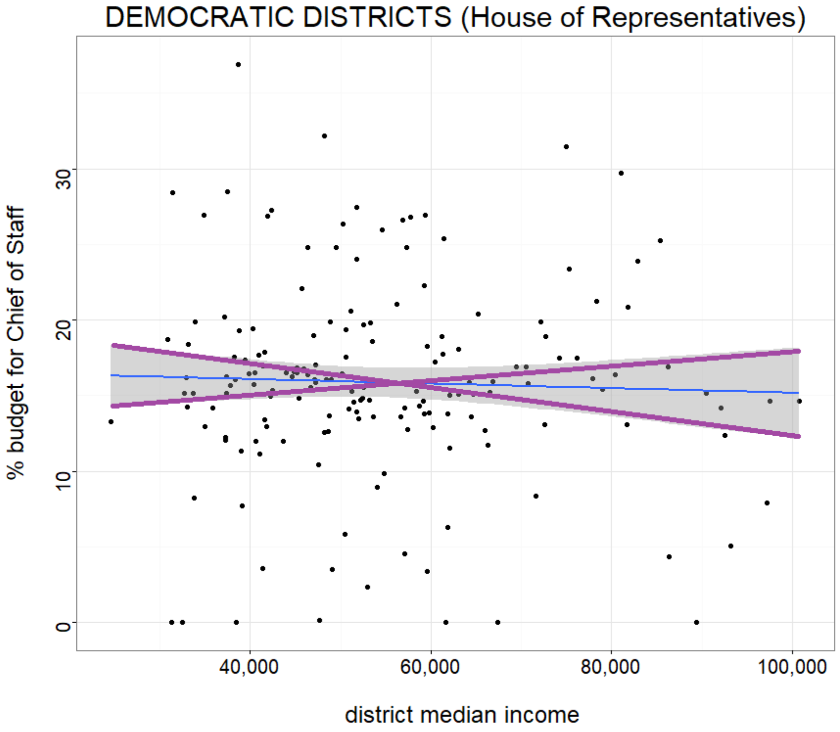
So 95% of the time the ‘true’ coefficient is positive (between 0.00508 and 0.11092),

We expect that the average, expected value of the coefficient is 0.058.

Step 1 (Democrats): No stars

The number in parentheses is the **standard error**. The larger this number is relative to the estimated effect size, the less certain the estimate is.

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.058*	−0.030
	(0.027)	(0.028)
Constant	14.969***	17.389***
	(1.705)	(1.643)
Observations	233	201
Note:	*p<0.05; **p<0.01; ***p<0.001	



95% Confidence Interval (95% of all expected regression coefficients will be in this range):

$$\begin{aligned}\hat{\beta} \pm 1.96 * \sigma &= -0.030 \pm 1.96 * 0.028 \\ &= (-0.030 - 1.96 * 0.028, -0.030 + 1.96 * 0.028) \\ &= (-0.08488, 0.02488)\end{aligned}$$

So some of the time the ‘true’ coefficient is negative (between -0.08488 and 0), and sometimes it is positive (between 0 and 0.02488)

We expect that the average, expected value of the coefficient is -0.030.

More Interpretation Facts: Multivariate vs Univariate Regression

Multivariate Regression

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.070**	−0.033
	(0.031)	(0.027)
years of seniority	−0.268	−0.394*
	(0.247)	(0.217)
age	−0.018	−0.136***
	(0.049)	(0.051)
Constant	15.908***	26.838***
	(2.757)	(3.078)
Observations	233	201

Multivariate v Univariate Regression

Coefficients change when you add new variables:

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.070**	−0.033
	(0.031)	(0.027)
years of seniority	−0.268	−0.394*
	(0.247)	(0.217)
age	−0.018	−0.136***
	(0.049)	(0.051)
Constant	15.908***	26.838***
	(2.757)	(3.078)
Observations	233	201

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.058*	−0.030
	(0.027)	(0.028)
Constant	14.969***	17.389***
	(1.705)	(1.643)
Observations	233	201
Note:	*p<0.05; **p<0.01; ***p<0.001	

Comparing Coefficients

- ONLY compare coefficients if the variables are measured in the same units

	% Budget for Chief of Staff:	
	Republicans	Democrats
median income ($10,000 dollars)	0.070**	−0.033
	(0.031)	(0.027)
years of seniority	−0.268	−0.394*
	(0.247)	(0.217)
age	−0.018	−0.136***
	(0.049)	(0.051)
Constant	15.908***	26.838***
	(2.757)	(3.078)
Observations	233	201

Intercept/constant/b

- This is not a variable of interest. It represents the y-intercept... where the estimated line crosses the y-axes.
- Why have it? Because it is needed to write out the regression function.
- The intercept often doesn't have real world significance. For example, no district has $0 median income.

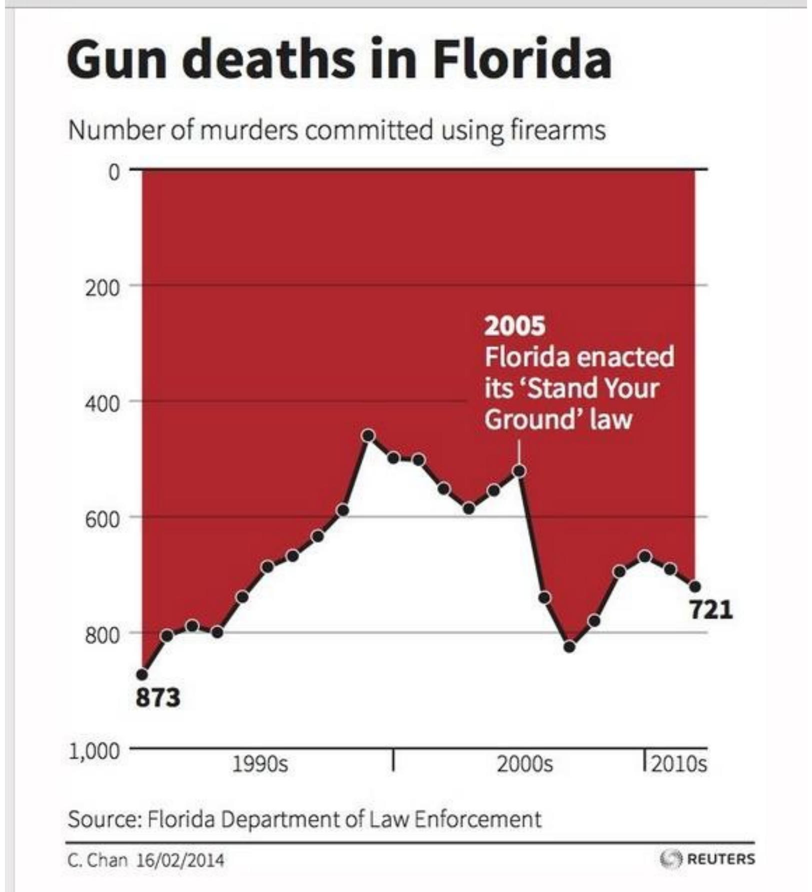
Size of effect

- For now:
  - You can only estimate the size of the effect if it is a linear regression
  - Otherwise, rely on the text in the paper to provide insight into what the size of the effect means

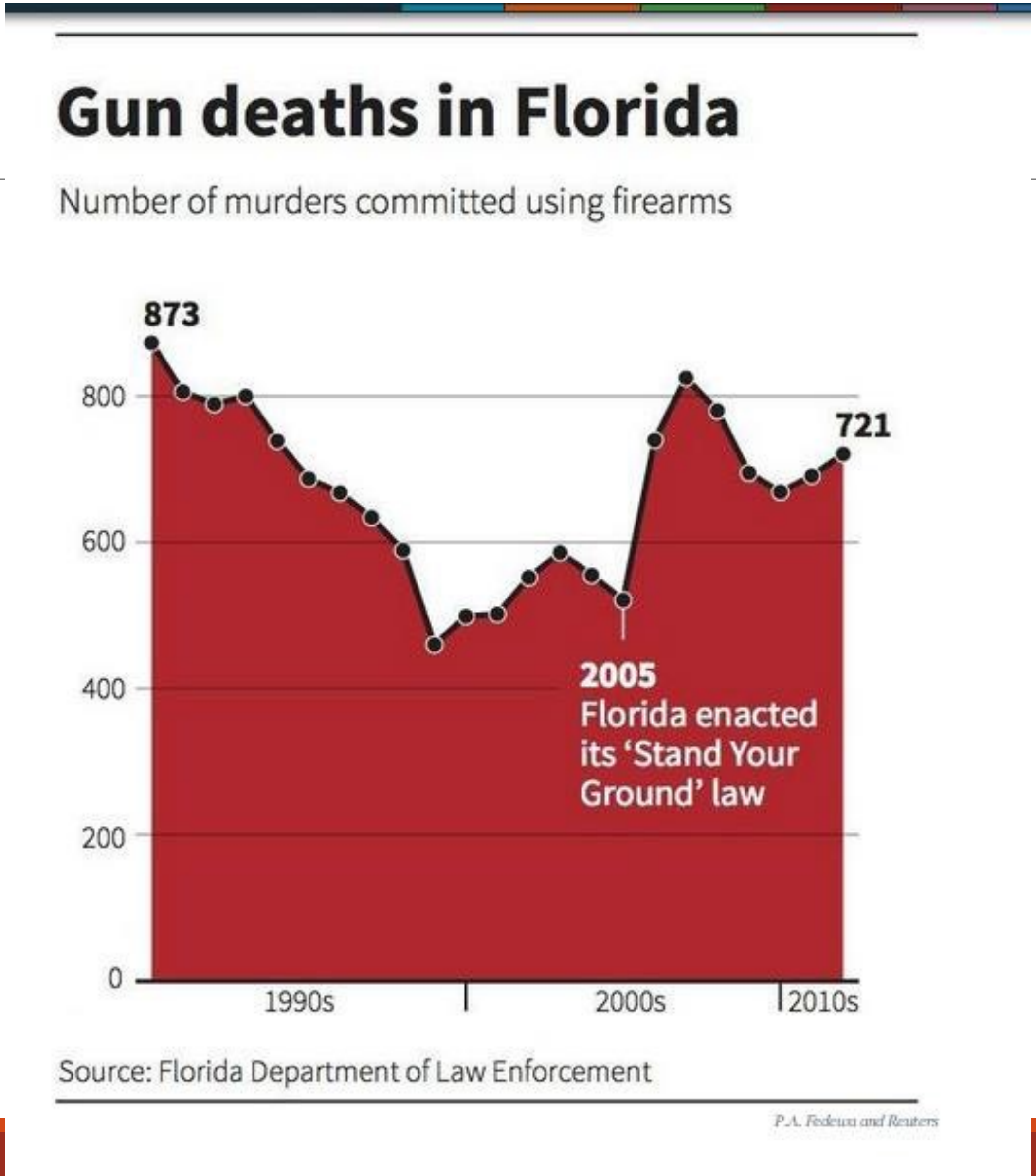
How to Lie With Graphs

What's Wrong with This Picture?

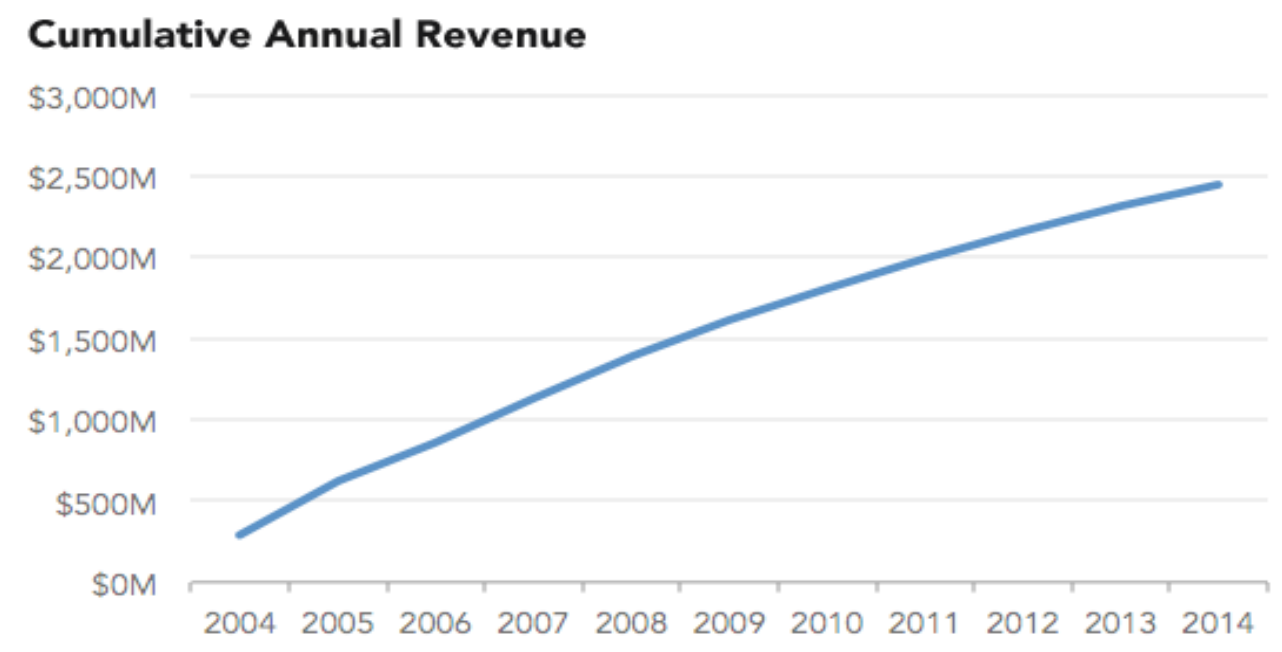
This graph was created and printed by Reuters



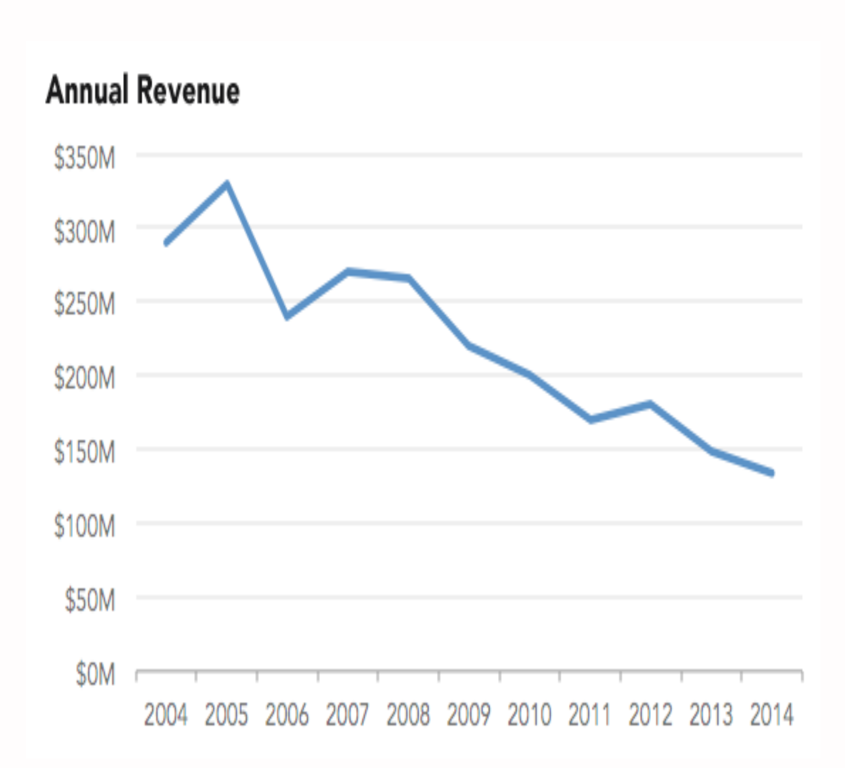
Corrected Graph

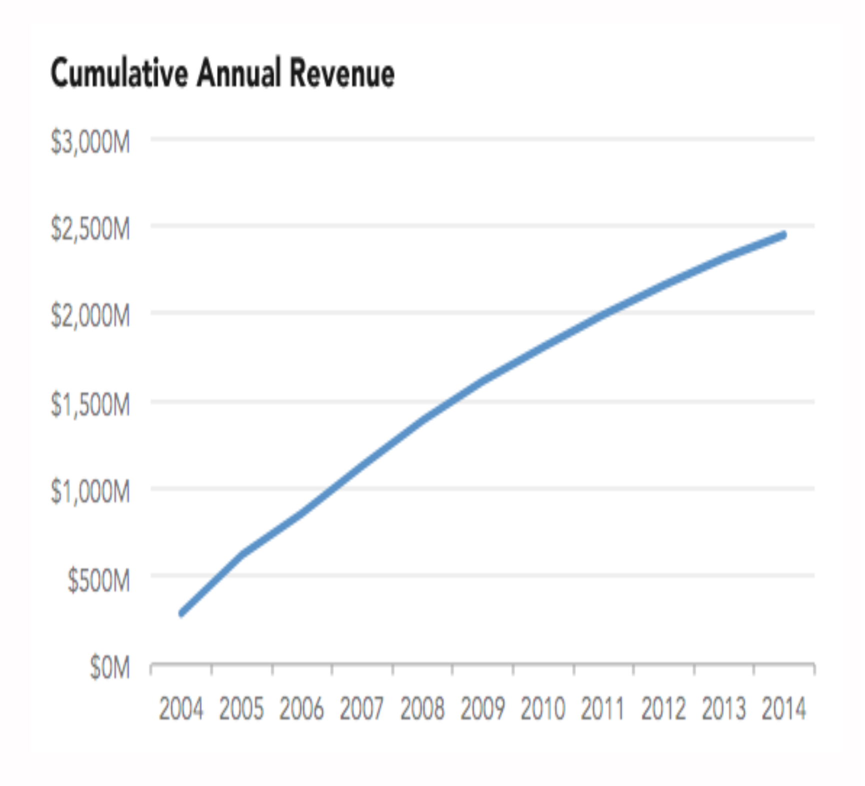


Size of effect



Size of effect





Spurious Correlation

http://www.tylervigen.com/spurious-correlations

More about the dinosaur graphic

What This Graph of a Dinosaur Can Teach Us about Doing Better Science

Jack Murtagh, Scientific American 2023