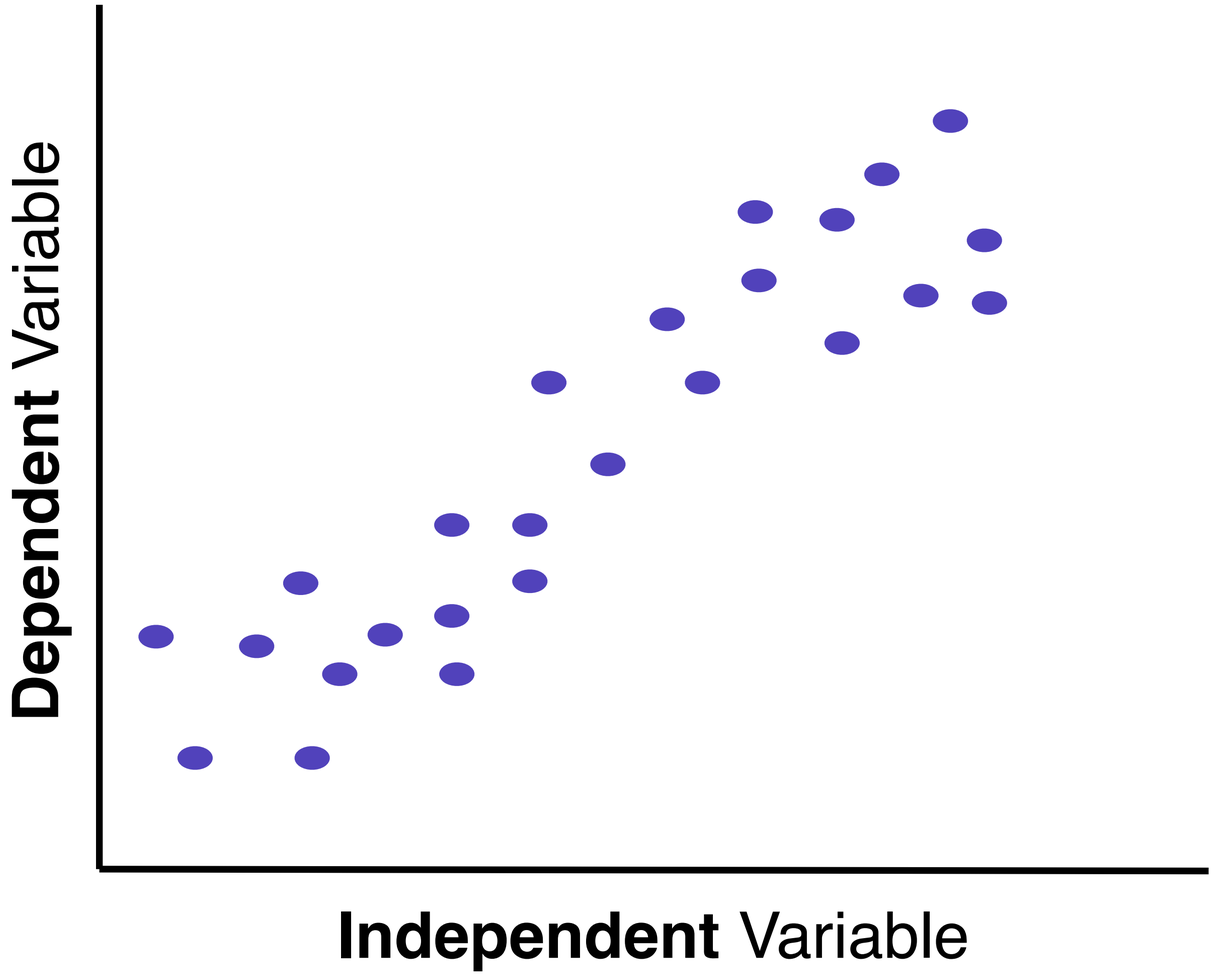
Reminders

Upcoming due dates
Mon Oct 27th Quiz 4
Wed Oct 29th Project Proposal

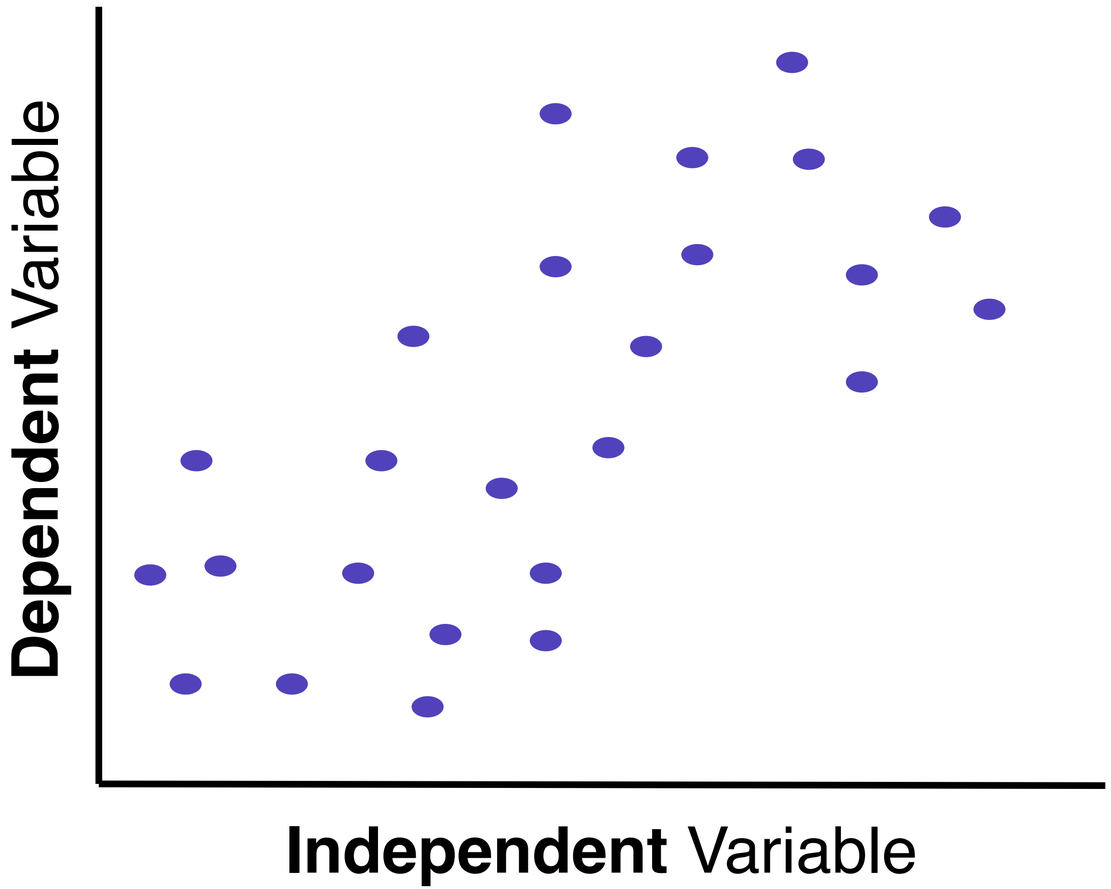
Statistical inference II

Data Science in Practice

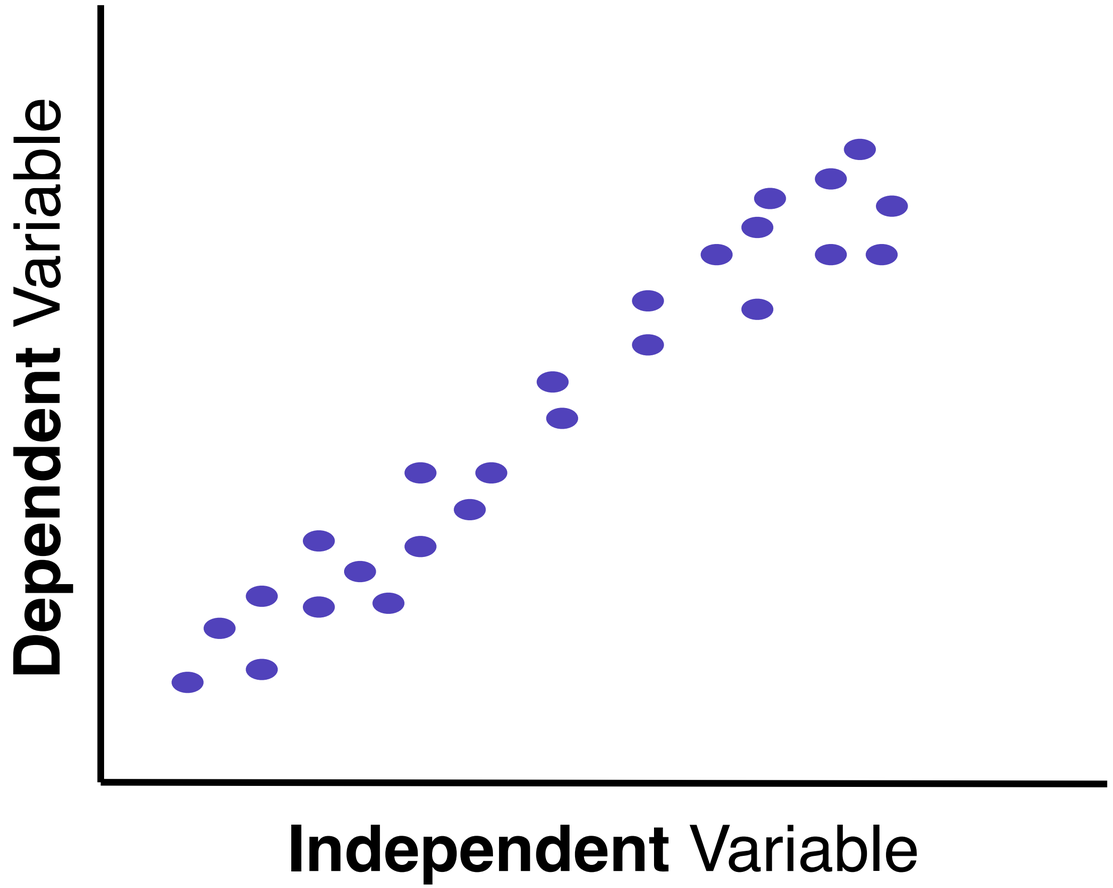
Jason G. Fleischer, PhD Dept. of Cognitive Science UC San Diego https://jgfleischer.com


Dependent Variable
Independent Variable

weaker relationship

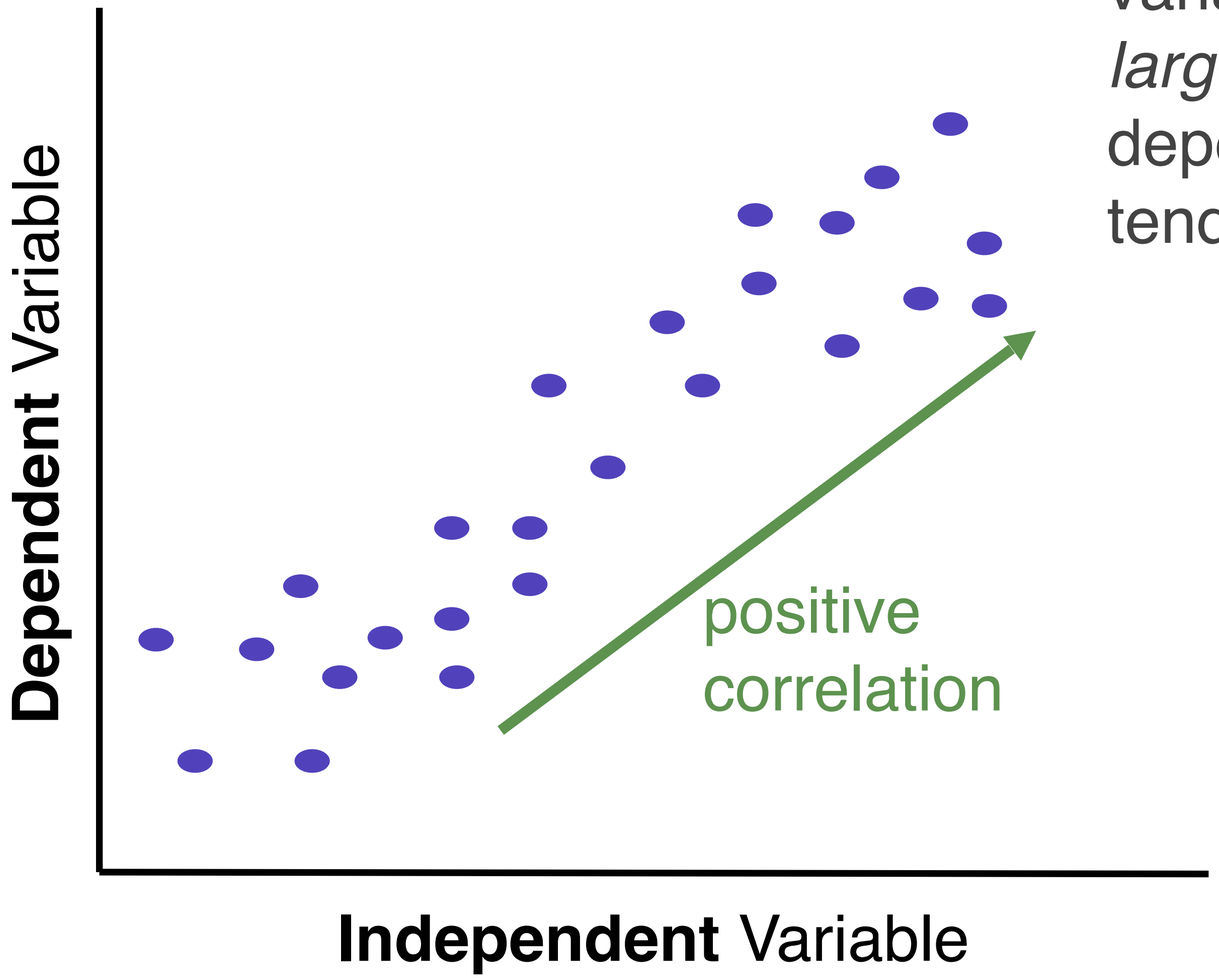


stronger relationship



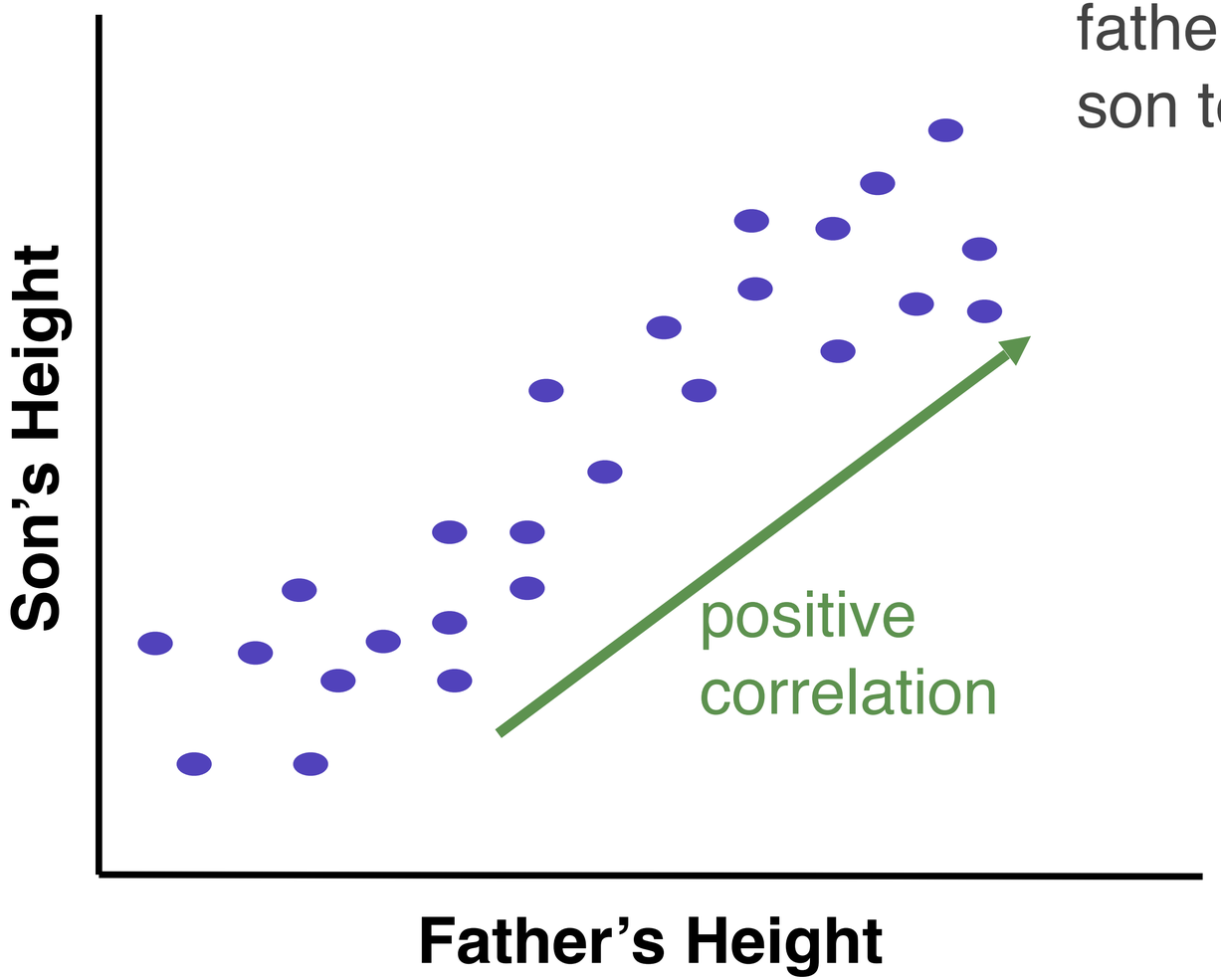
stronger relationship = higher correlation

The *larger* the independent variable value, the *larger* the dependent variable tends to be



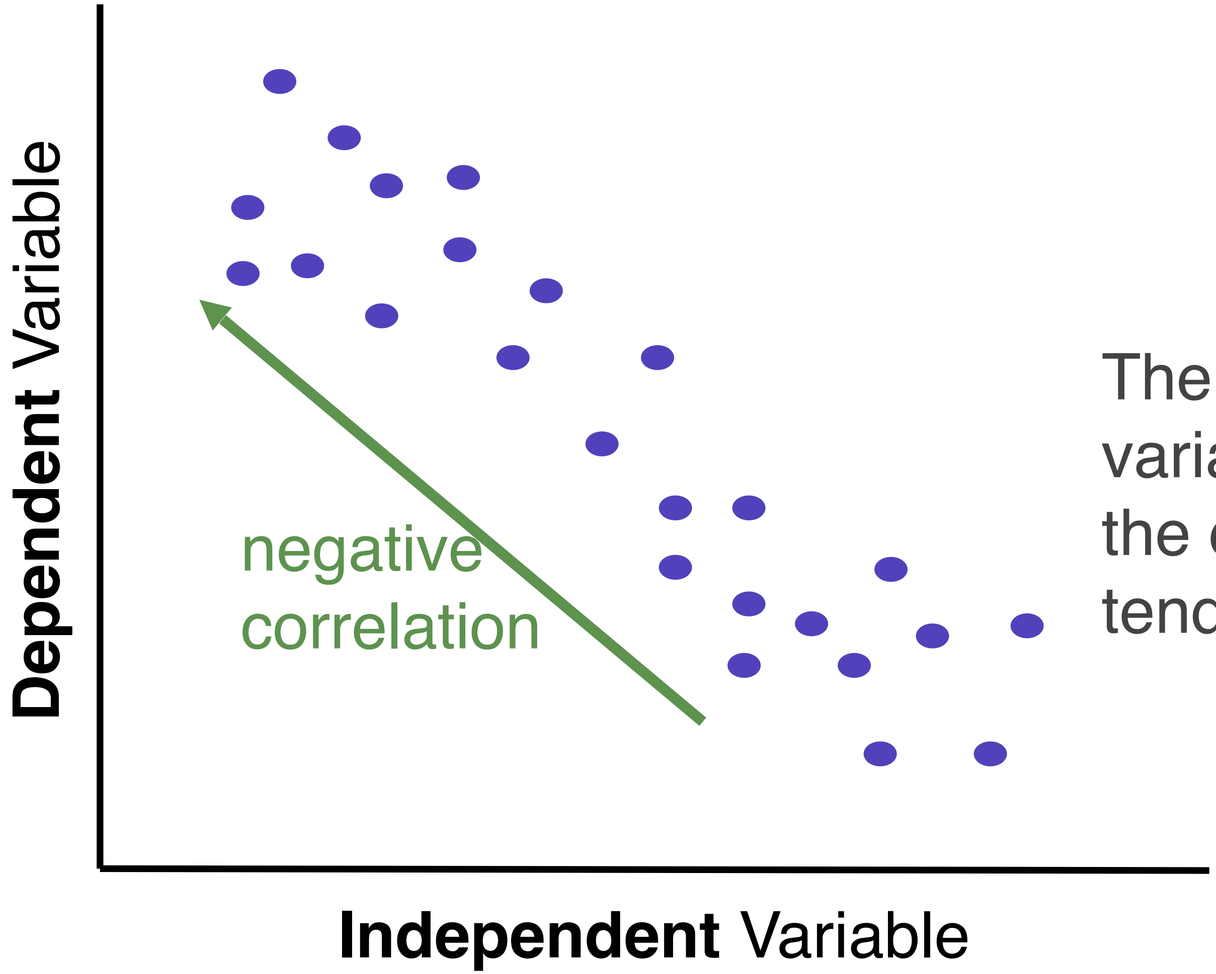
The *smaller* the independent variable value, the *smaller* the dependent variable tends to be

The *taller* the father, the taller his son tends to be



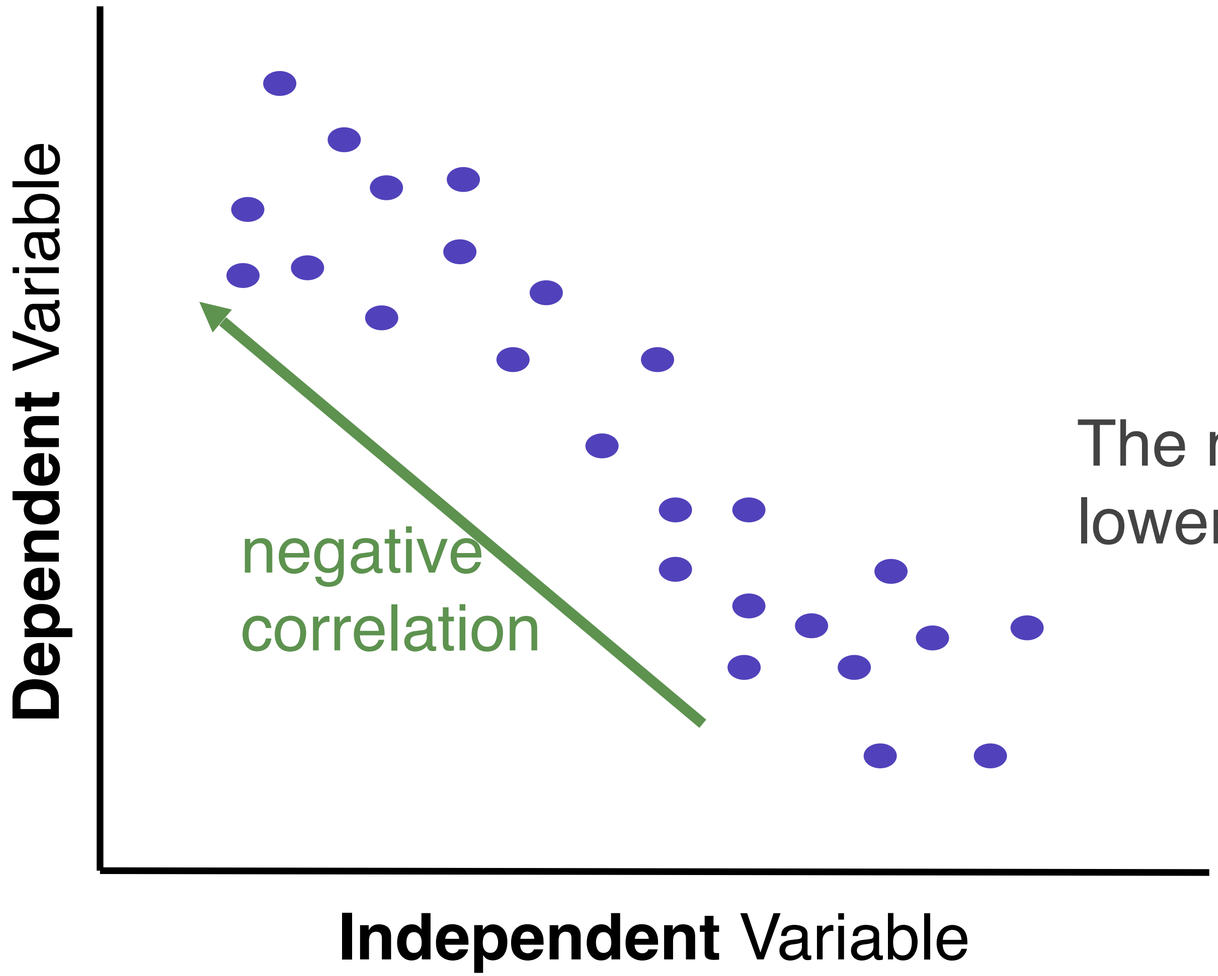
The *shorter* the father, the shorter his son tends to be

The *smaller* the independent variable value, the *larger* the dependent variable tends to be



The *larger* the independent variable value, the *smaller* the dependent variable tends to be

The *lower* the number of absences, the *higher* the students' grades tend to be



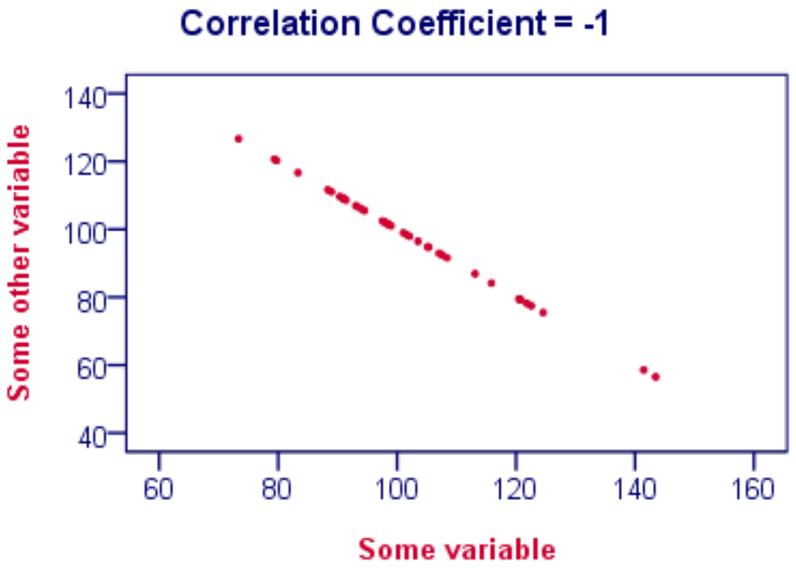
The more absences, the lower the students' grades

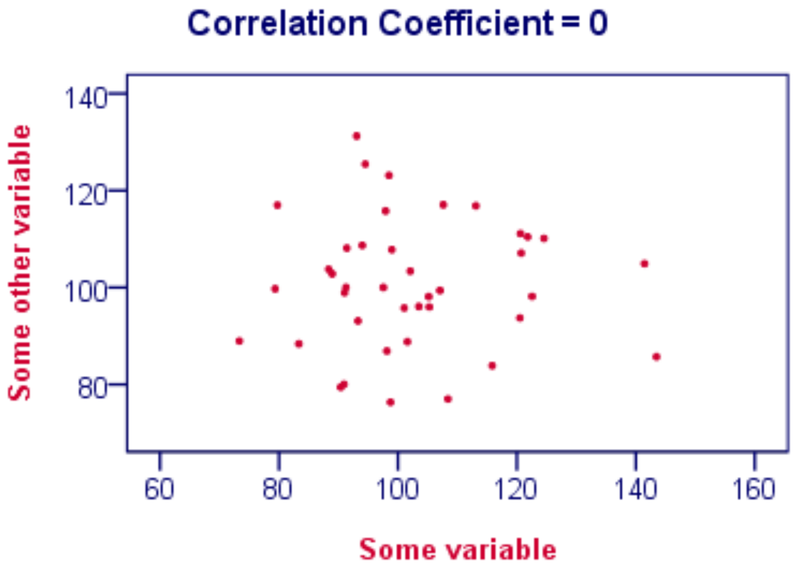
Pearson's *r* :

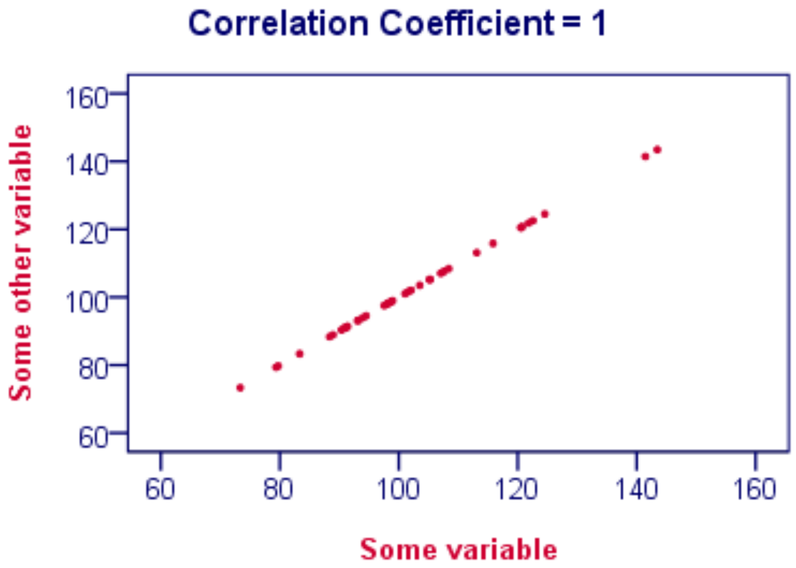
linear correlation between two variables

takes values [-1,1]

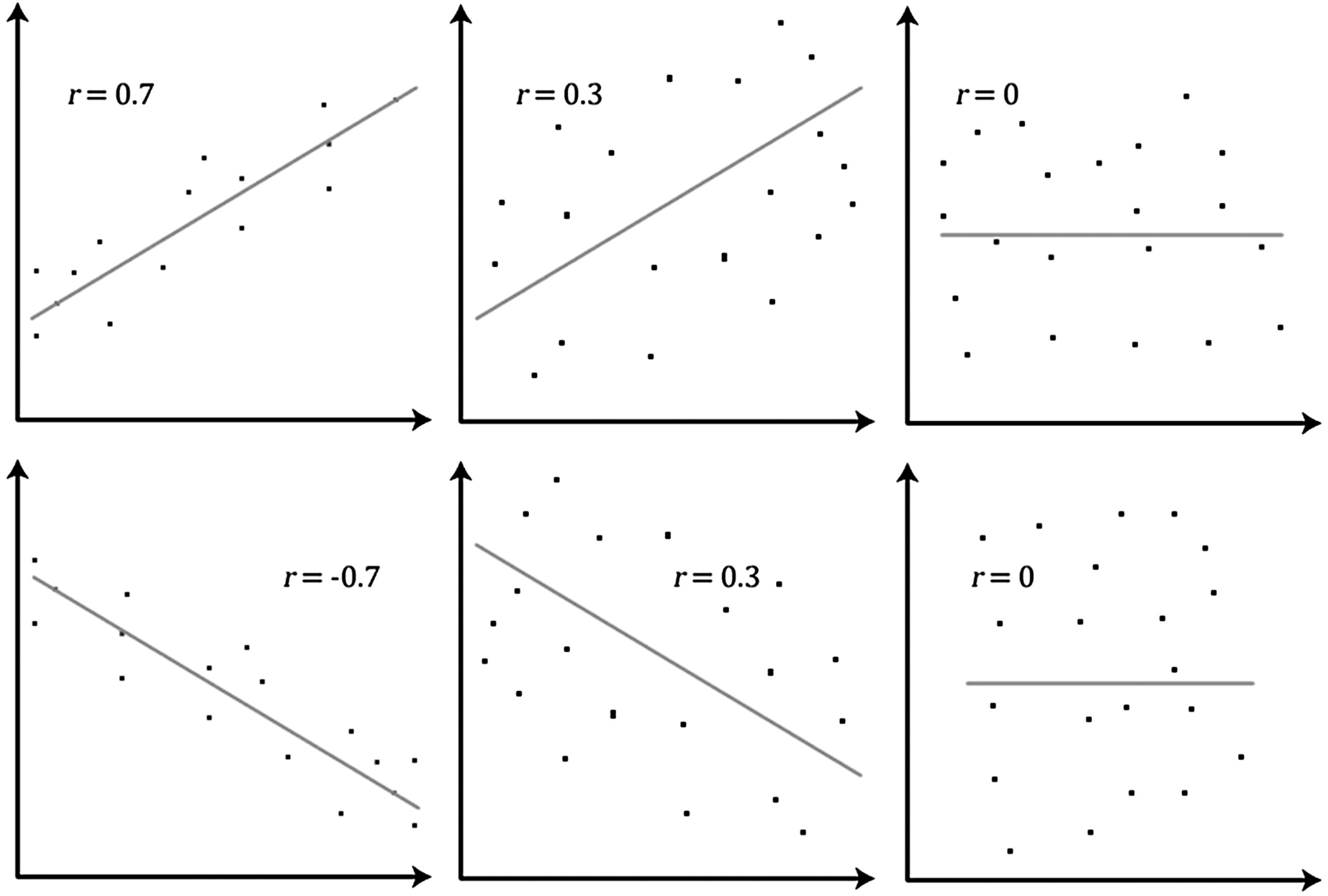
Correlation is how close the data are to being in a line…
BUT IT HAS NOTHING TO DO WITH THE SLOPE





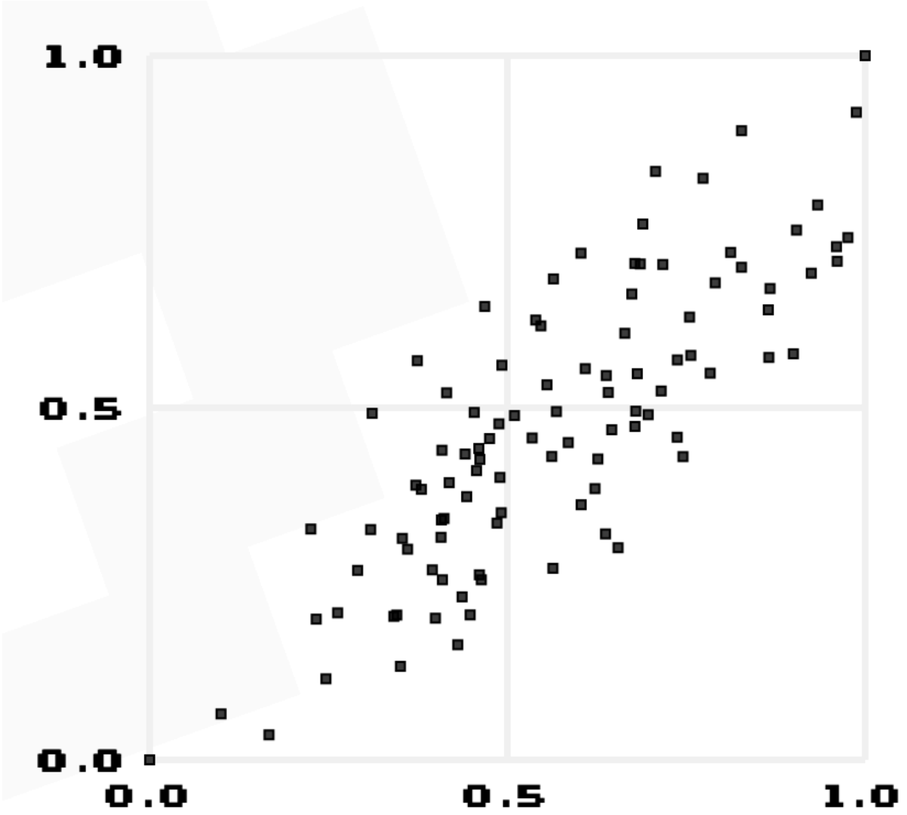


https://www.spss-tutorials.com/pearson-correlation-coefficient/

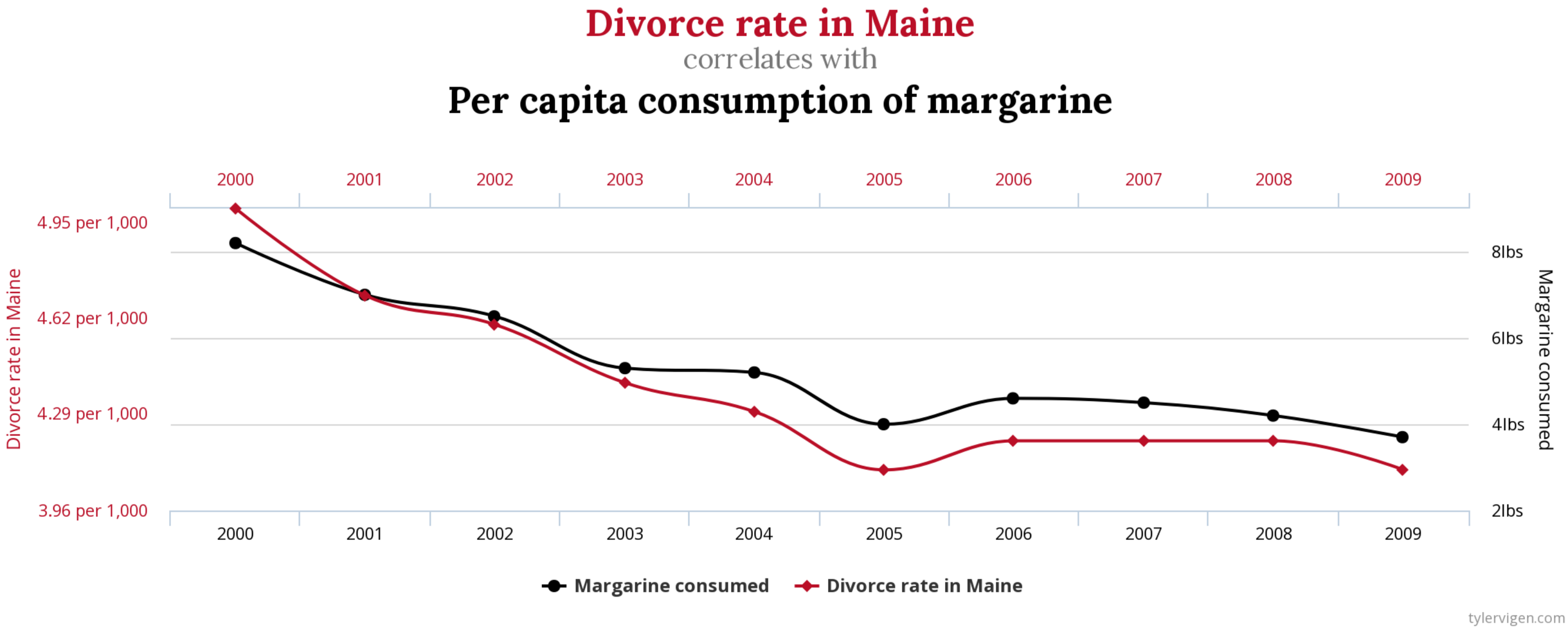

r = 0.7
r = 0.3
r = 0
r = -0.7
r = 0.3
r = 0

https://www.guessthecorrelation.com/

Which of the following is the Pearson correlation coefficient (*r*) for this relationship?



https://forms.gle/M4Q3hFkeoLKCixUX8



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

What about correlation between categorical variables?

- Make a contingency table …

Handed-ness / Sex	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- Then calculate phi (based on chi-squared statistic) …

Phi coefficient [edit]

Main article: Phi coefficient

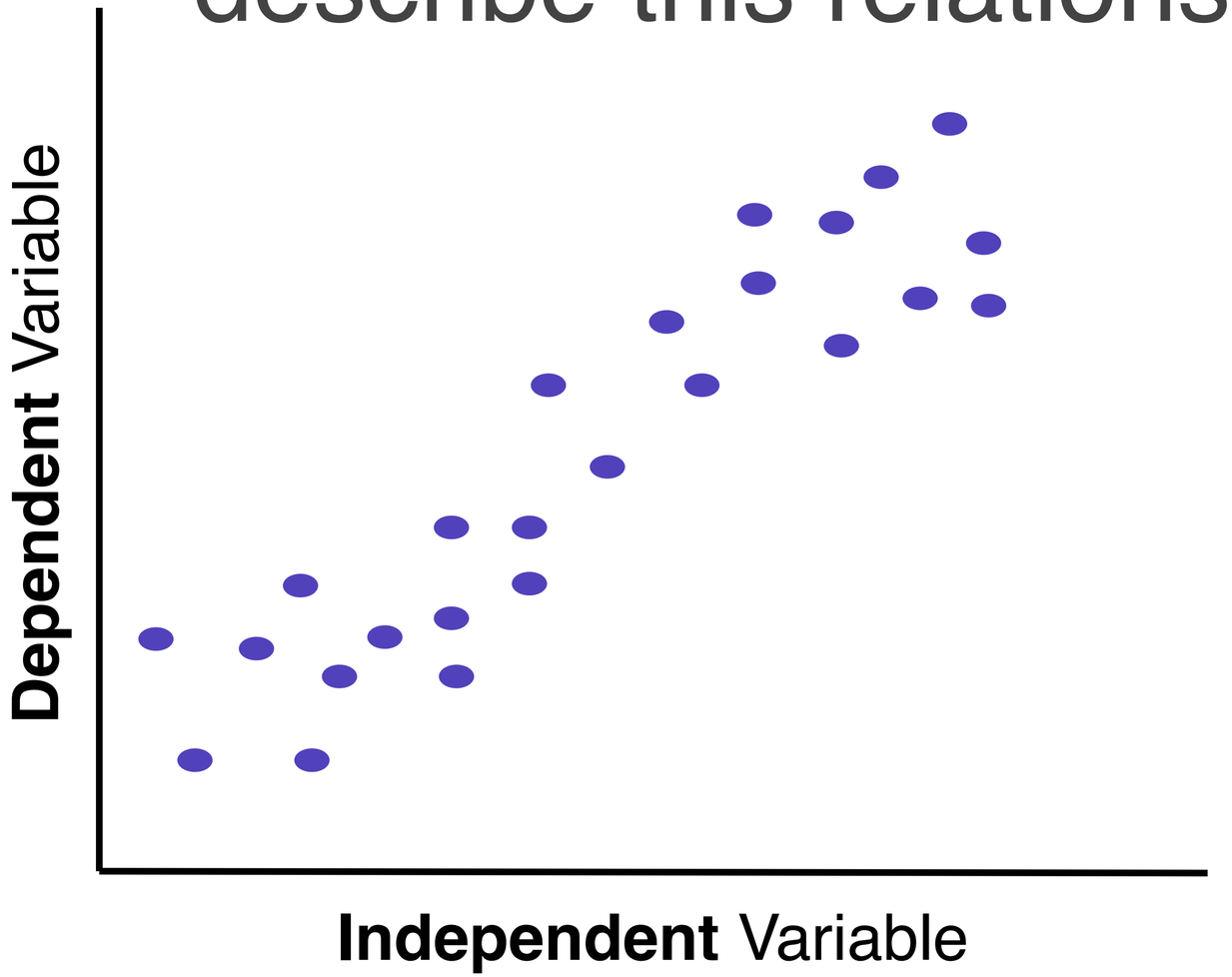
A simple measure, applicable only to the case of 2 × 2 contingency tables, is the phi coefficient (φ) defined by

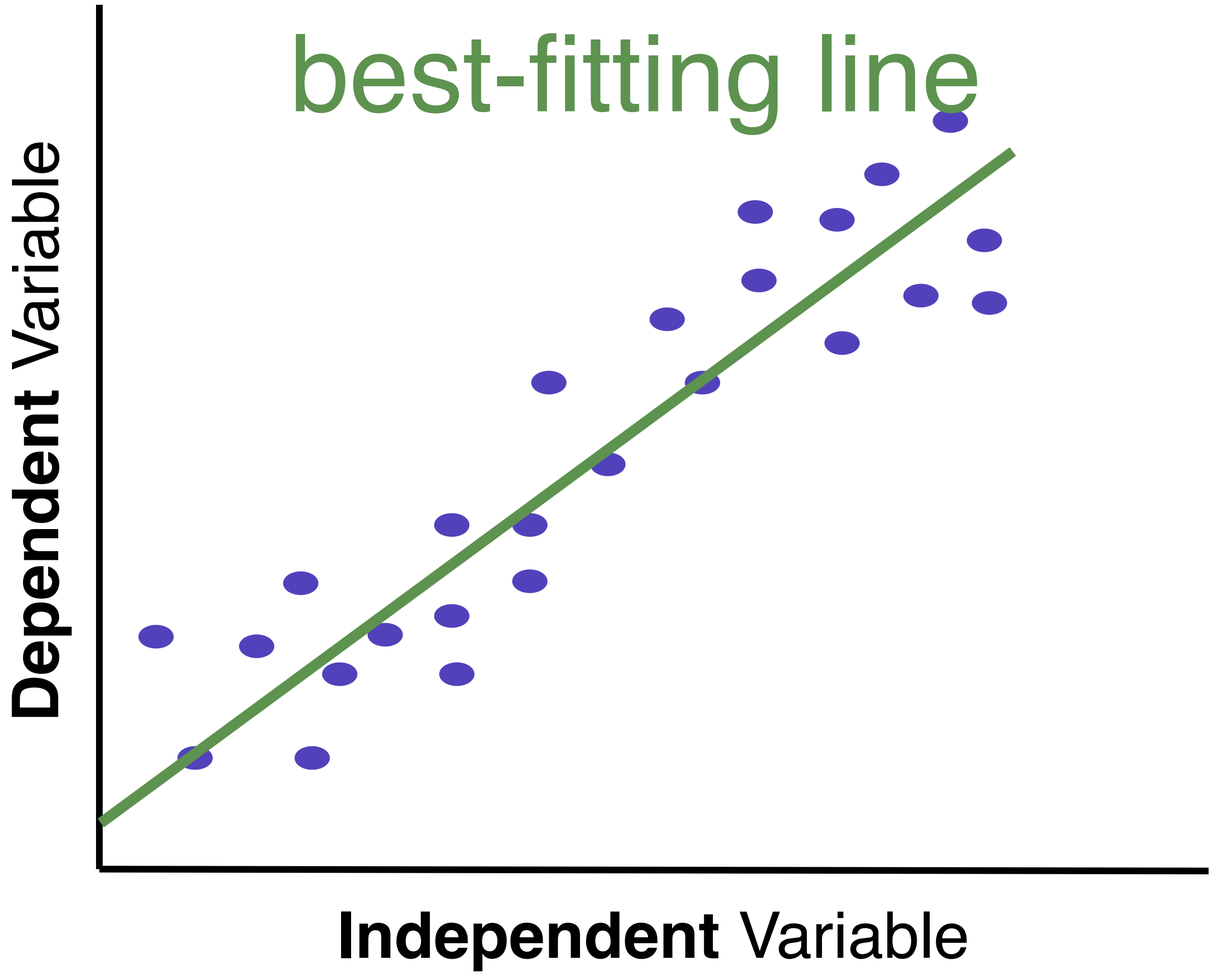
$$\phi = \pm\sqrt{\frac{\chi^2}{N}},$$

where χ^2 is computed as in Pearson's chi-squared test, and N is the grand total of observations. φ varies from 0 (corresponding to no association between the variables) to 1 or −1 (complete association or complete inverse association), provided it is based on frequency data represented in 2 × 2 tables. Then its sign equals the sign of the product of the main diagonal elements of the table minus the product of the off–diagonal elements. φ takes on the minimum value −1.0 or the maximum value of +1.0 if and only if every marginal proportion is equal to 0.5 (and two diagonal cells are empty).[2]

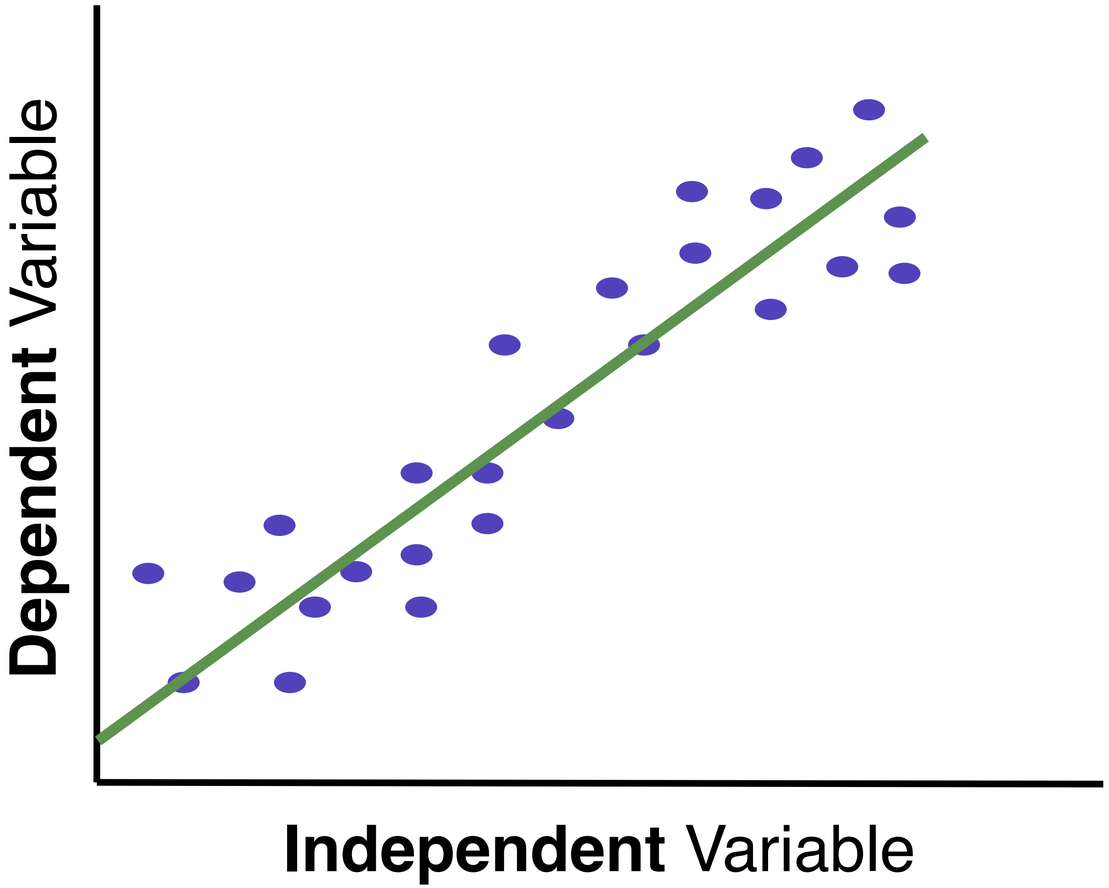
- Yes there are ways to do this in pandas/numpy and other libraries!

Linear regression can be used to describe this relationship

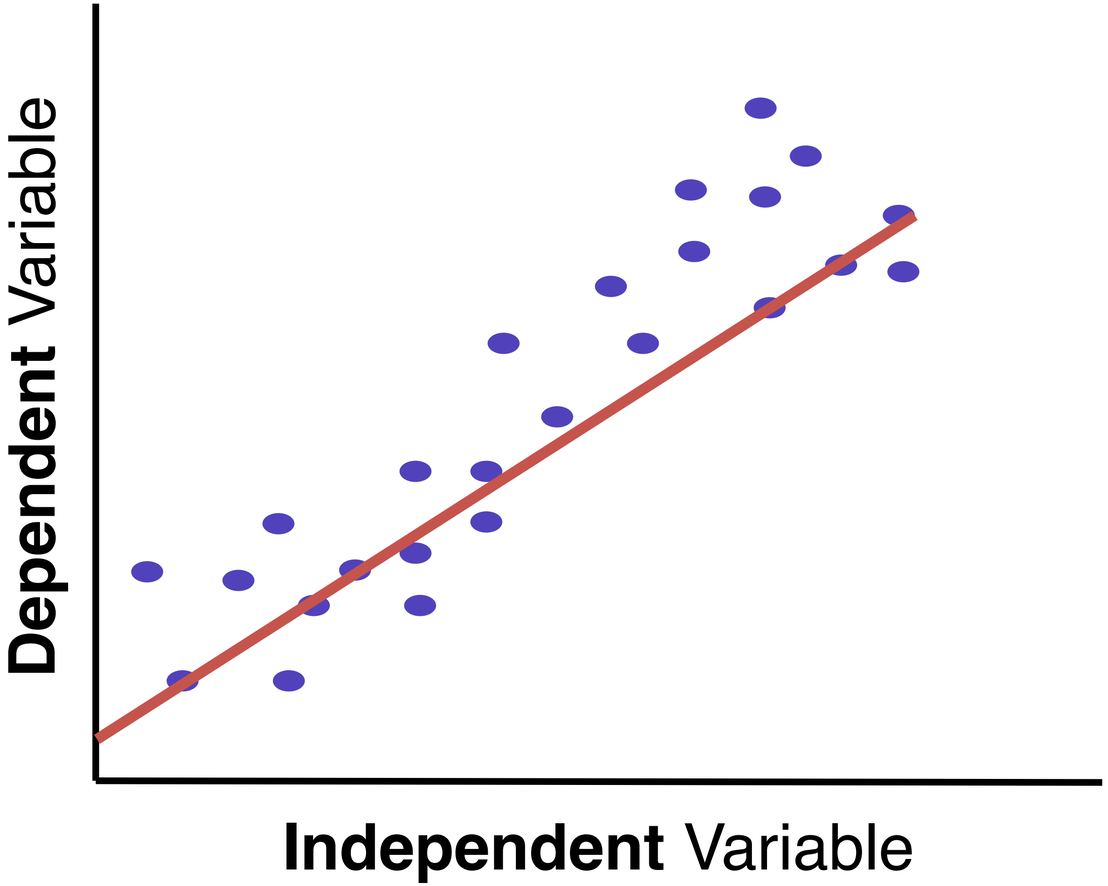


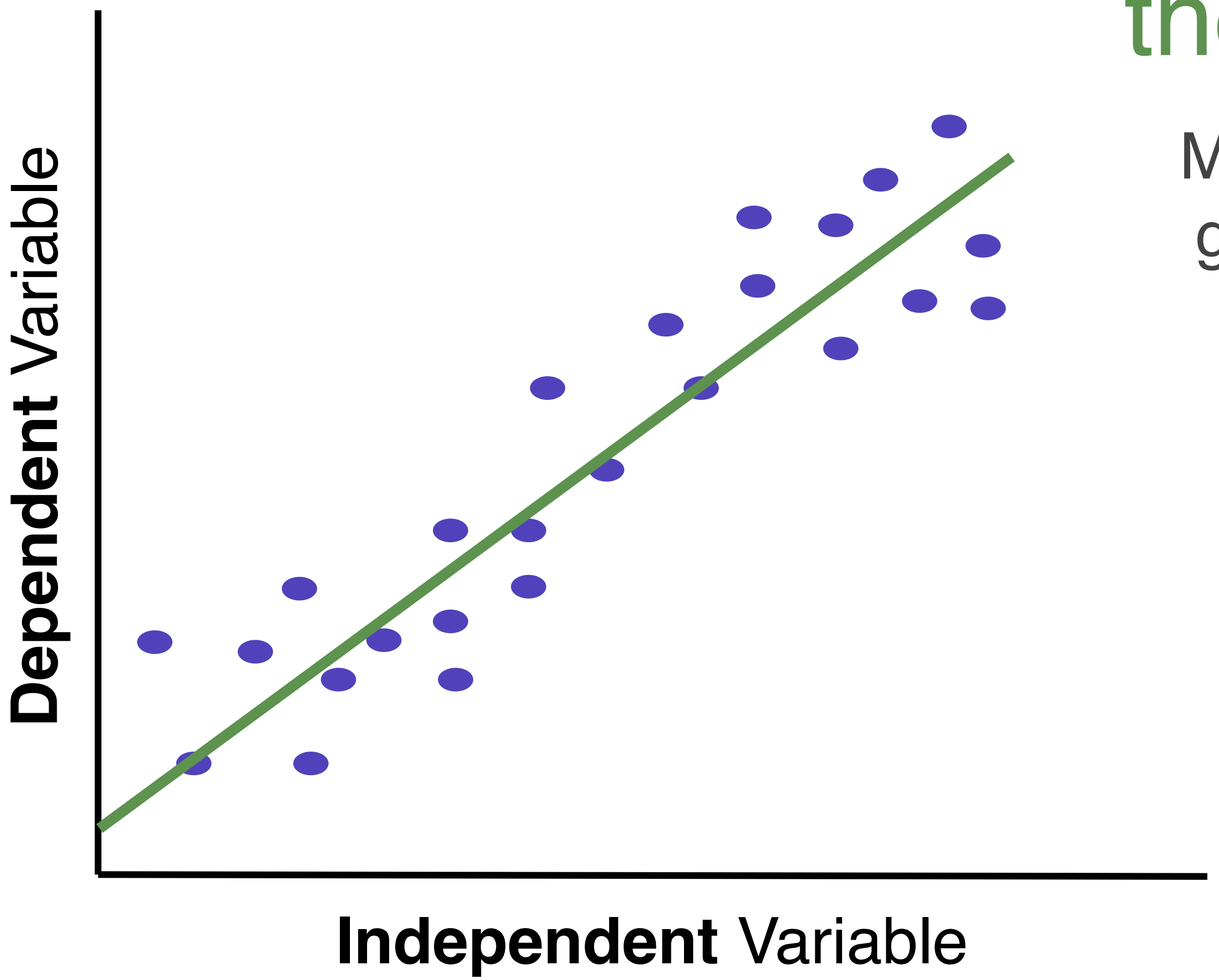

best-fitting line
Dependent Variable
Independent Variable

Best-fitting line



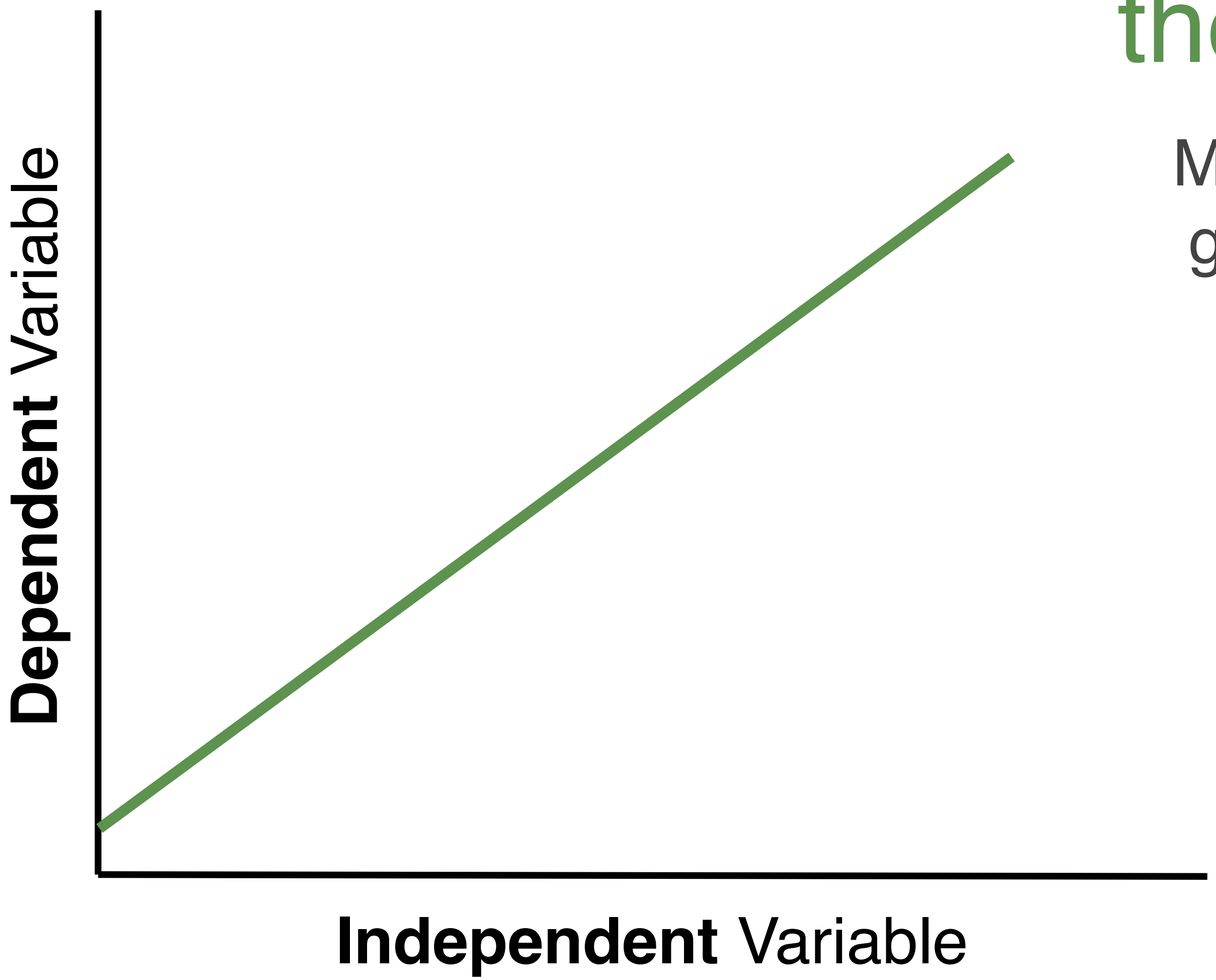
NOT a best-fitting line





This line is a model of the data

Models are mathematical equations generated to *represent* the real life situation



This line is a model of the data

Models are mathematical equations generated to *represent* the real life situation

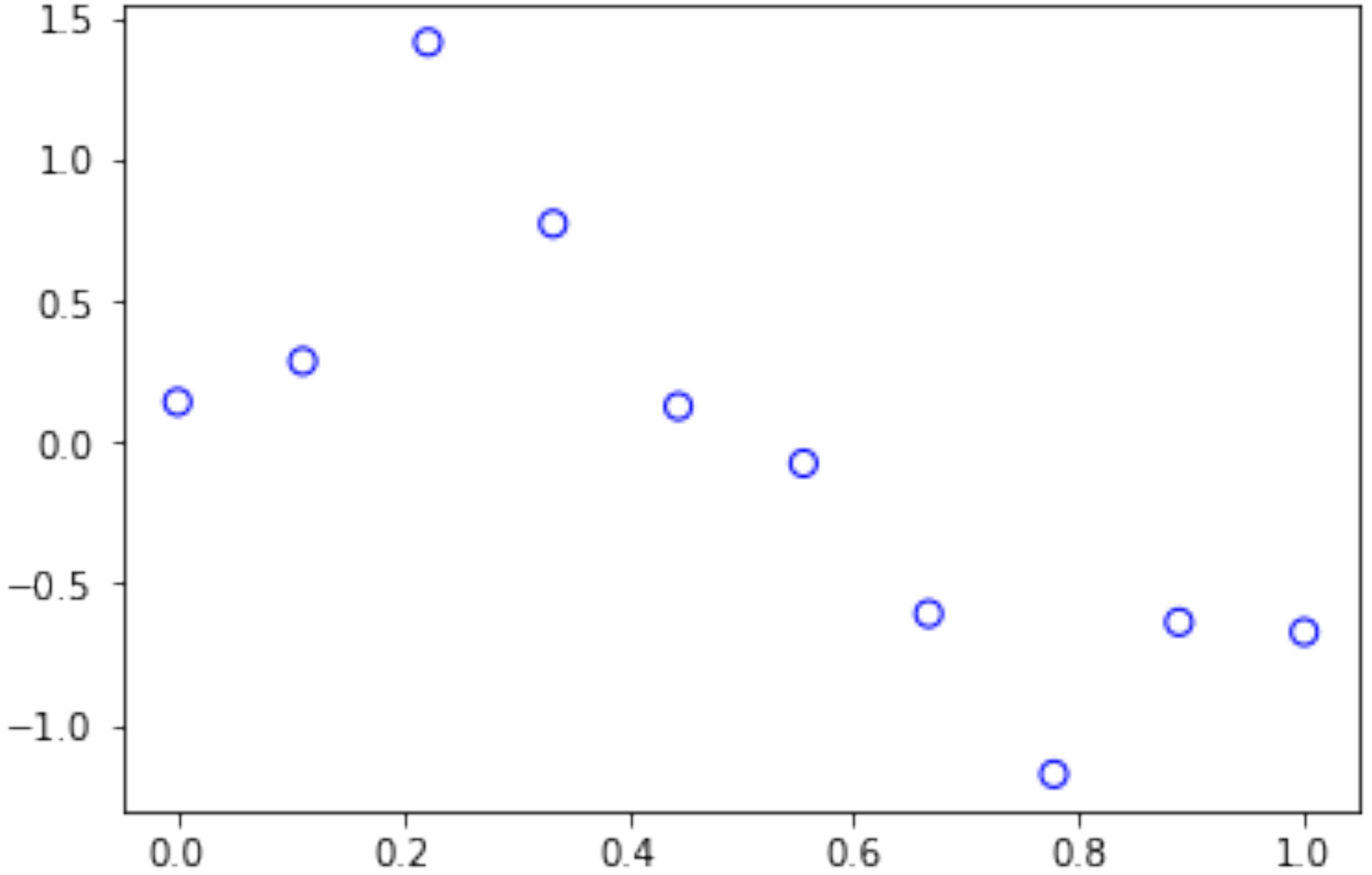
2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

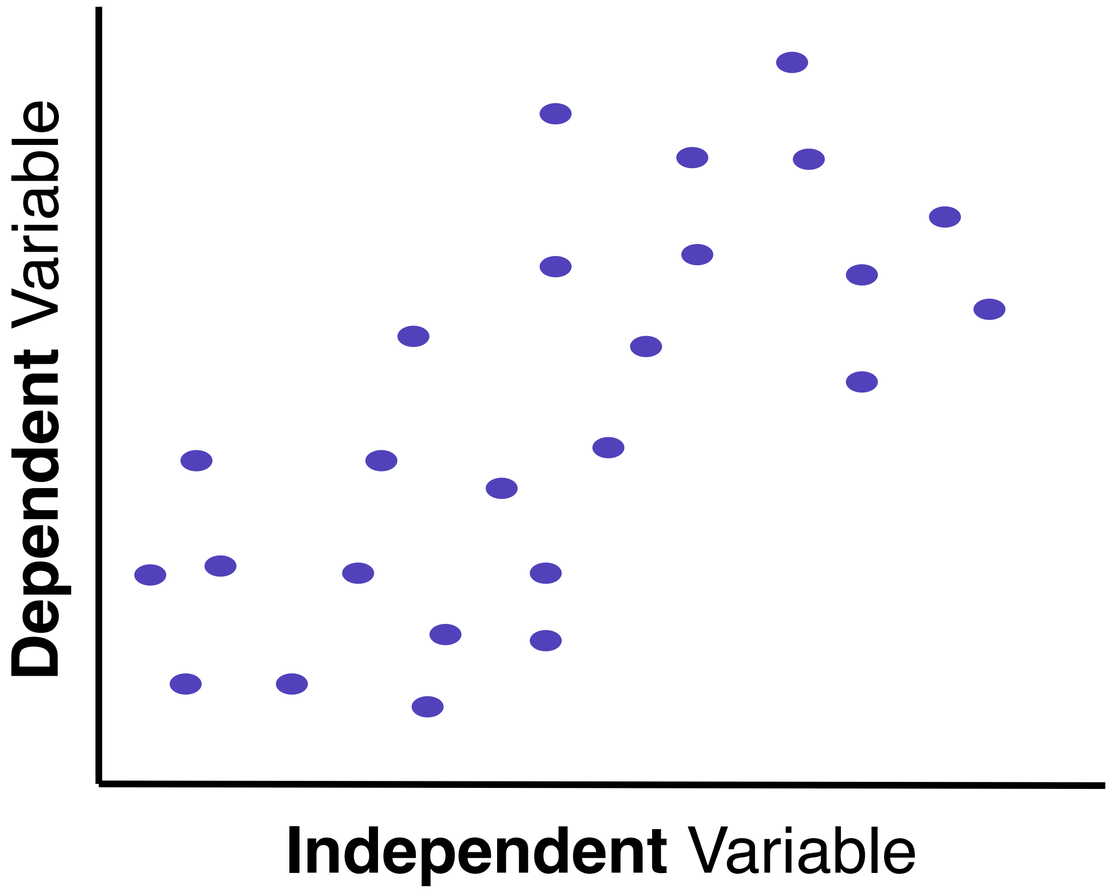
2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

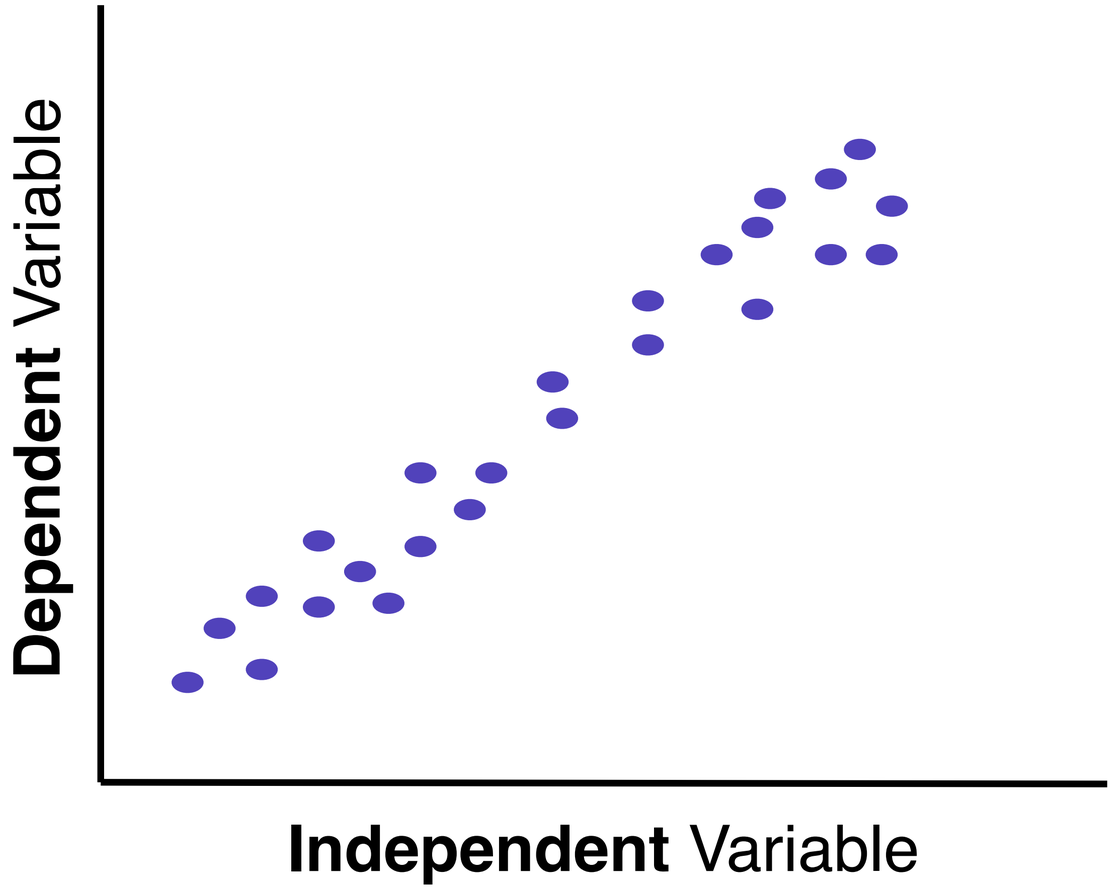
George Box


1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0

weaker relationship



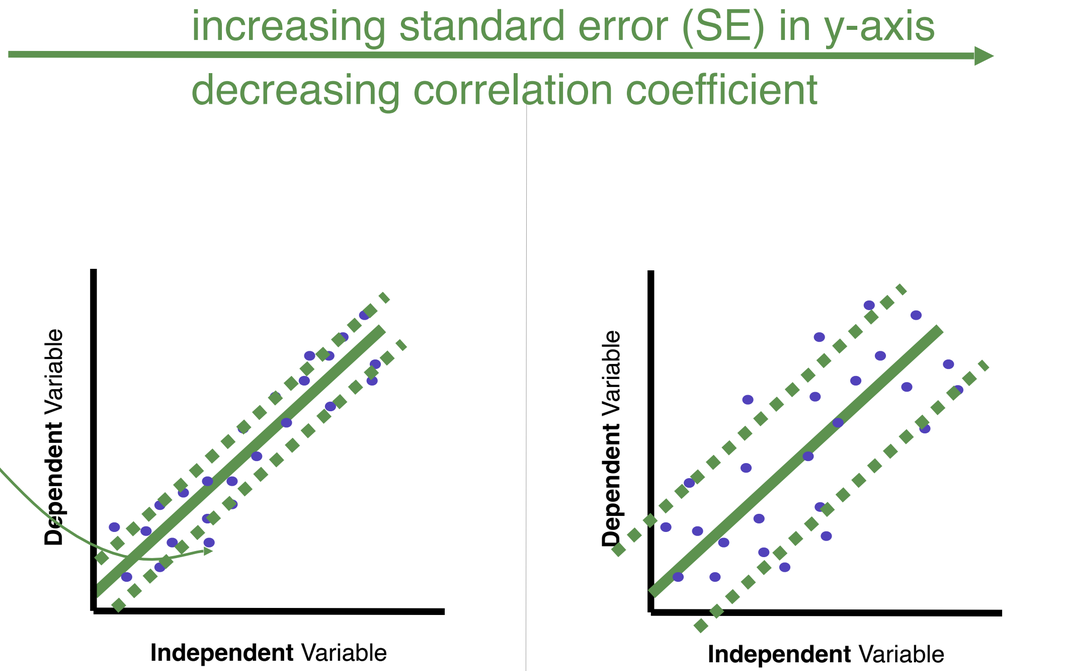
stronger relationship



stronger relationship = higher correlation

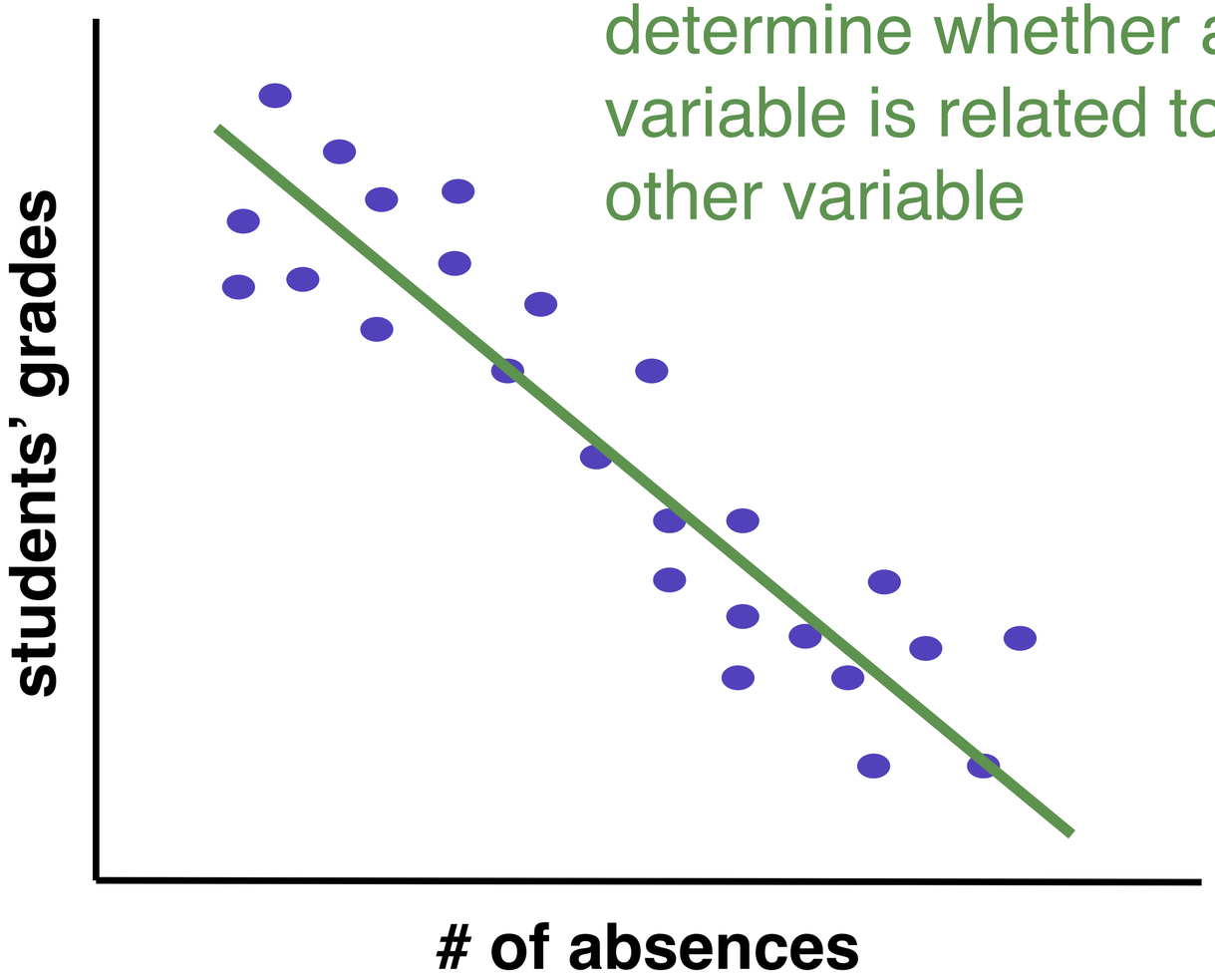
This is a kind of effect size

The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

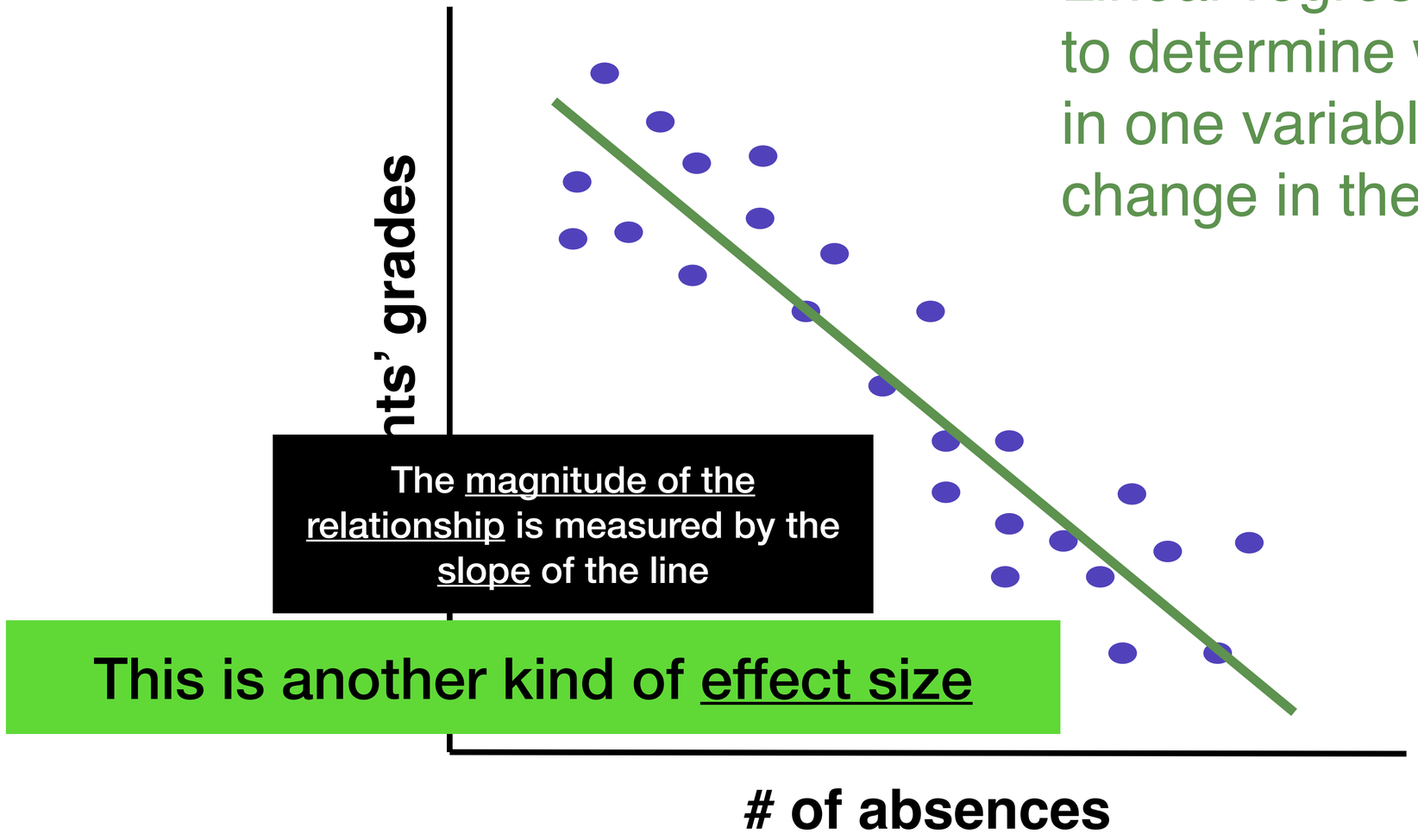


Standard error is standard deviation / sqrt(n)

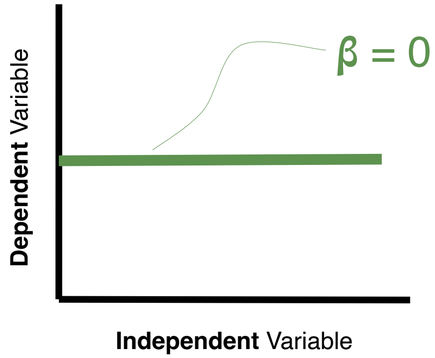
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



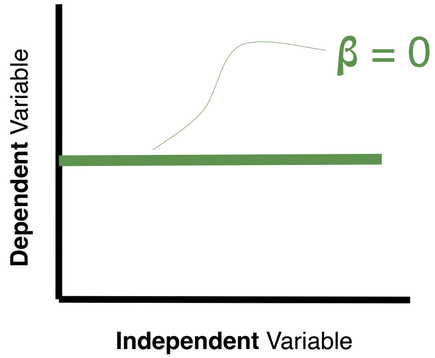
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable

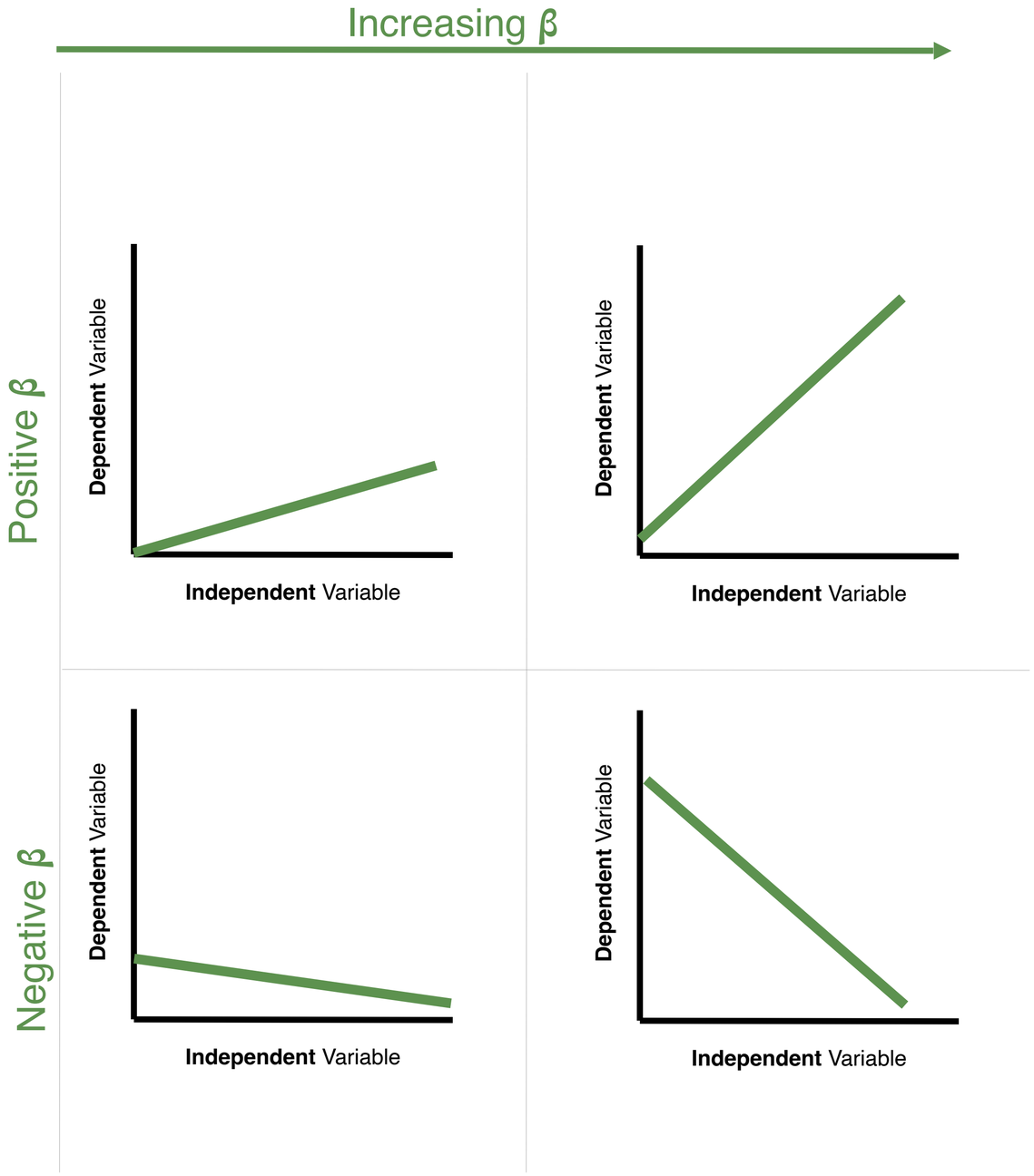


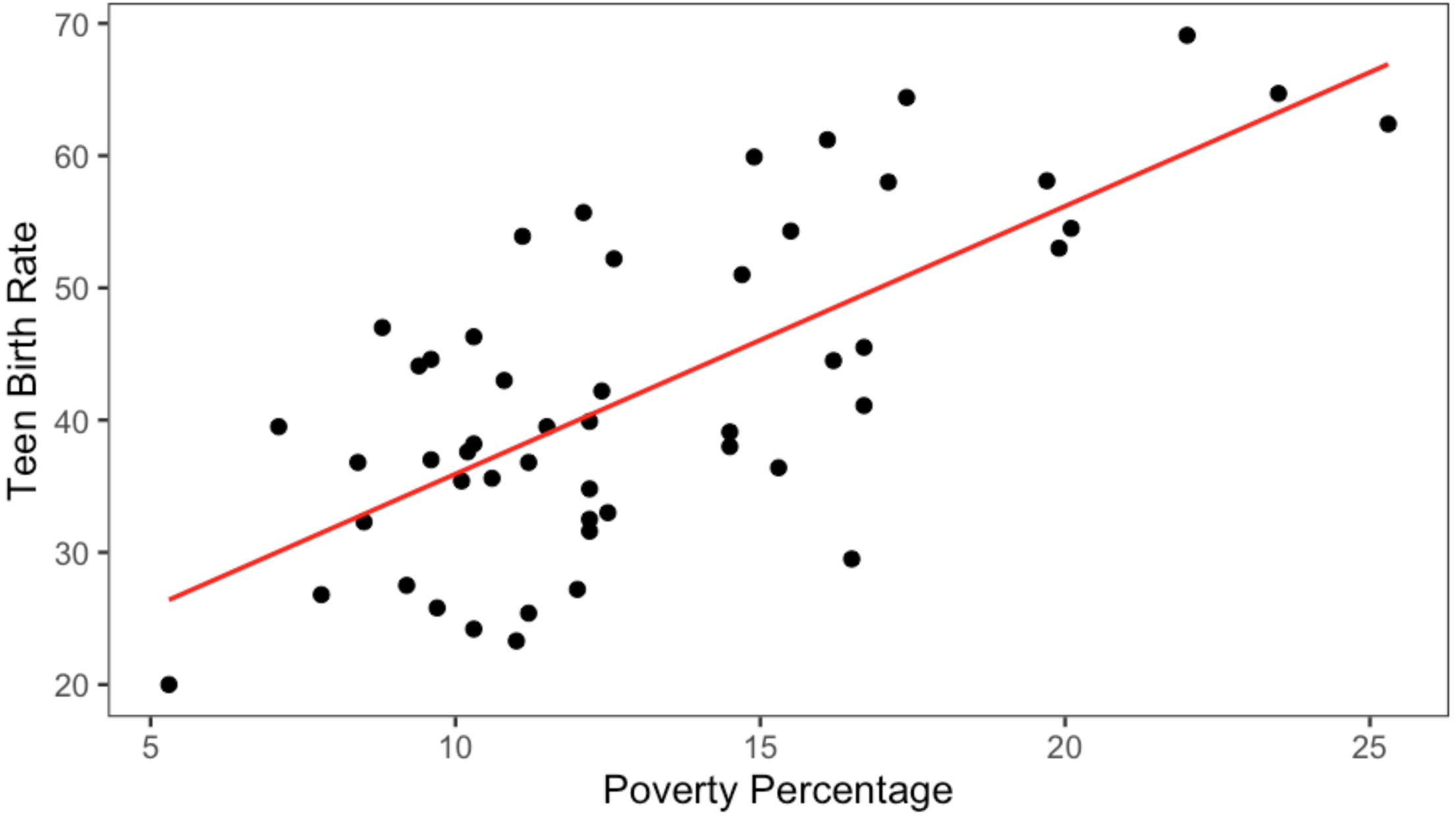
Effect size (β) can be estimated using the slope of the line

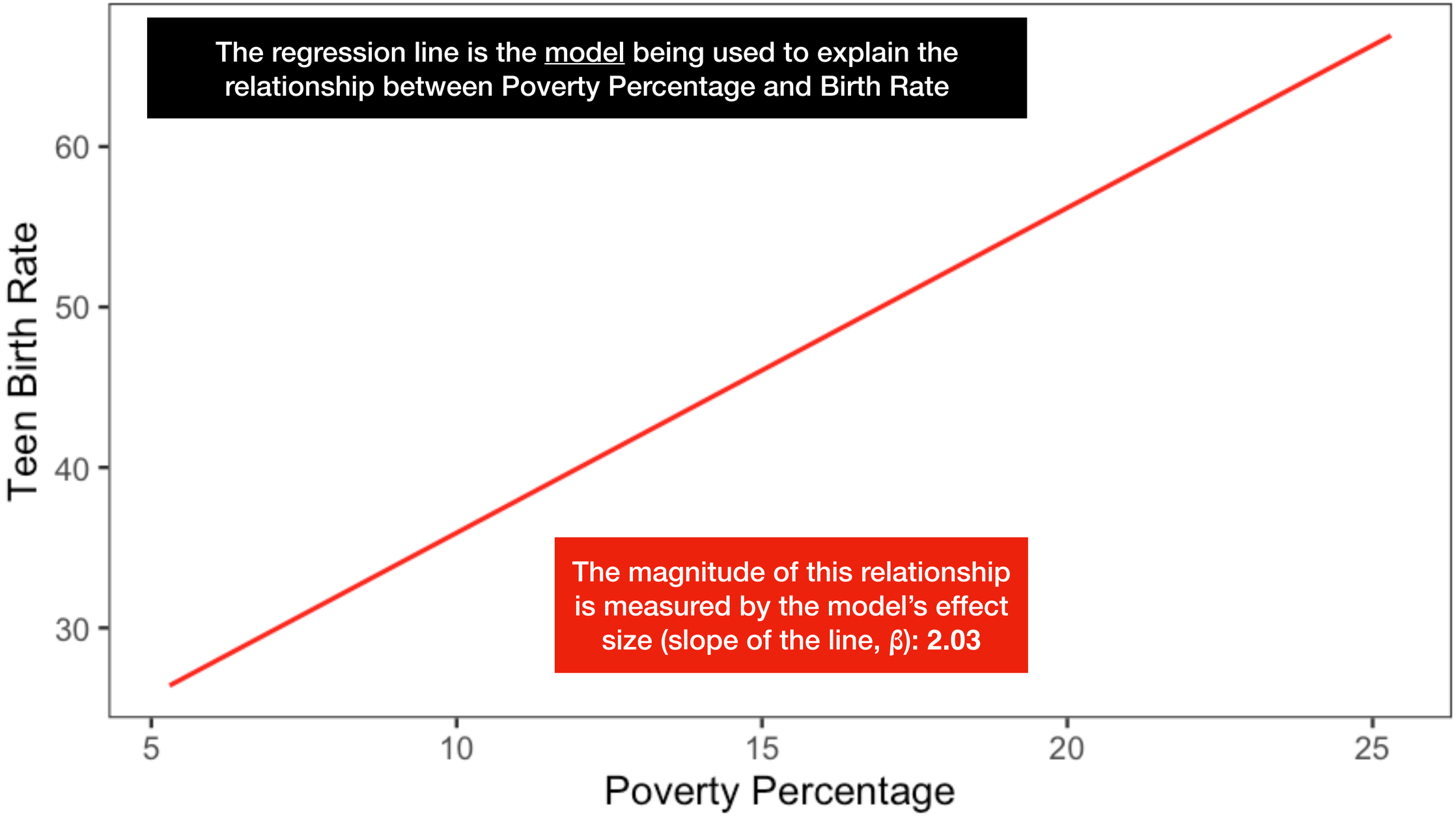


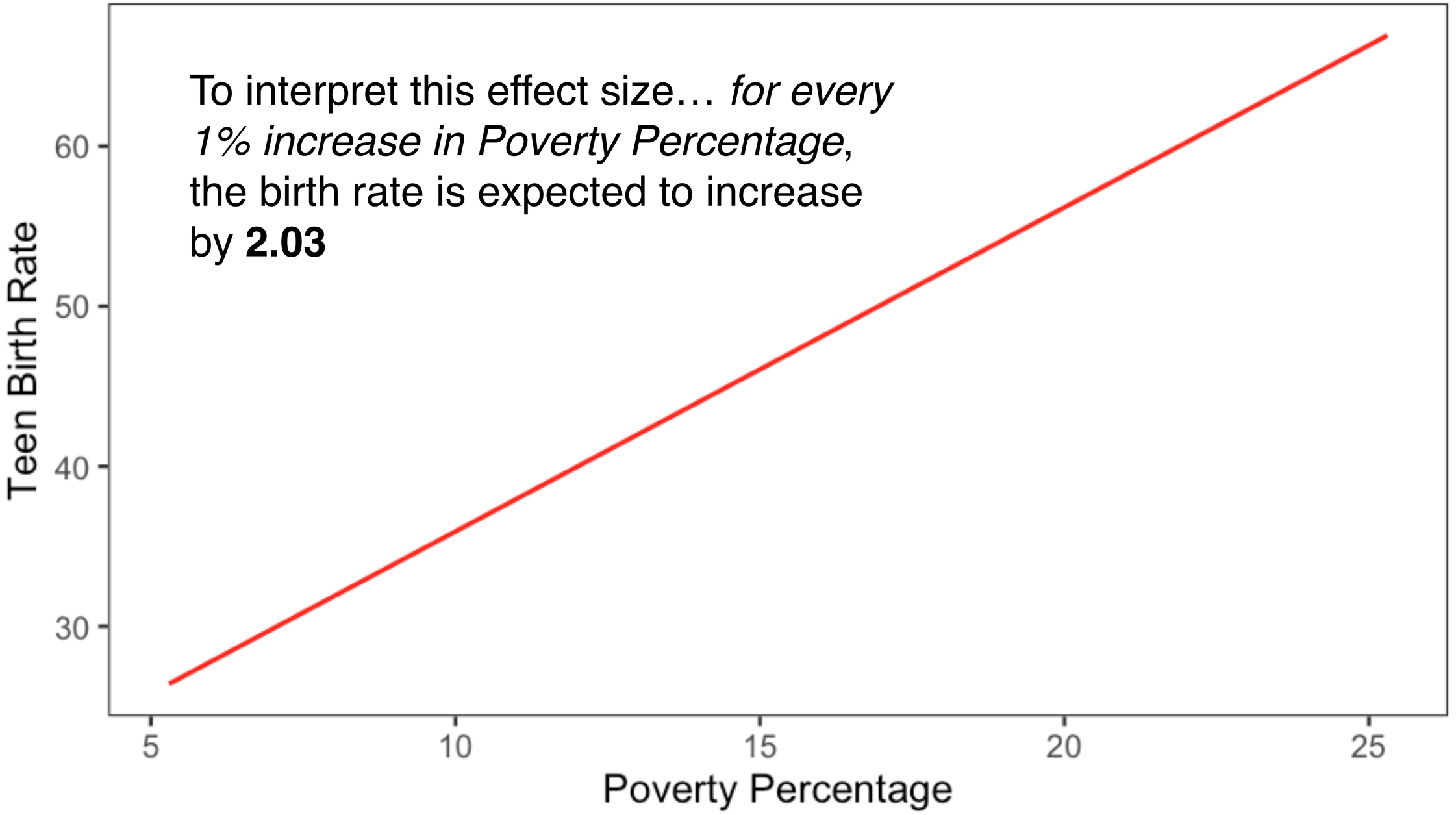
Effect size (β) can be estimated using the slope of the line





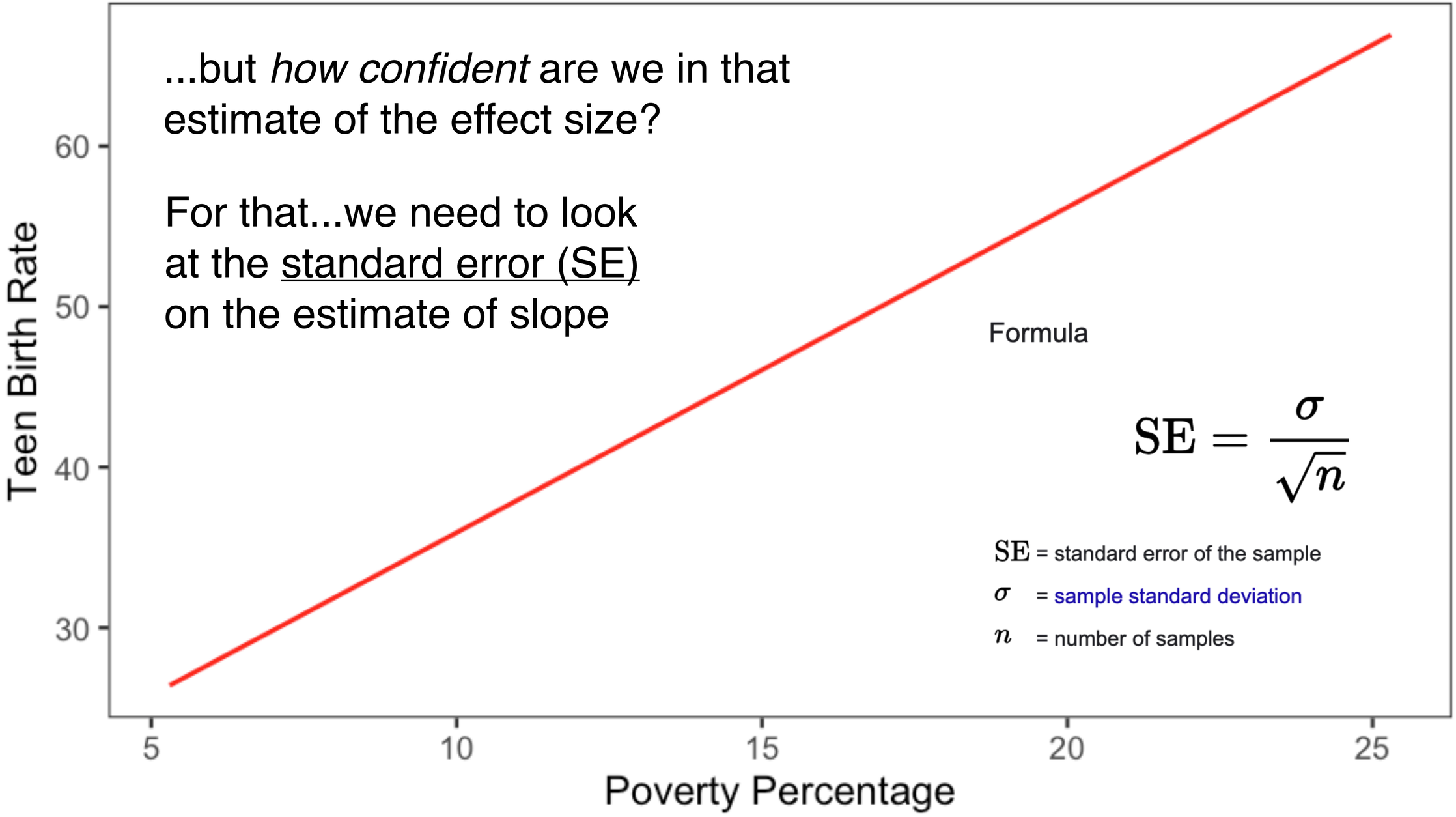

Teen Birth Rate
70
60
50
40
30
20
5
10
15
20
25
Poverty Percentage

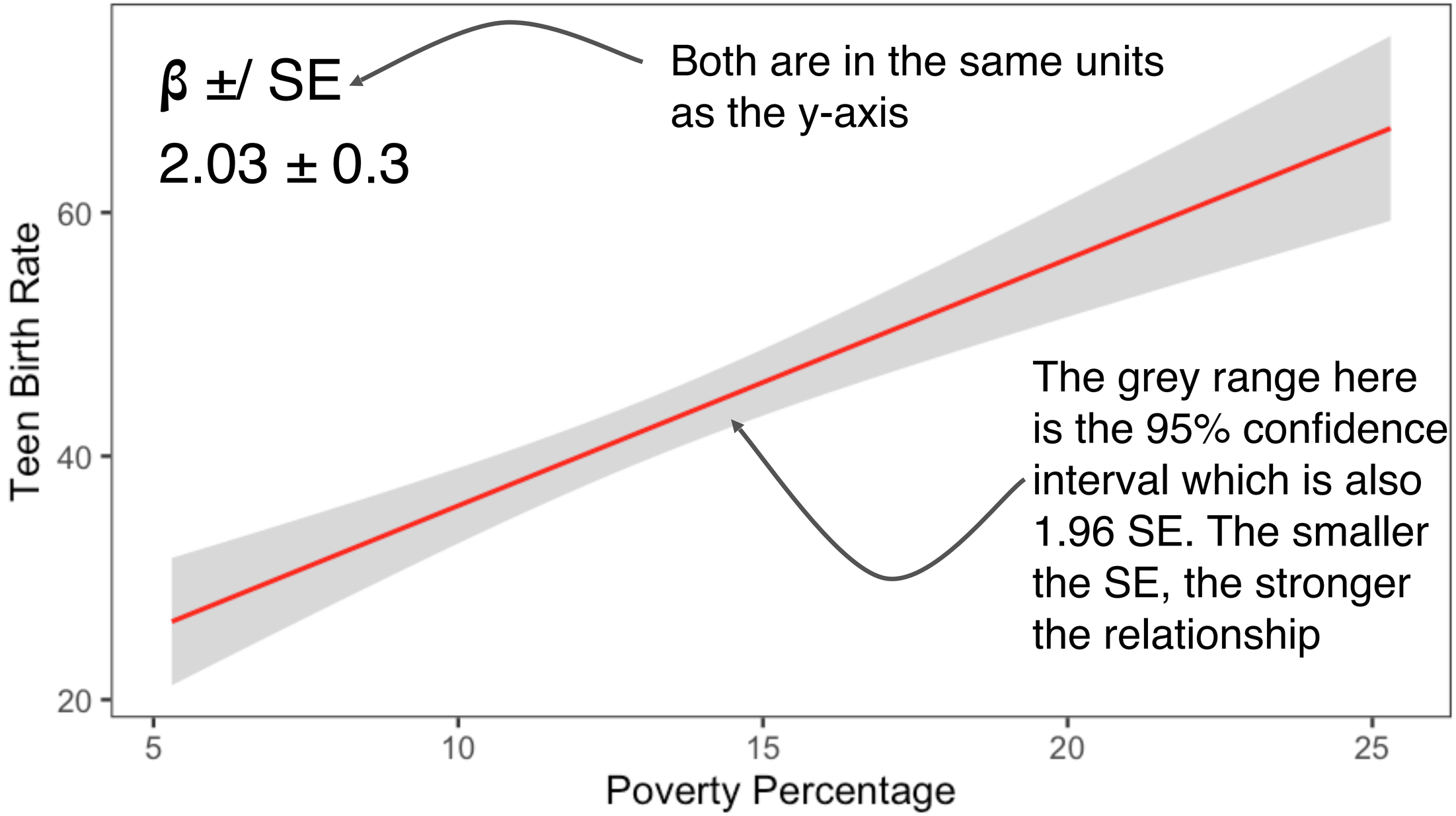

The regression line is the model being used to explain the relationship between Poverty Percentage and Birth Rate
The magnitude of this relationship is measured by the model's effect size (slope of the line, β): 2.03
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage


To interpret this effect size… *for every 1% increase in Poverty Percentage*, the birth rate is expected to increase by **2.03**
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage

...but *how confident* are we in that estimate of the effect size?

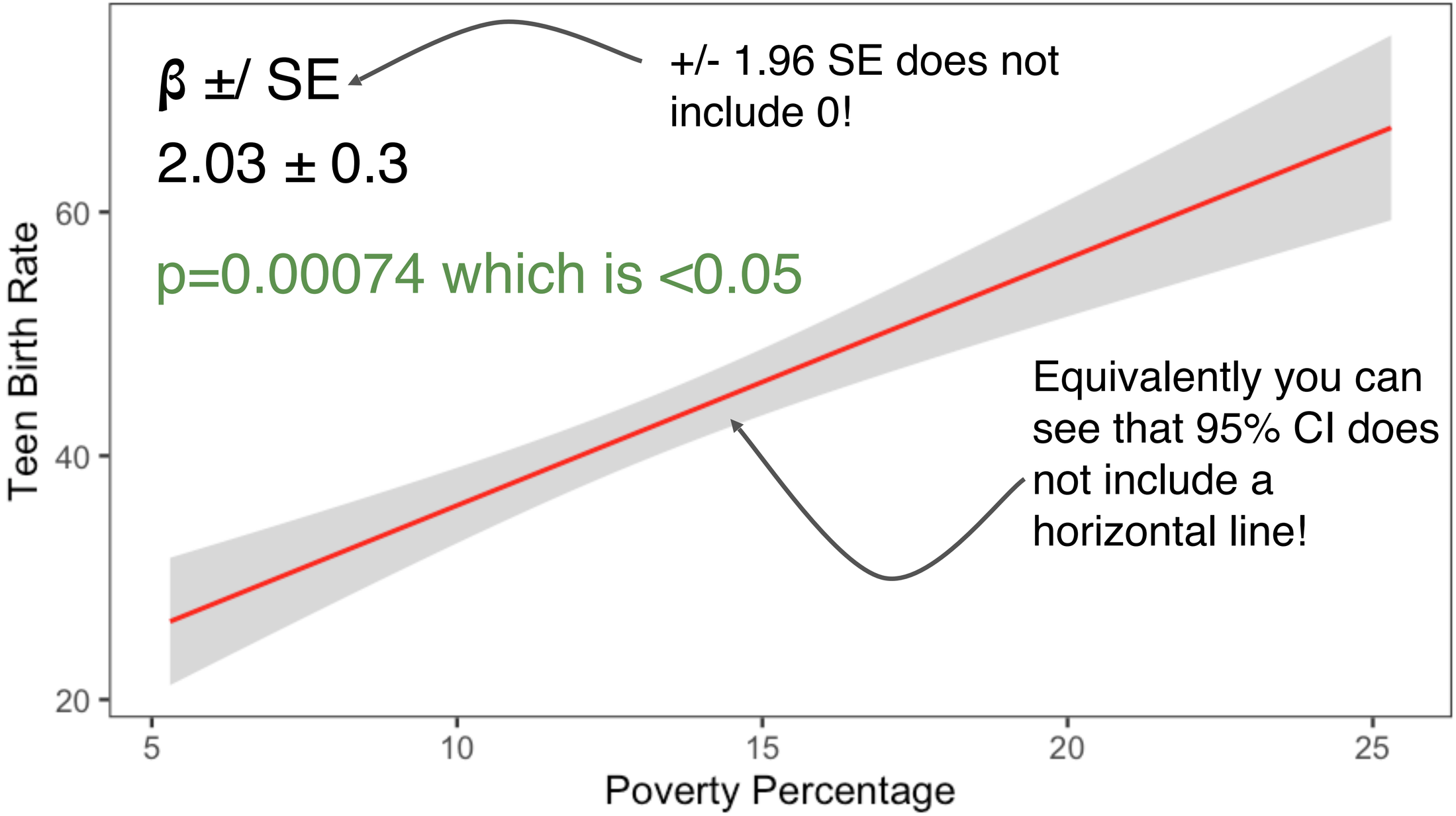
For that...we need to look at the standard error (SE) on the estimate of slope




β ±/ SE
2.03 ± 0.3
Both are in the same units as the y-axis
The grey range here is the 95% confidence interval which is also 1.96 SE. The smaller the SE, the stronger the relationship
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage

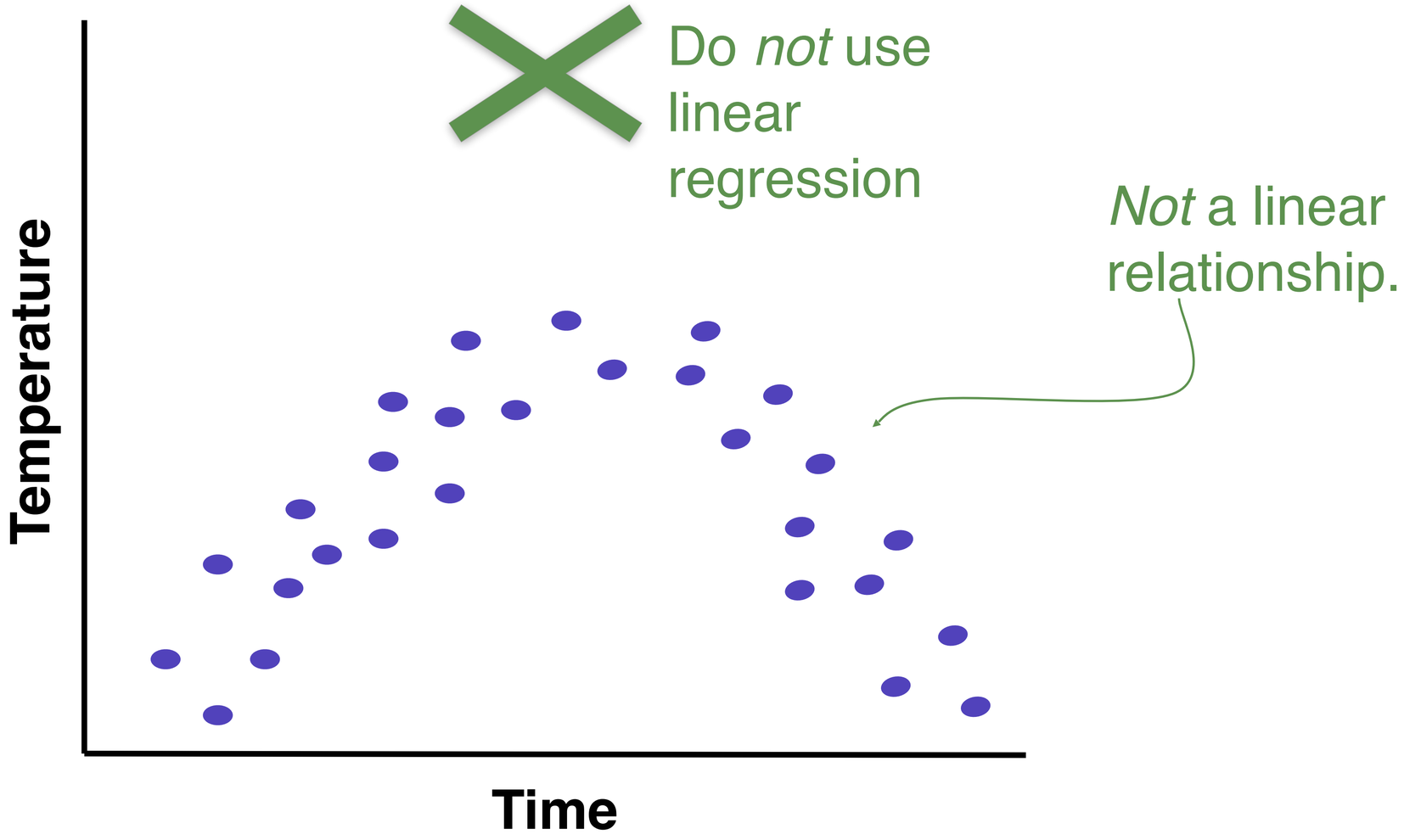
p-value : the probability of getting the observed results (or results more extreme) by chance alone

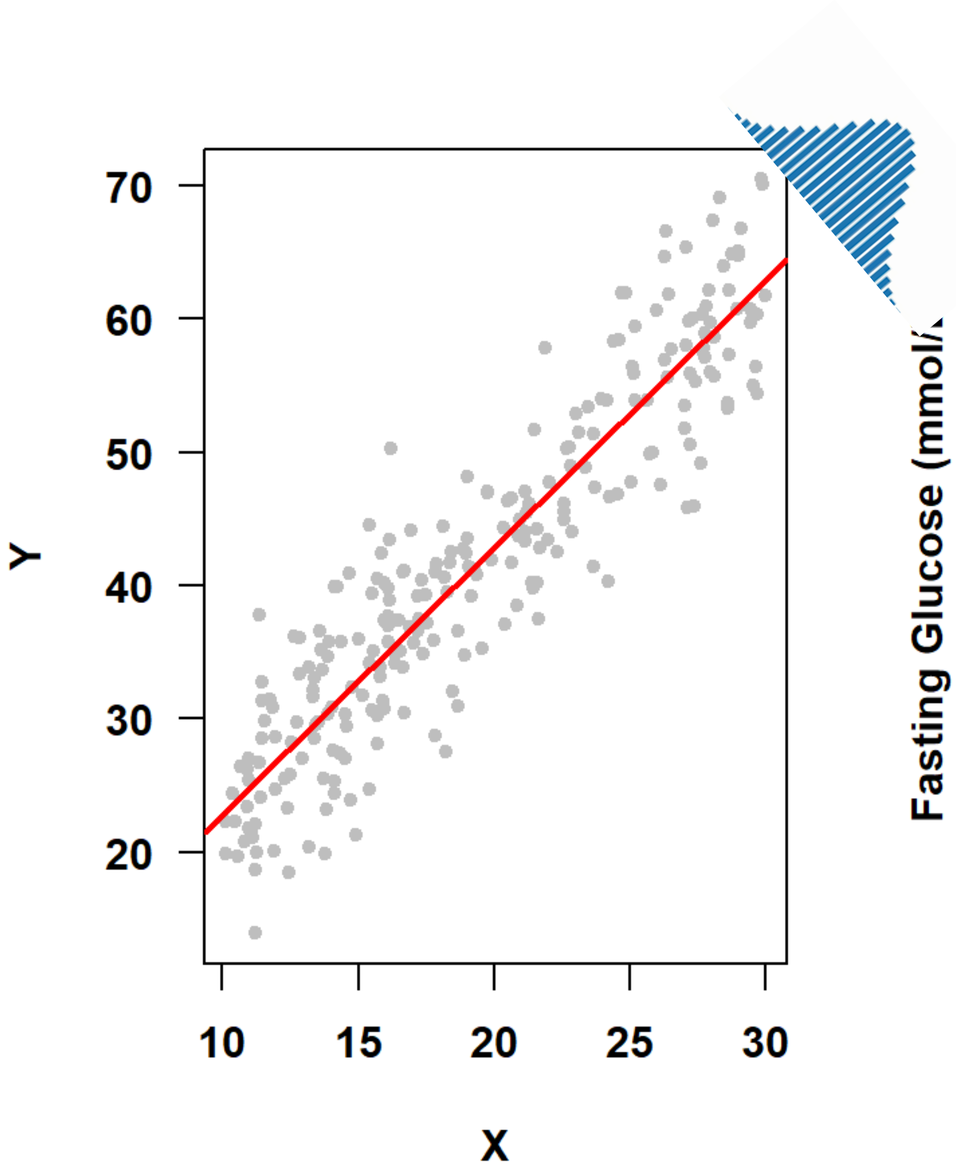
Takes into account the effect size (β) and the SE

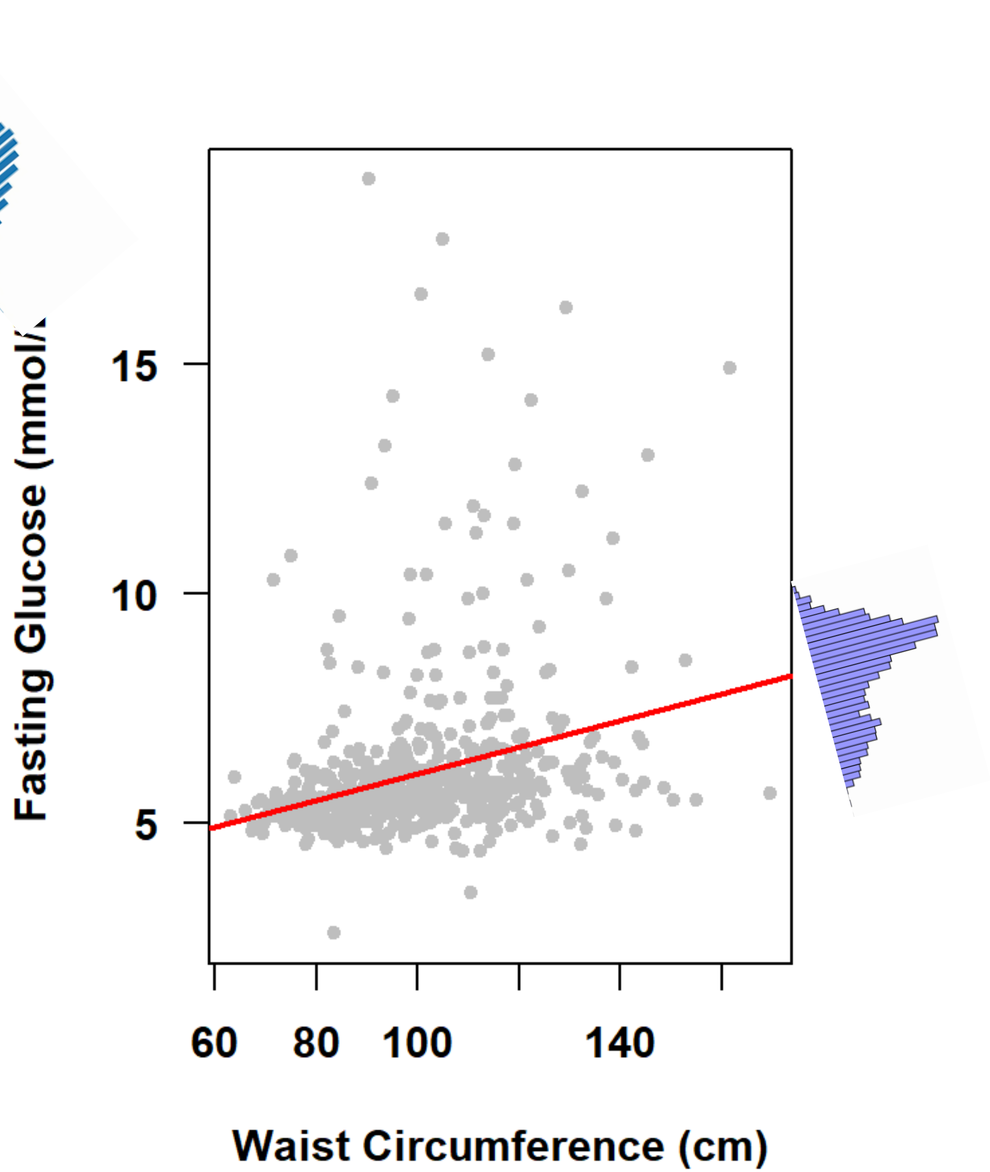

β ±/ SE
2.03 ± 0.3
+/- 1.96 SE does not include 0!
p=0.00074 which is <0.05
Equivalently you can see that 95% CI does not include a horizontal line!
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage

Assumptions of linear regression using Ordinary Least Squares method

1. Linear in parameters
2. Normality of residuals
3. No multicollinearity
4. No autocorrelation
5. Homoscedasticity


Do *not* use linear regression
Not a linear relationship.
Temperature
Time

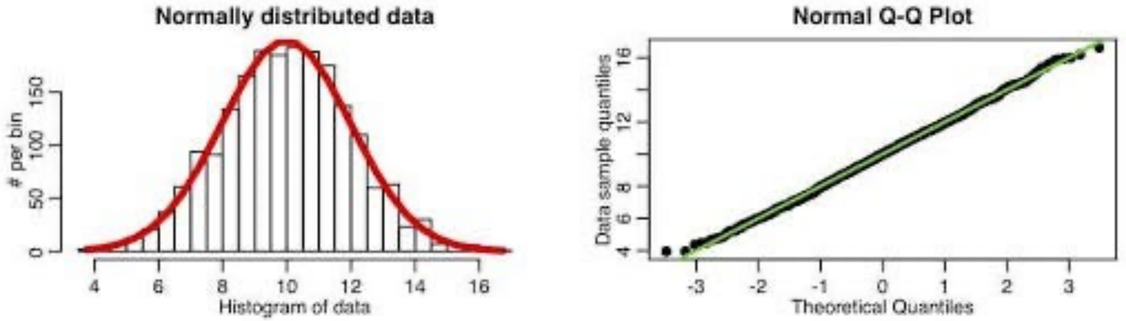


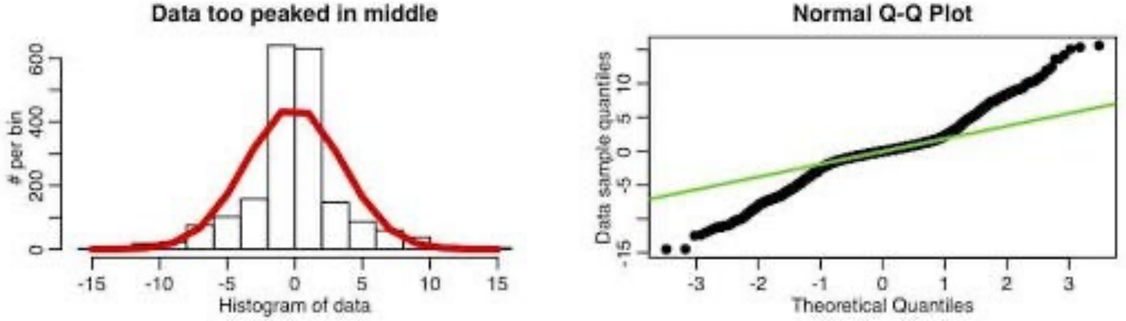


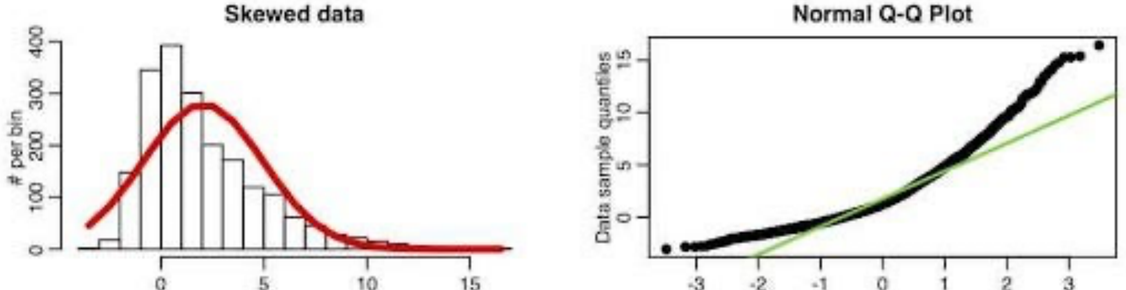
Non-normal residuals

QQ plot

Theoretical normal distribution quantiles vs sample quantiles

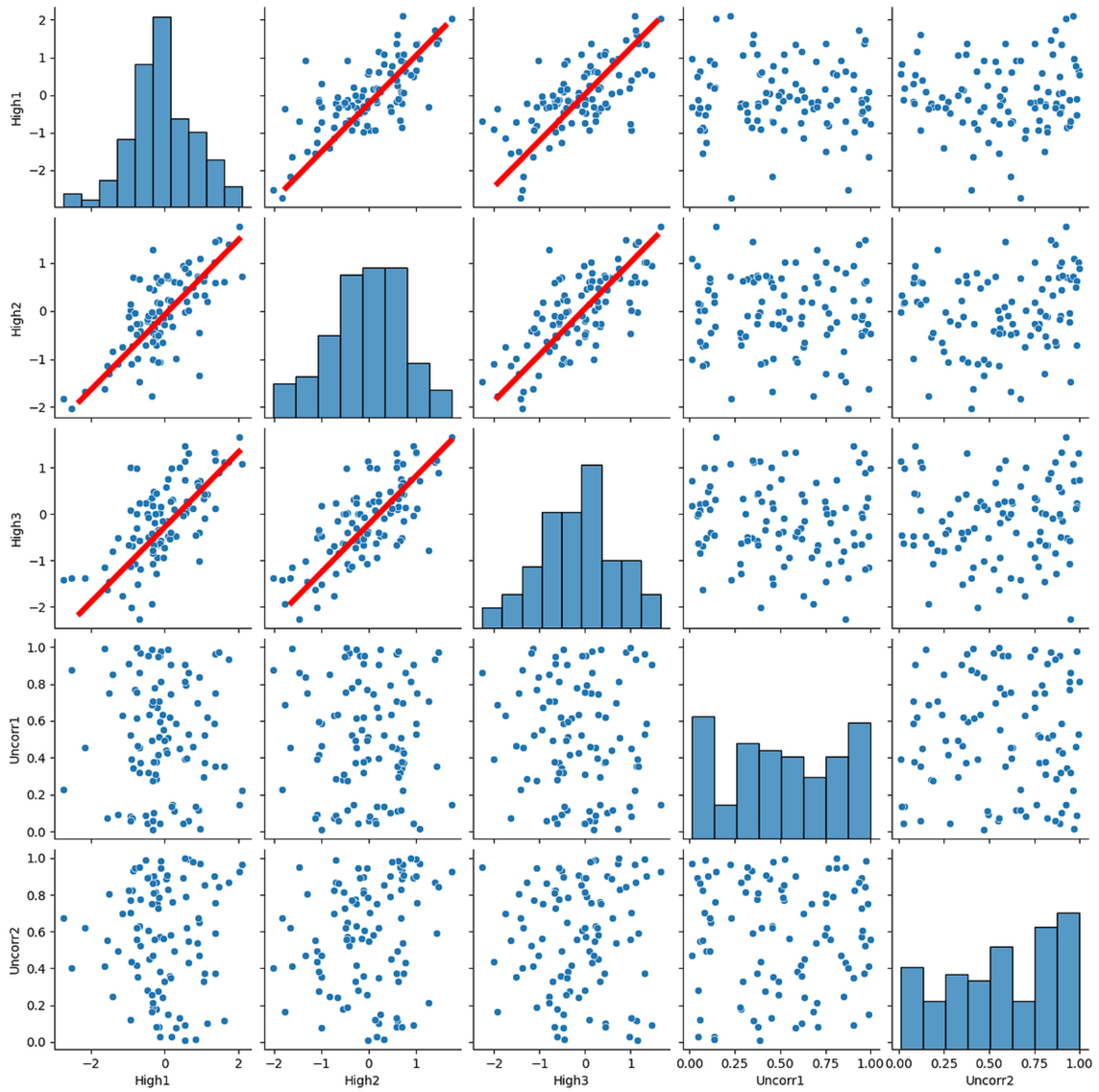




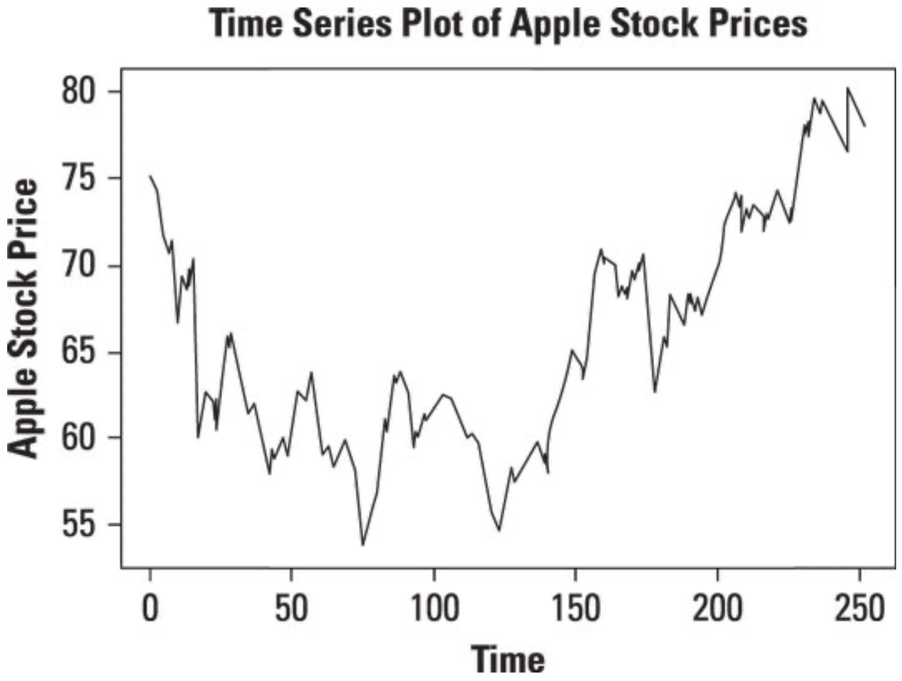


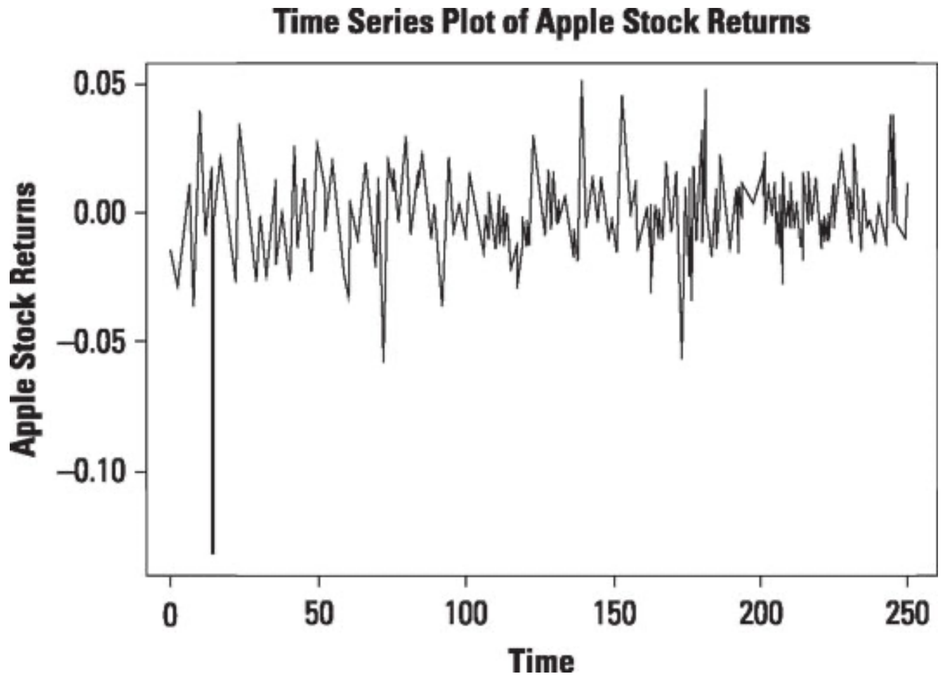
Multicollinearity

occurs when one or more independent variables (in multiple linear regression) are too highly correlated with each other.



Daily returns are
$\ln(price_t / price_{t-1})$

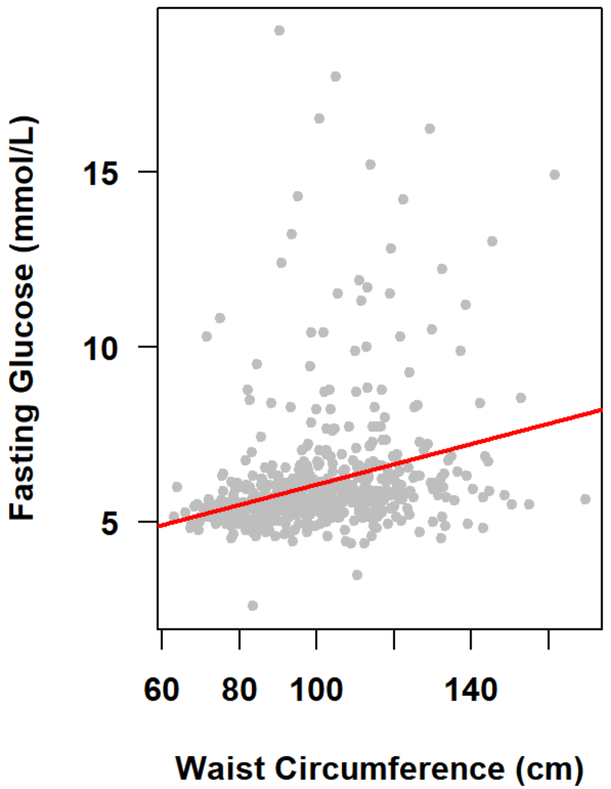


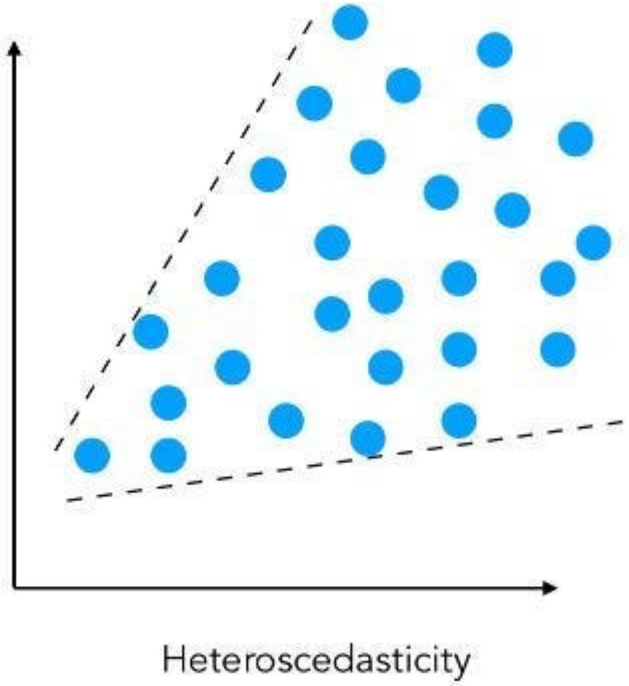


Autocorrelation occurs when the observations are *not* independent of one another (i.e. stock prices)

https://www.dummies.com/programming/big-data/data-science/autocorrelation-plots-graphical-technique-for-statistical-data//

For heteroskedasticity you can’t use OLS to fit a regression on the raw data



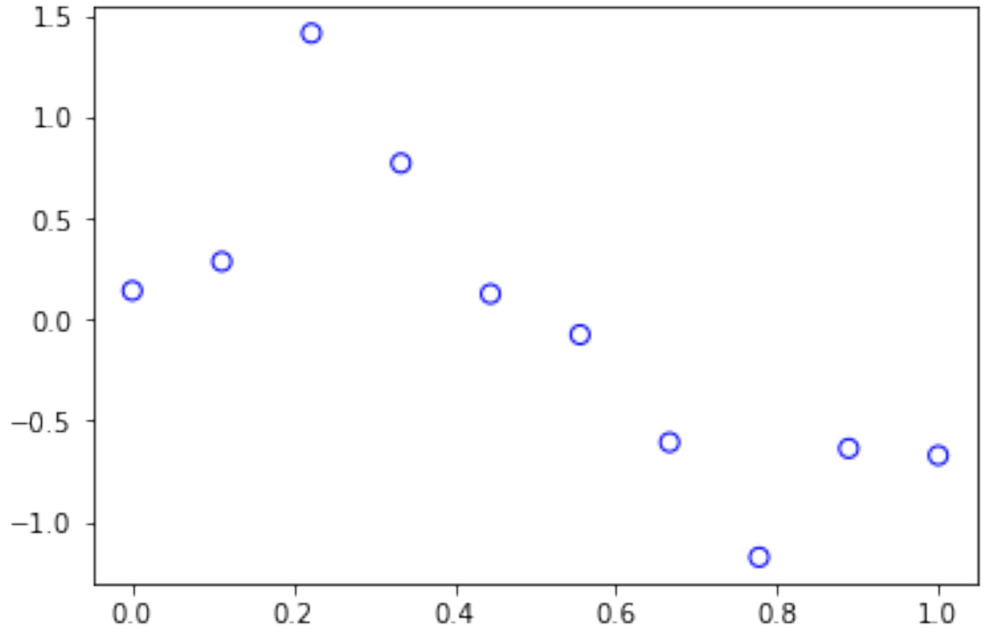


Options:

- Robust regression
- Piece-wise linear
- Transform the data

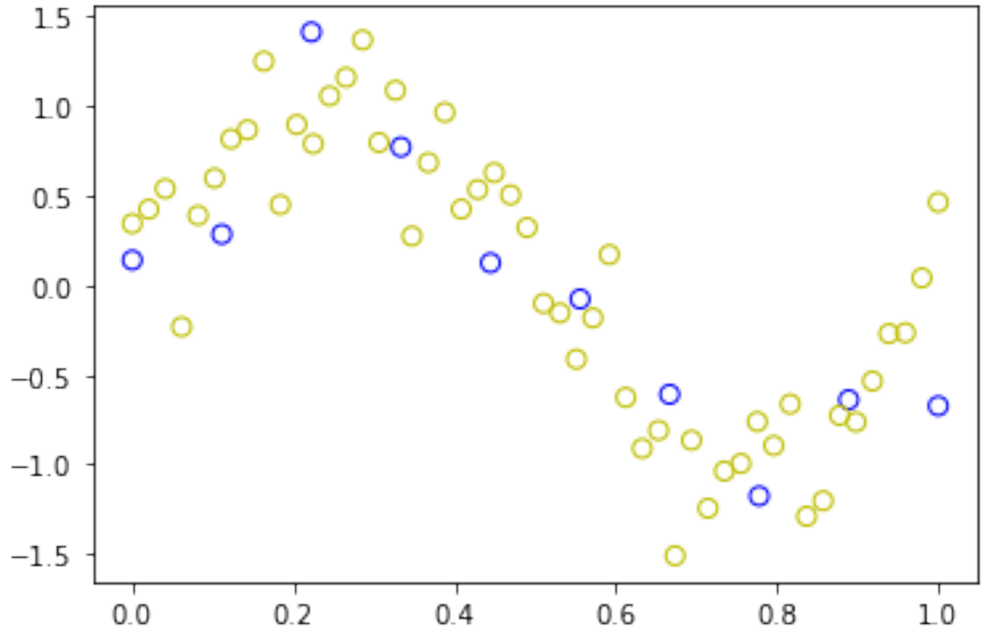
OLS can fit non-linear data

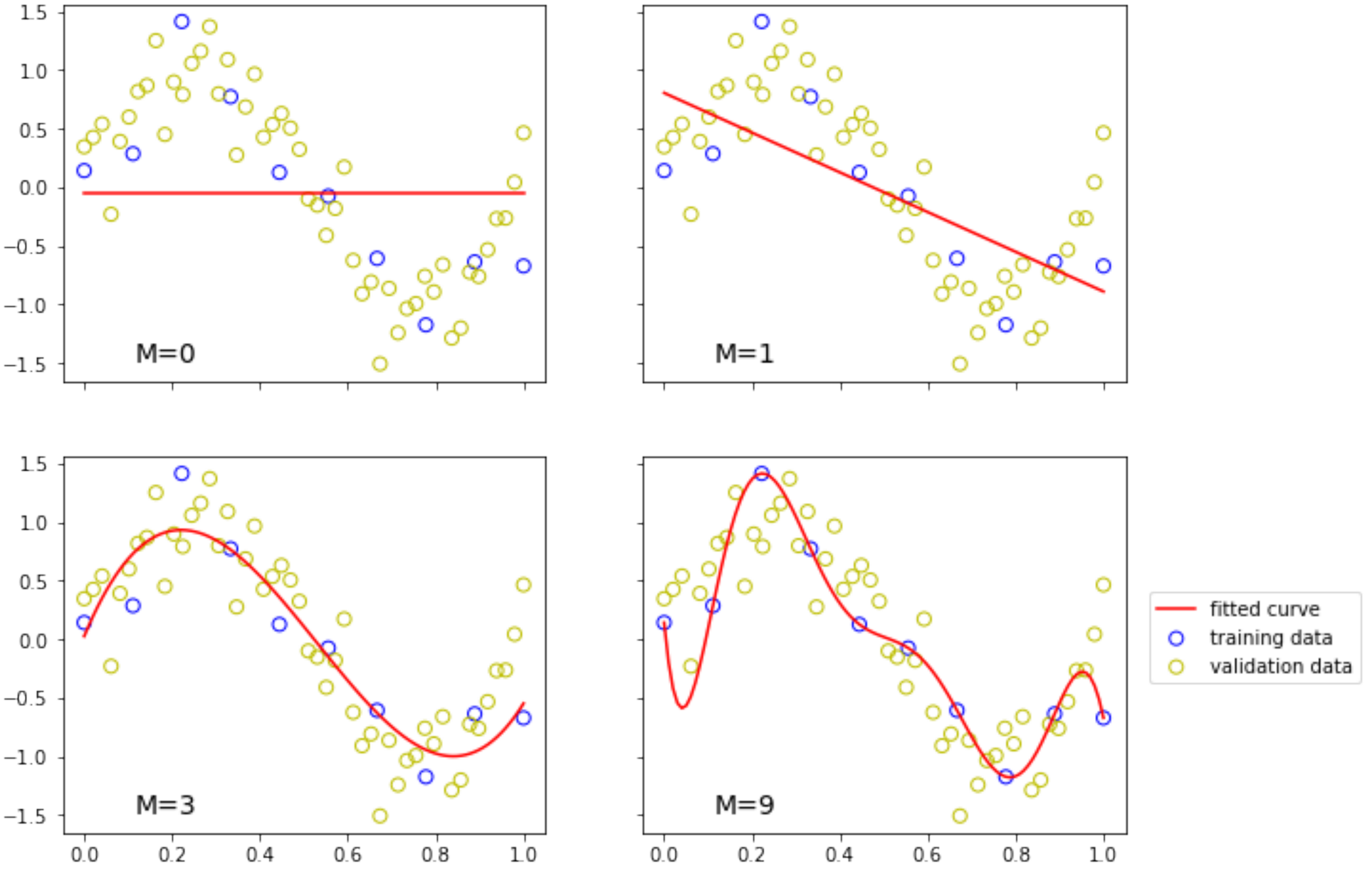
And now you have a model selection problem!



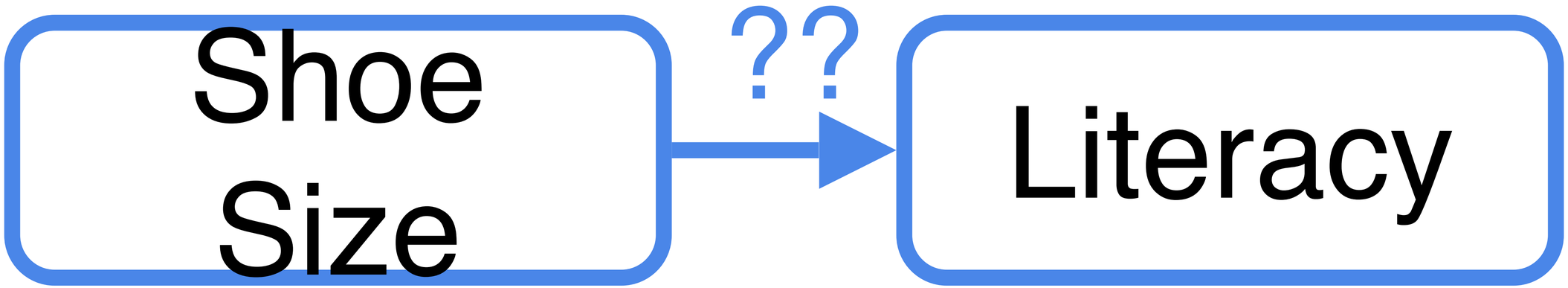
OLS can fit non-linear data

And now you have a model selection problem!

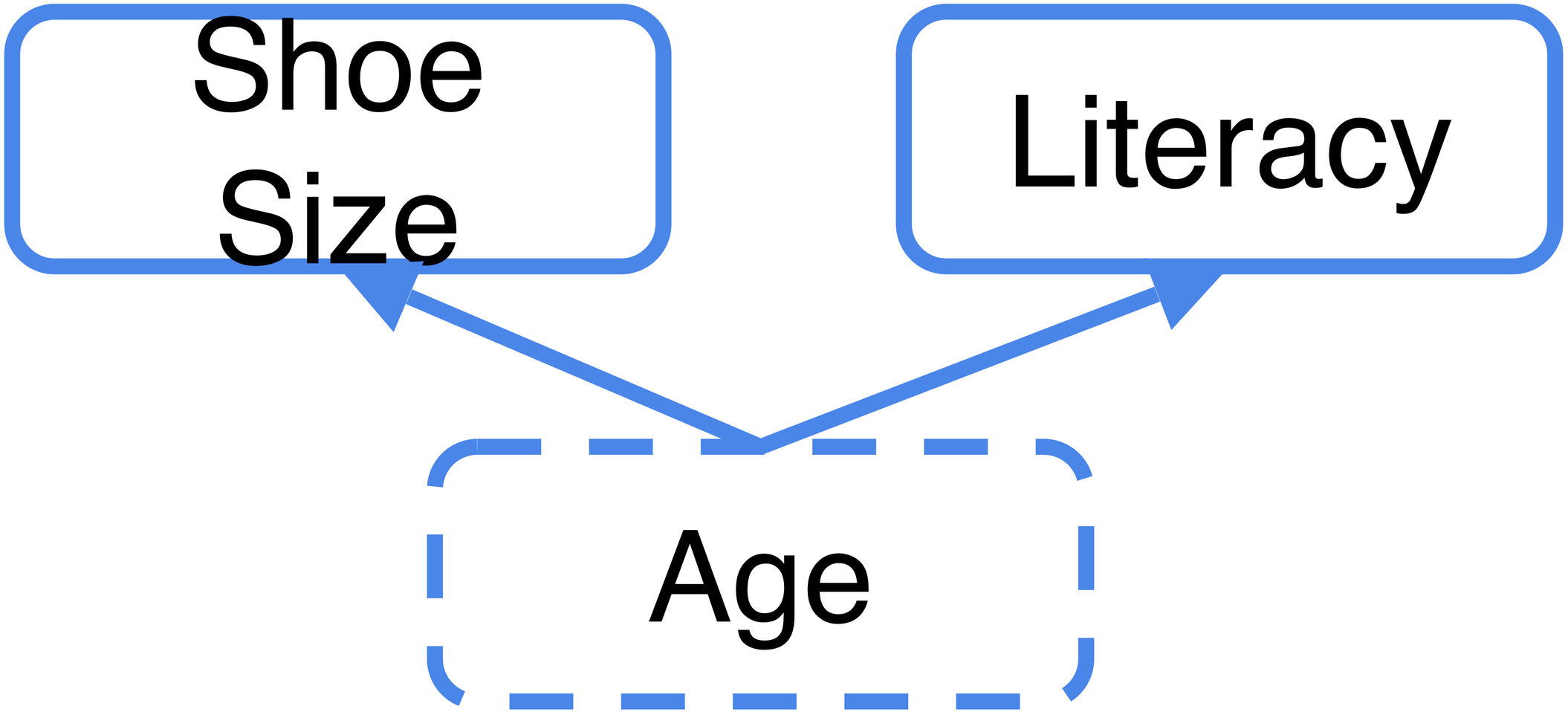


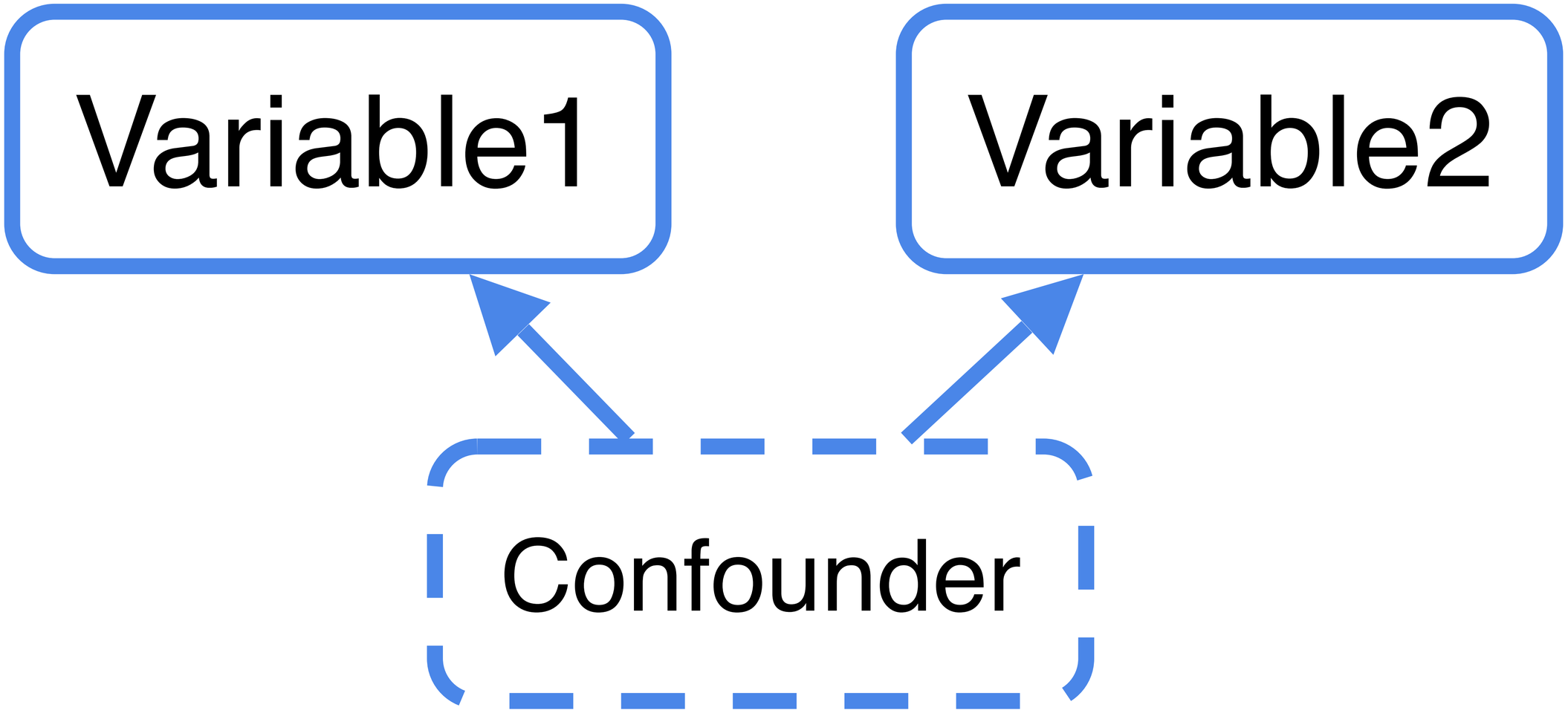

M=0
M=1
M=3
M=9
fitted curve
training data
validation data

Confounding


Shoe Size
??
Literacy


Small shoes
Not literate
Child
Big shoes
Literate
Adult


Shoe Size
Literacy
Age


Variable1
Variable2
Confounder

Confounding

popsicles → crime rate

Your analysis sees an increase in crime rate whenever popsicle sales increase. What could confound this analysis?

A
popsicle preference

B
new gun laws

C
temperature

D
changes in
popsicle prices

E
new law
enforcement
officers

You can plan ahead to avoid confounding and/or include confounders in your models to account for their role on the outcome variable.

Ignoring confounders will lead you to draw incorrect conclusions

Stratification changes results

Sample: 400 patients with index vertebral fractures

…looks like vertebroplasty was *way* worse for patients!

Vertebroplasty	Conservative care	Relative risk (95% confidence interval)
30/200 (15%)	15/200 (7.5%)	2.0 (1.1–3.6)

subsequent fractures

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

But wait...at time of initial fracture...

	Vertebroplasty N = 200	**Conservative care N = 200**
Age, y, mean ± SD	78.2 ± 4.1	79.0 ± 5.2
Weight, kg, mean ± SD	54.4 ± 2.3	53.9 ± 2.1
Smoking status, No. (%)	110 (55)	16 (8)

Age and weight are similar between groups. **Smoking Status** differs vastly.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/

So…let's stratify those results real quick

Smoke			**No smoke**		
Vertebroplasty	Conservative	RR (95% confidence interval)	Vertebroplasty	Conservative	RR (95% confidence interval)
23/110 (21%)	3/16 (19%)	1.1 (0.4, 3.3)	7/90 (8%)	12/184(7%)	1.2 (0.5, 2.9)

Risk of re-fracture is now similar within group

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503514/