Reminders

Upcoming due dates
Mon Feb 02 Quiz 4
Wed Feb 04 Project Proposal

Repo invites: Click accept in email!

Statistical Inference 2
Linear models

Data Science in Practice

Jason G. Fleischer, PhD Dept. of Cognitive Science UC San Diego

CORRELATION

ASSOCIATION BETWEEN VARIABLES

i.e. Pearson Correlation, Spearman Correlation, chi-square test

COMPARISON OF MEANS

DIFFERENCE IN MEANS BETWEEN VARIABLES

i.e. t-test, ANOVA

REGRESSION

DOES CHANGE IN ONE VARIABLE MEAN CHANGE IN ANOTHER?

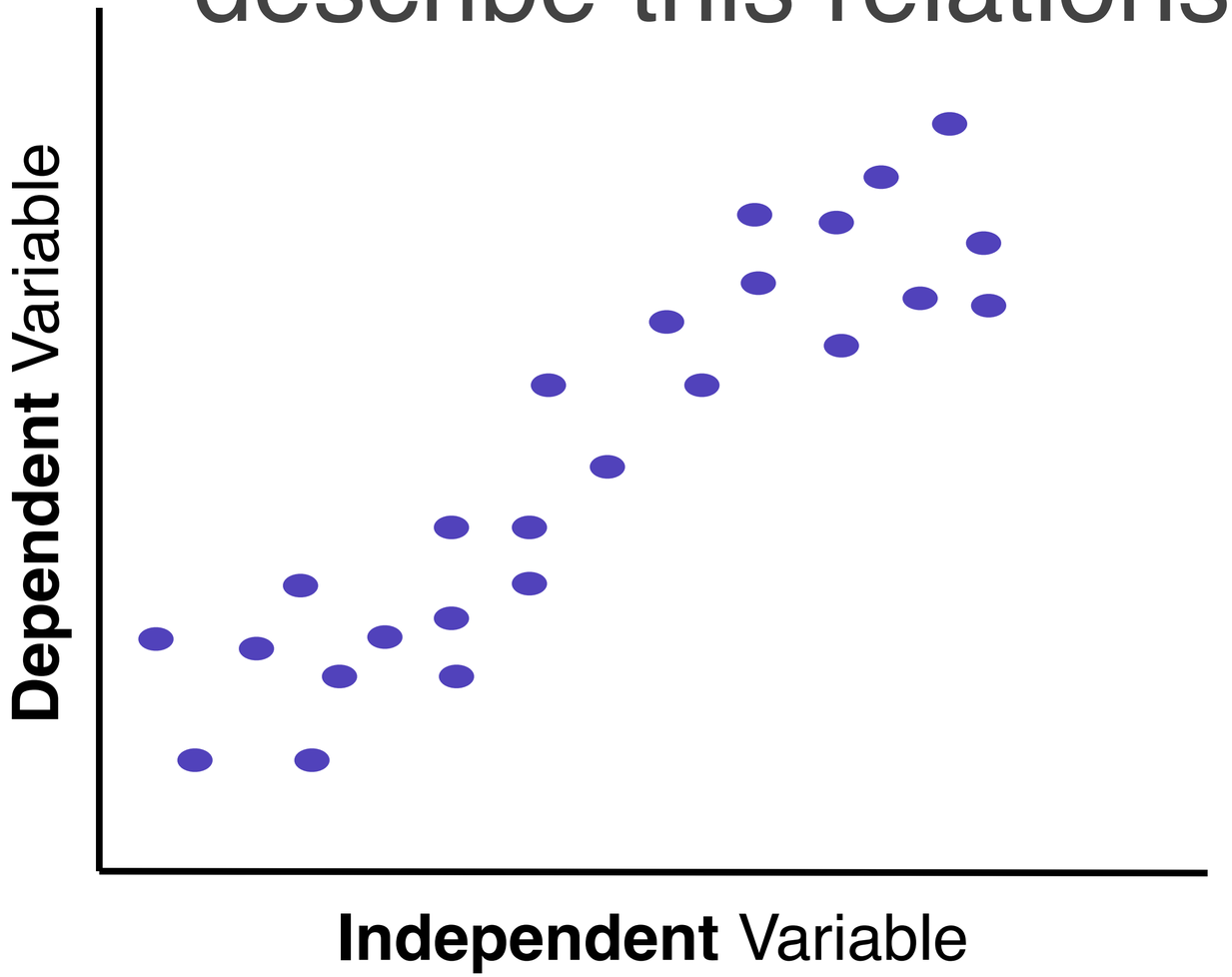
I.e. simple regression, multiple regression

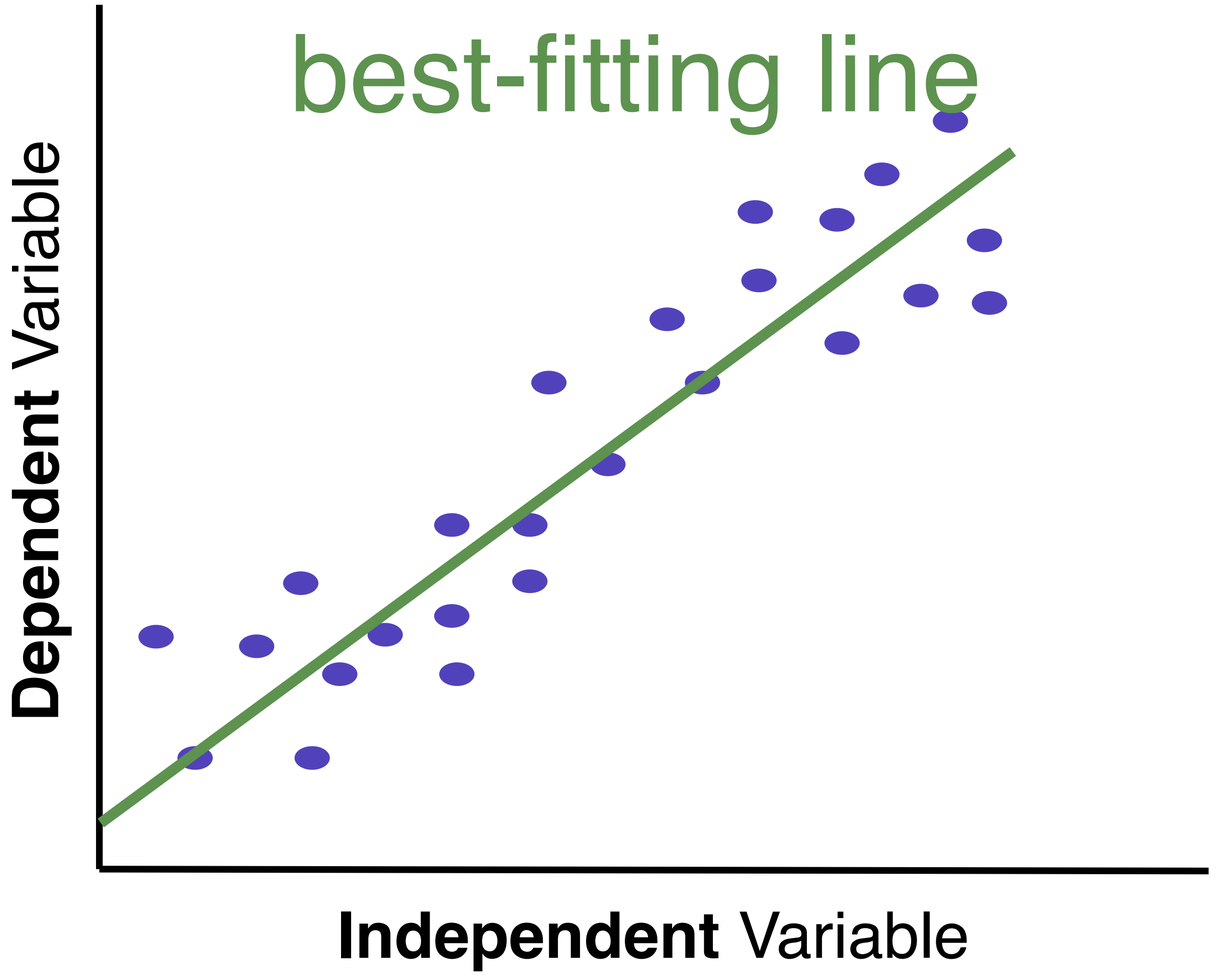
NON-PARAMETRIC TESTS

FOR WHEN ASSUMPTIONS IN THESE OTHER 3 CATEGORIES ARE NOT MET

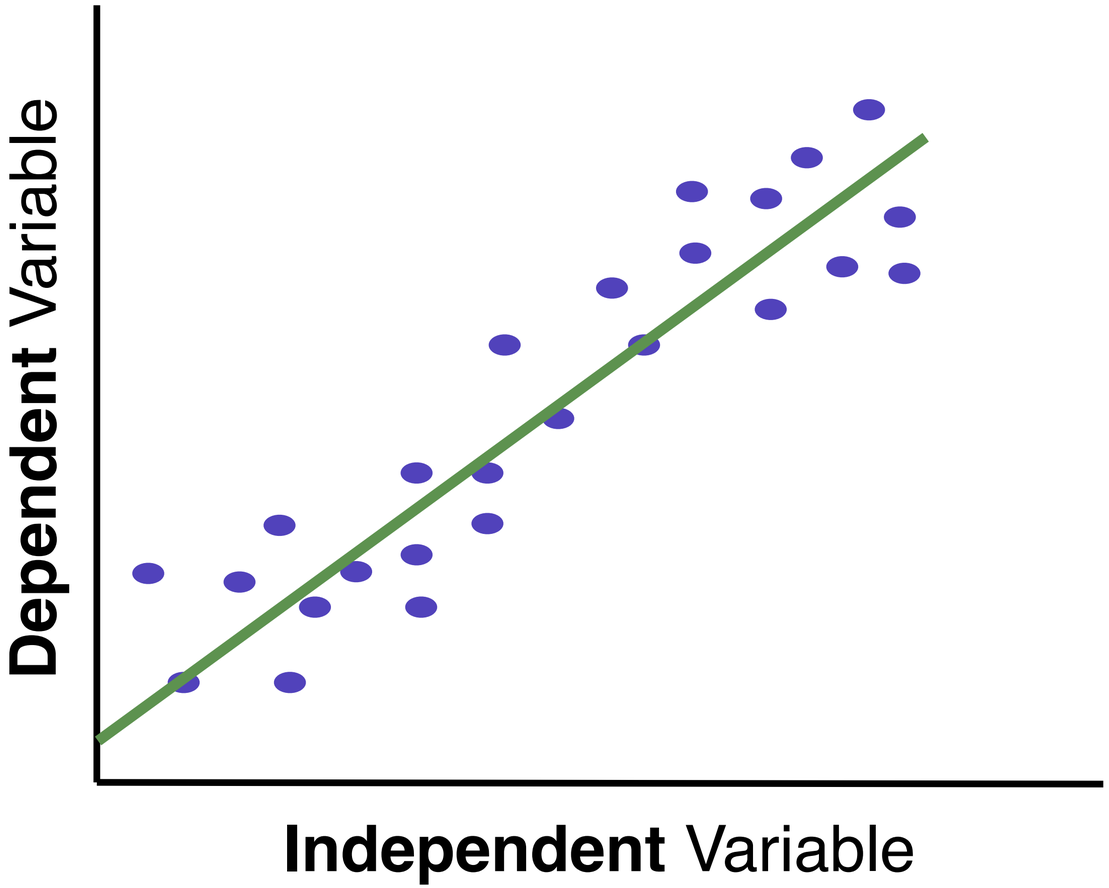
i.e. Wilcoxon rank-sum test, Wilcoxon sign-rank test, sign test

Linear regression can be used to describe this relationship

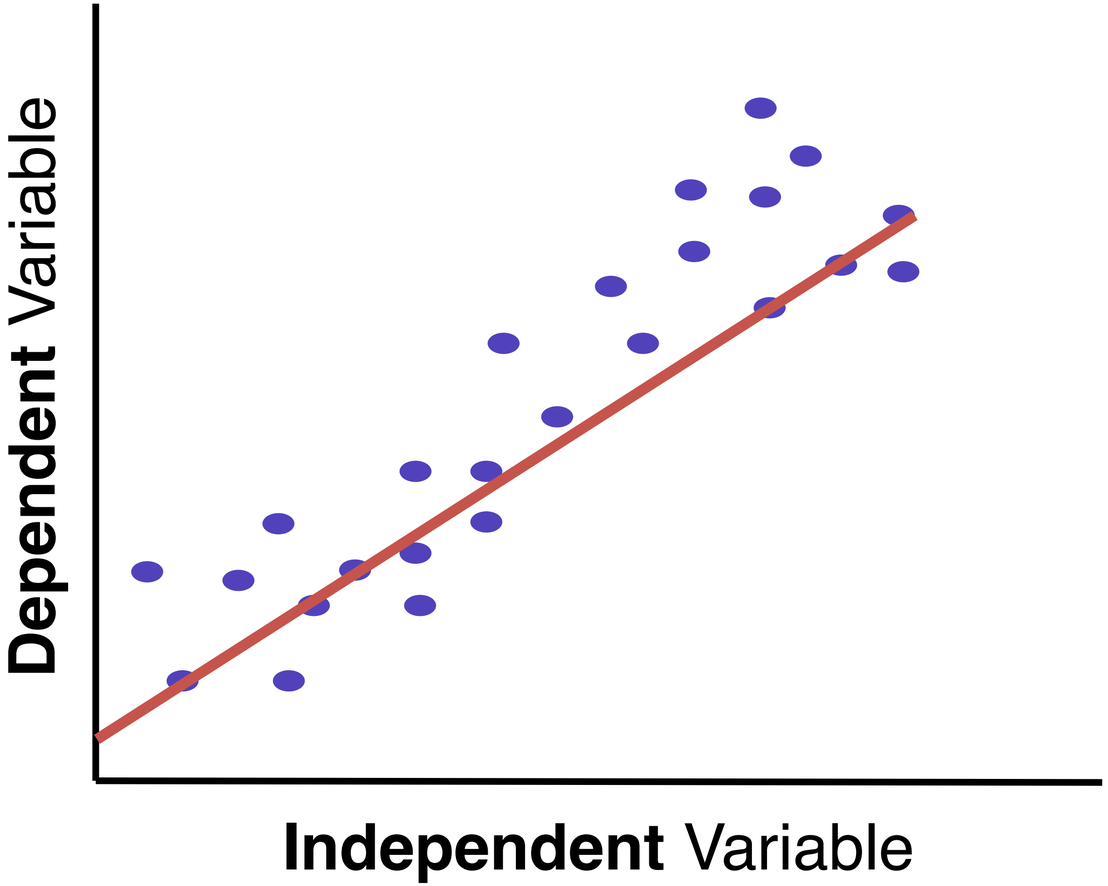



best-fitting line
Dependent Variable
Independent Variable

Best-fitting line



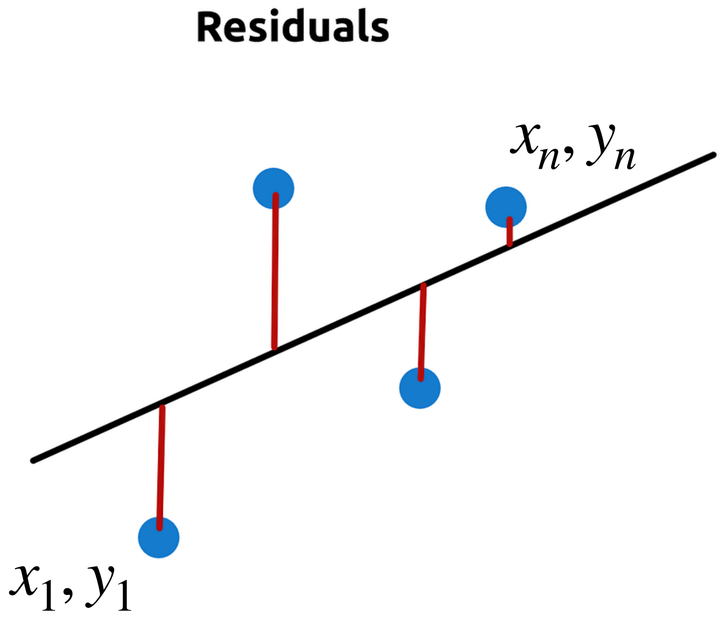
NOT a best-fitting line



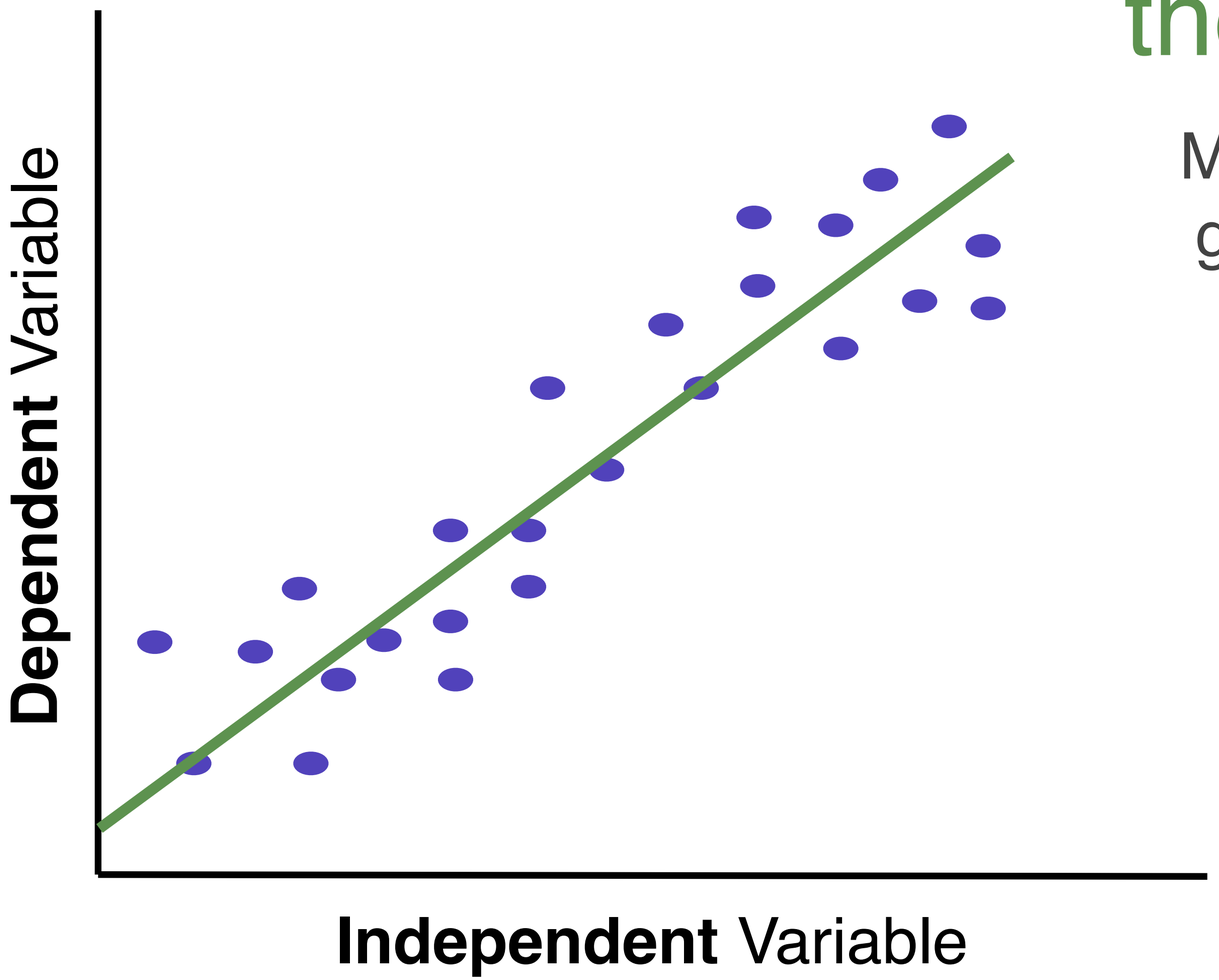
Ordinary Least Squares

Just one of several regression algorithms

- Minimizes Residual Sum of Squares (RSS)
- OLS residuals are VERTICAL only!!
  - Implies x predicts y values!
  - Compare with TLS regression which uses perpendicular distance
- OLS residuals are SQUARED!!
  - Outliers have large influence!
  - Compare with LAD (aka L1 loss) regression which uses absolute value instead of square

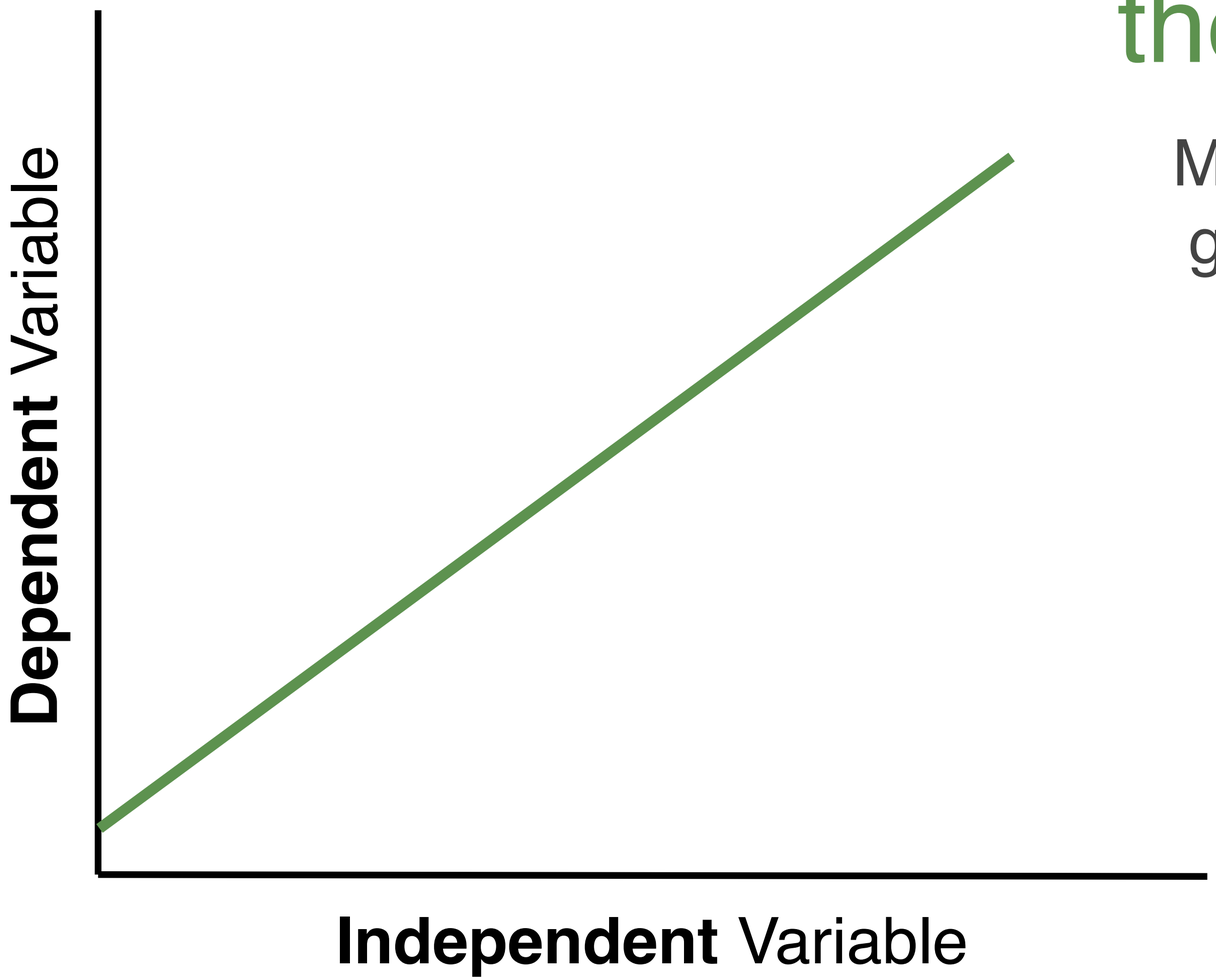


$$RSS = \sum_{k=1}^{n} (f(x_k) - y_k)^2$$



This line is a model of the data

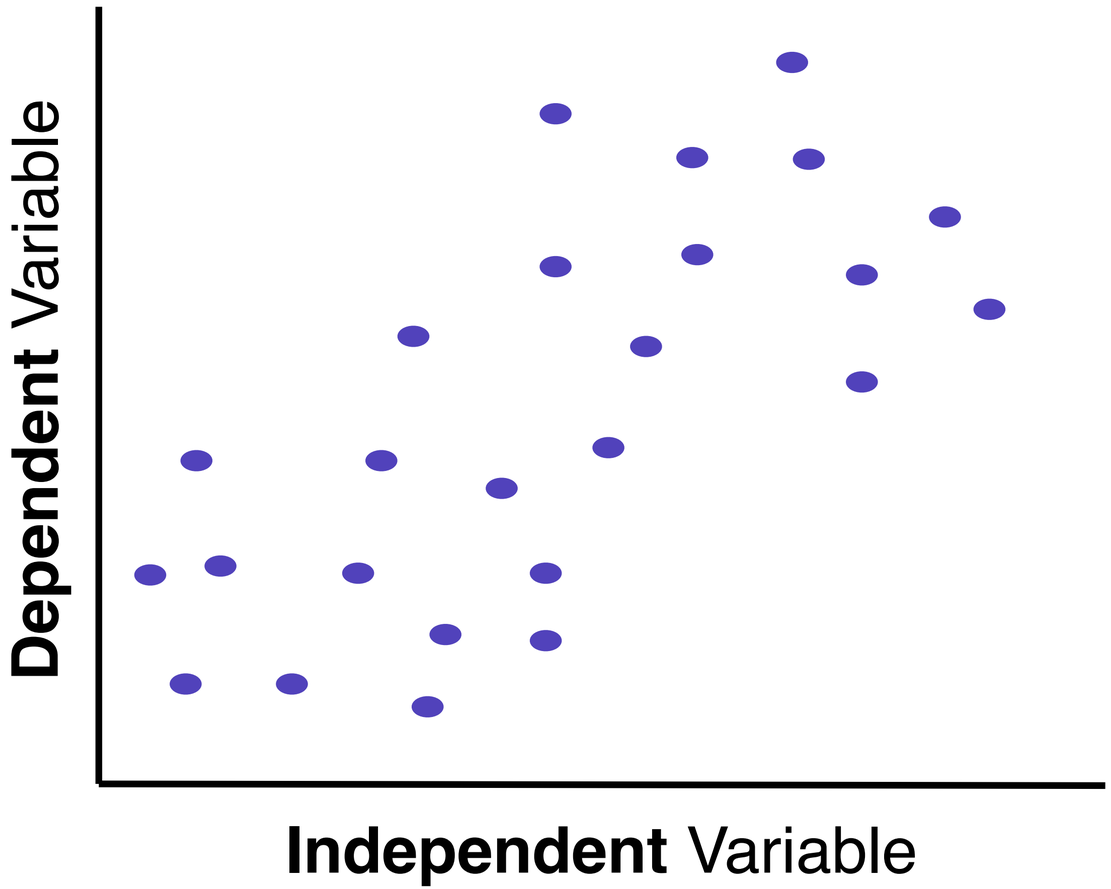
Models are mathematical equations generated to *represent* the real life situation



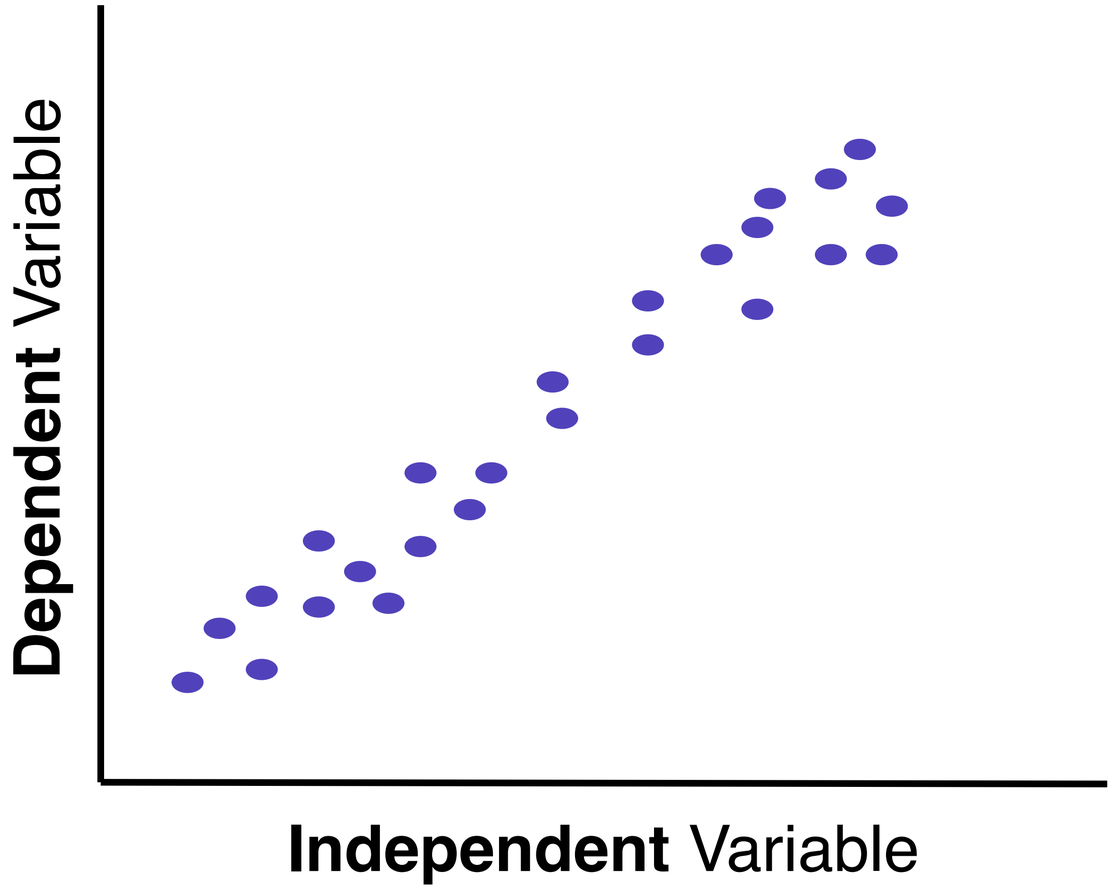
This line is a model of the data

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

weaker relationship



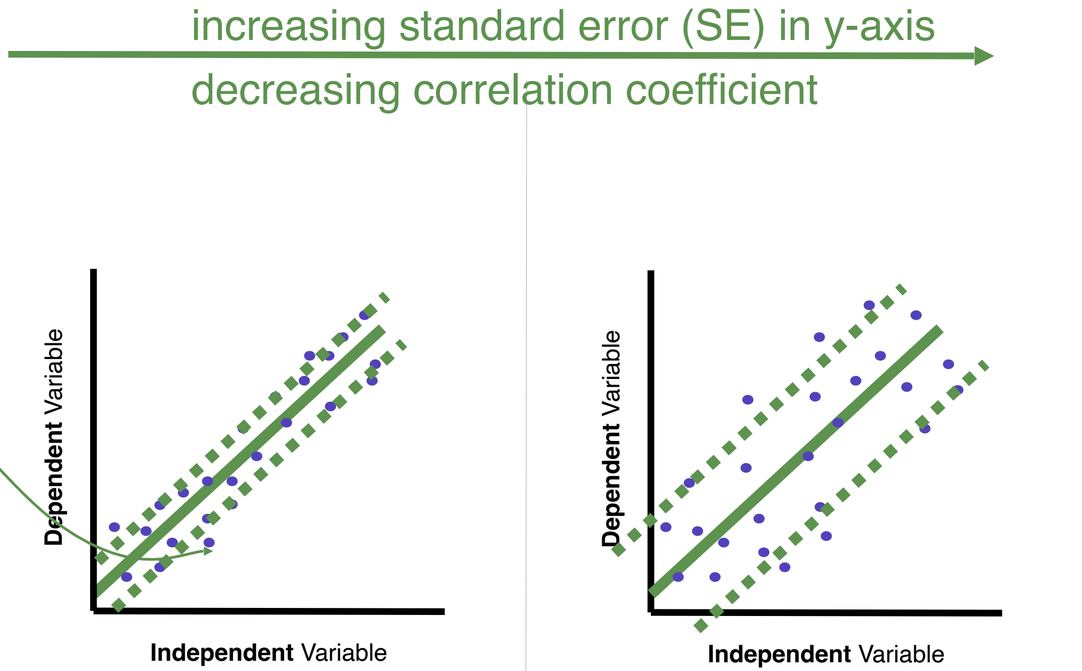
stronger relationship



stronger relationship = higher correlation

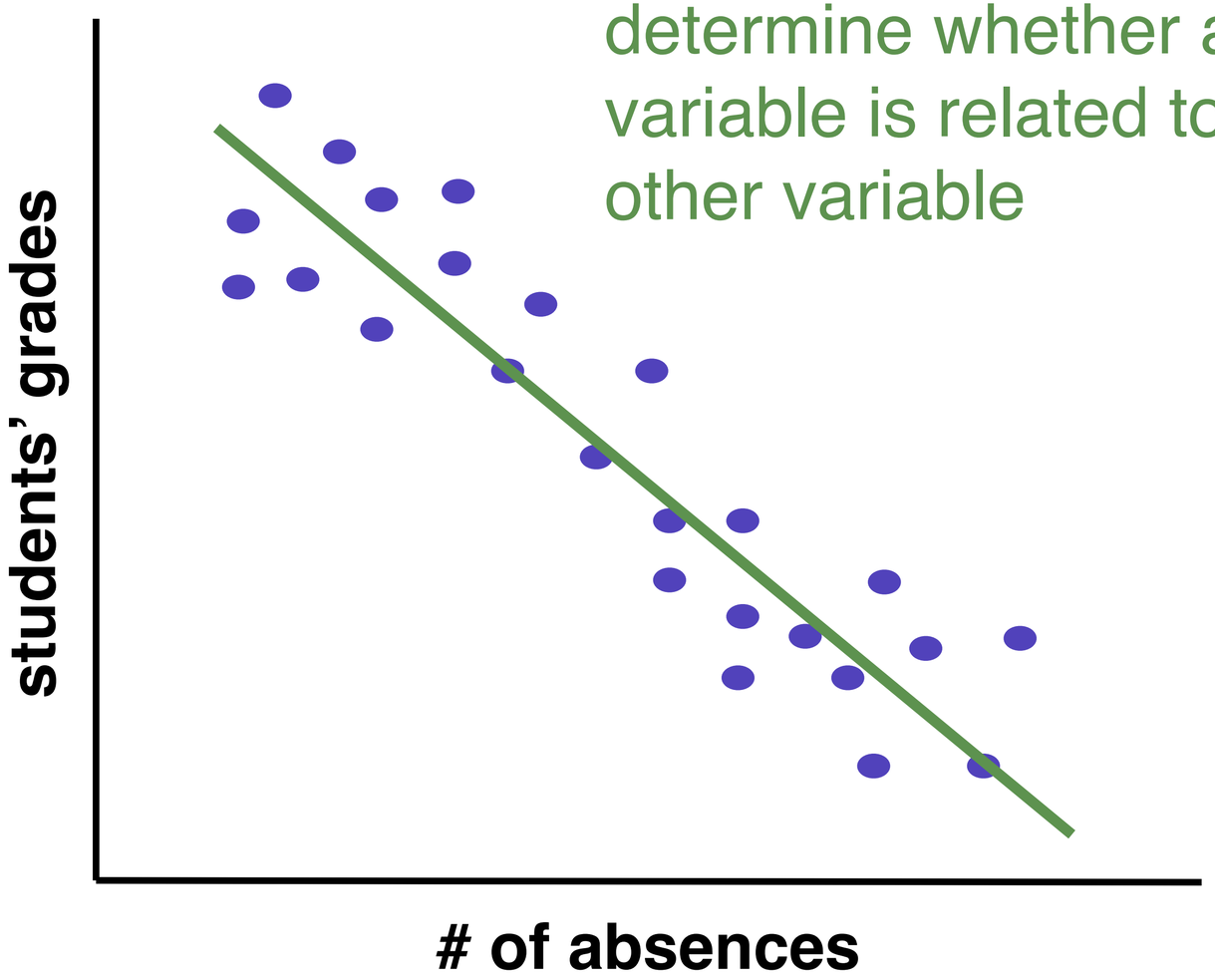
This is a kind of effect size

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

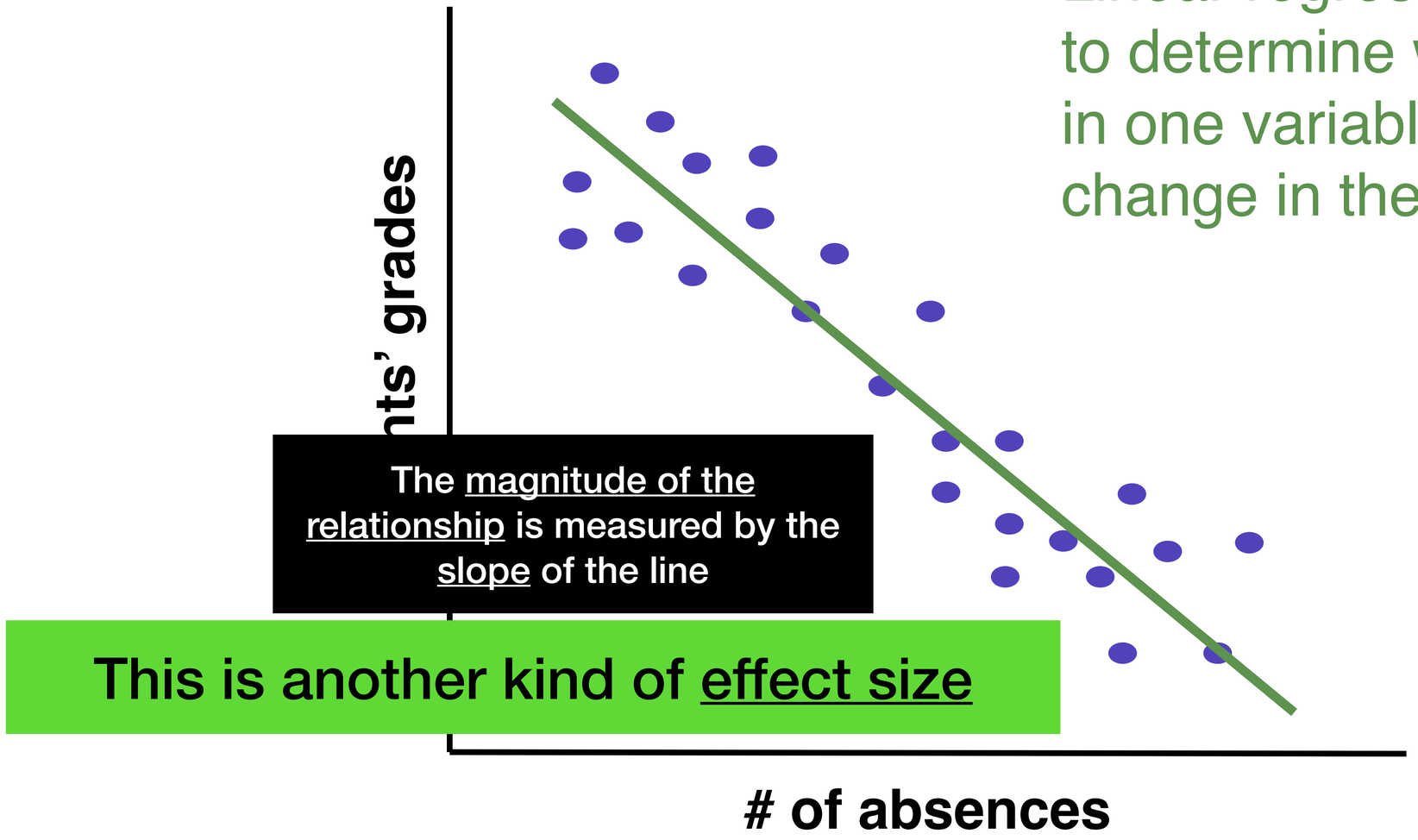


Standard error is standard deviation / sqrt(n)

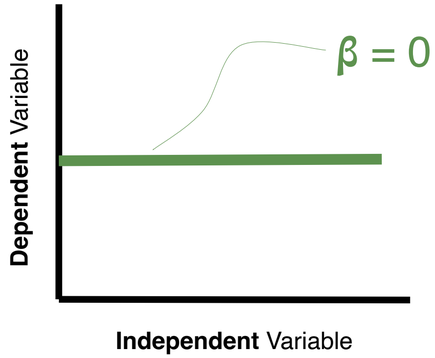
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable



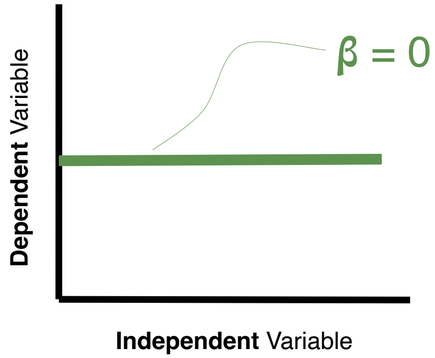
Linear regression can be used to determine whether a change in one variable is related to the change in the other variable

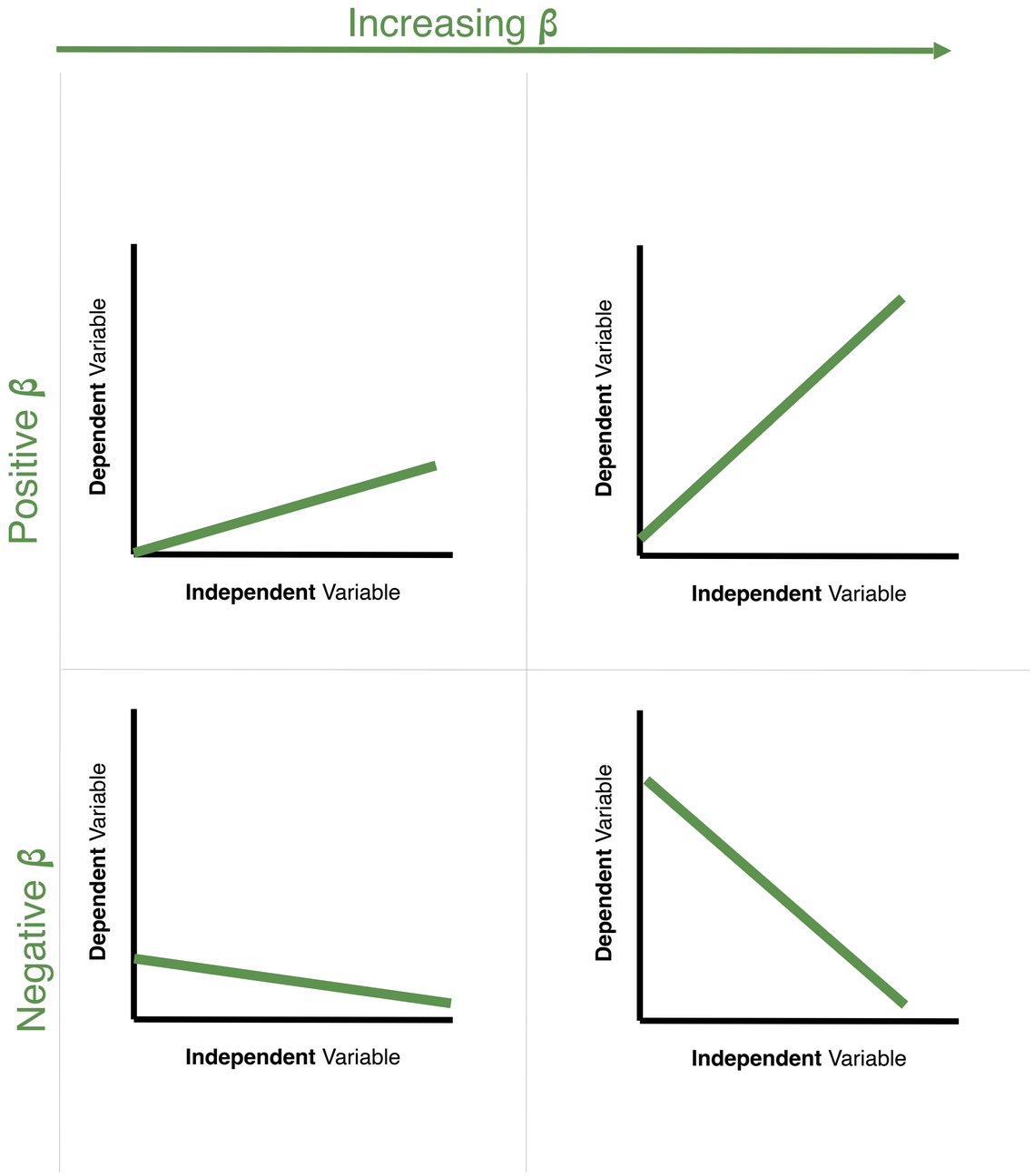


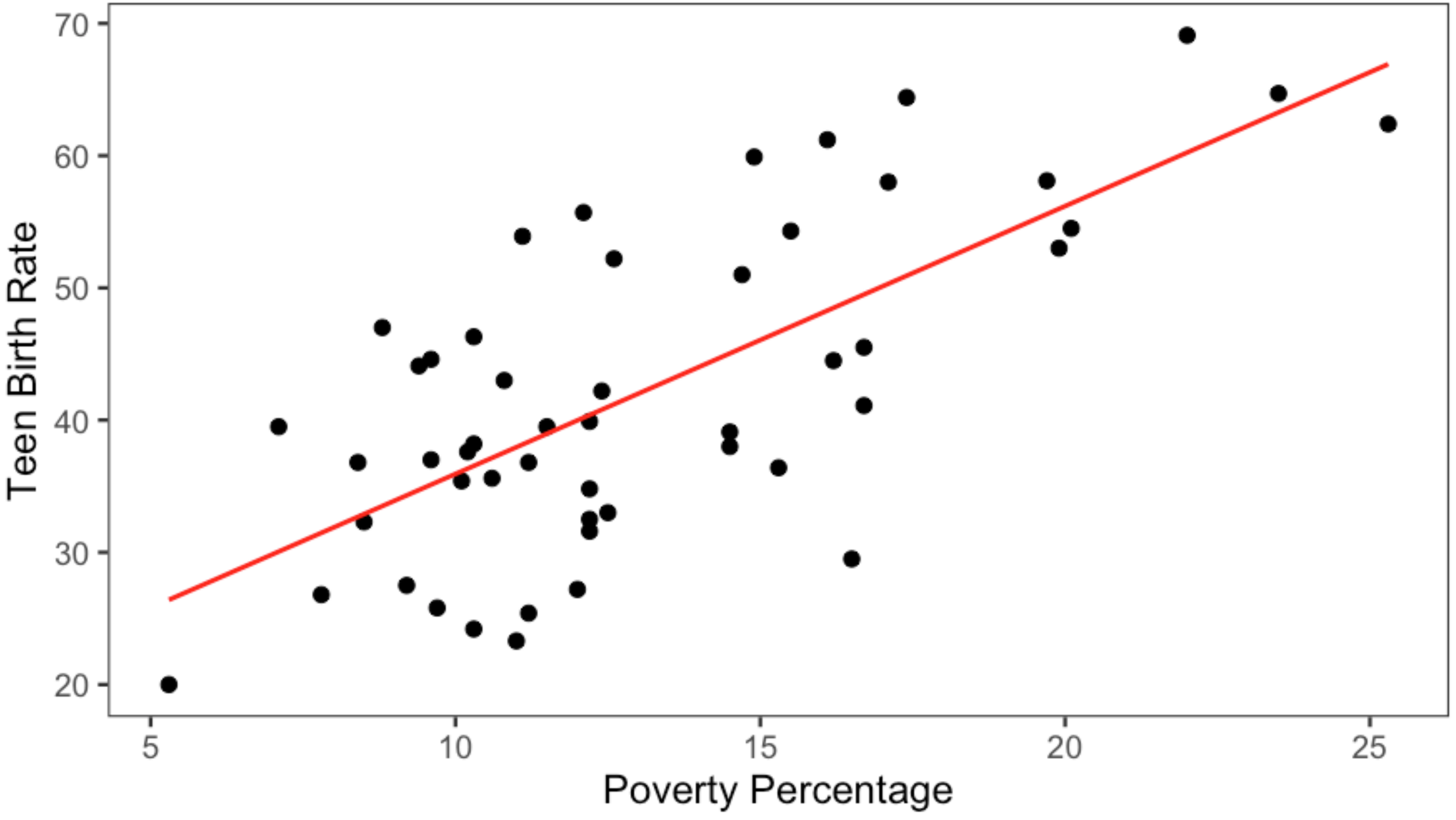
Effect size (β) can be estimated using the slope of the line

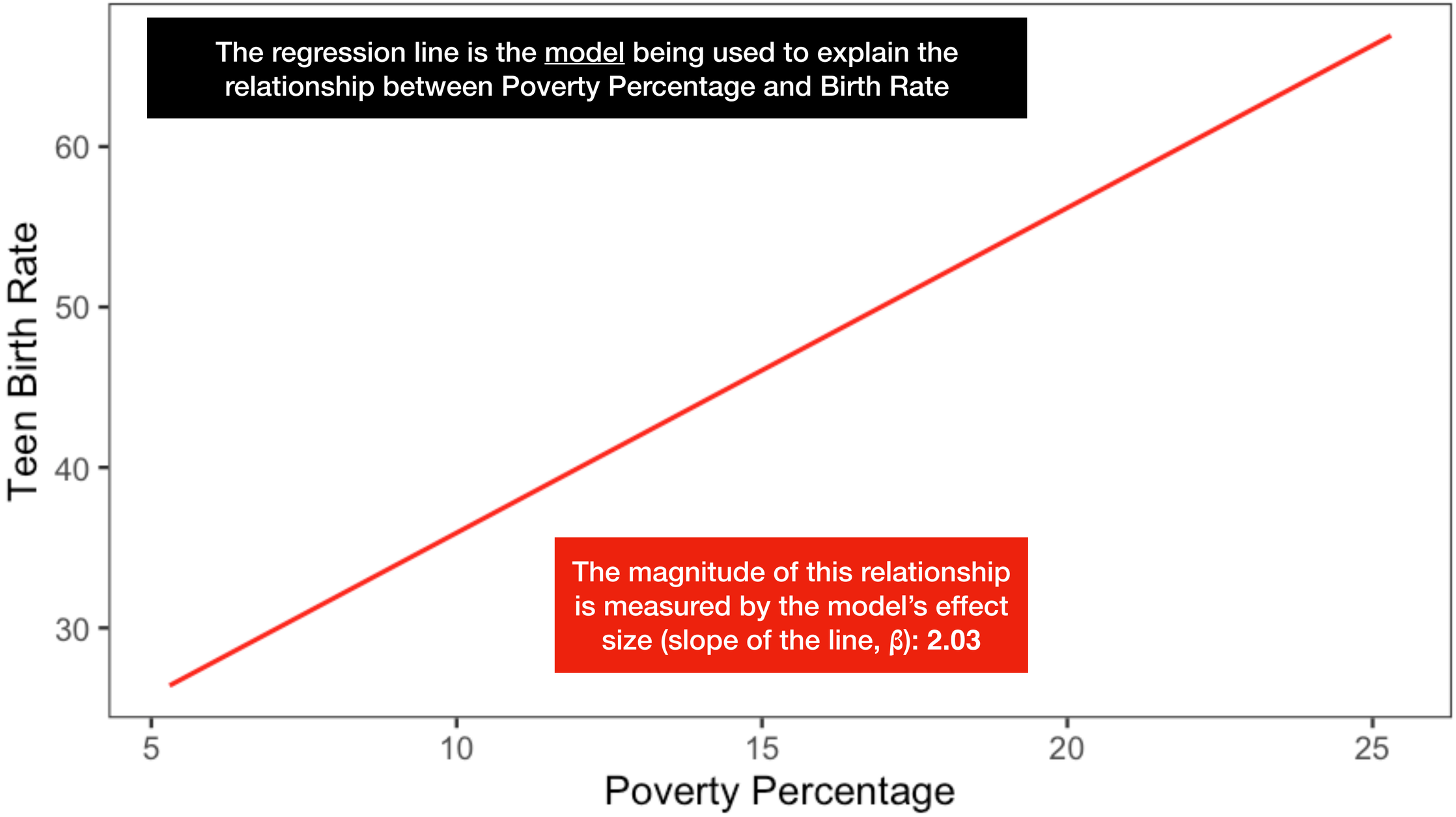


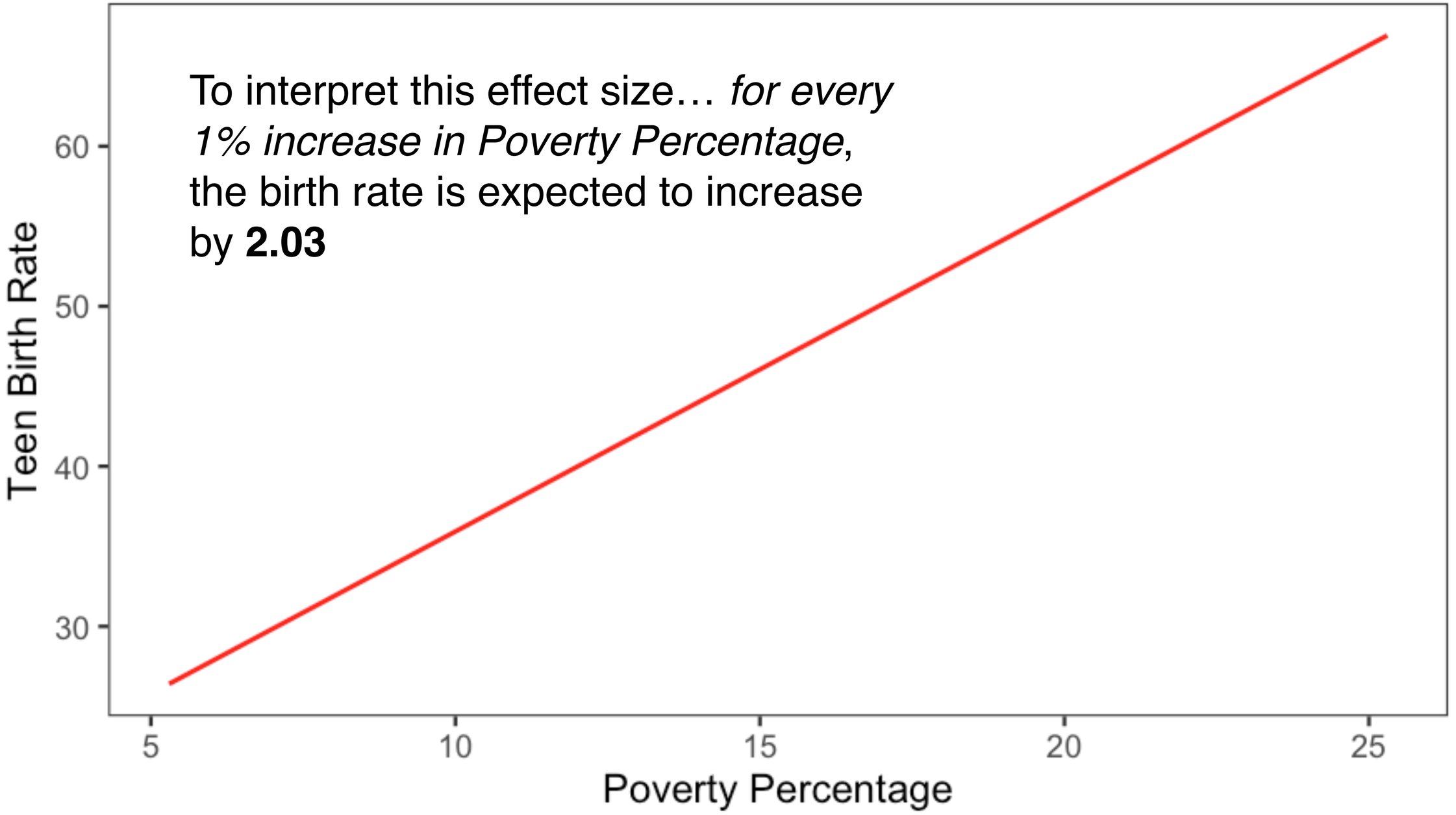
Effect size (β) can be estimated using the slope of the line





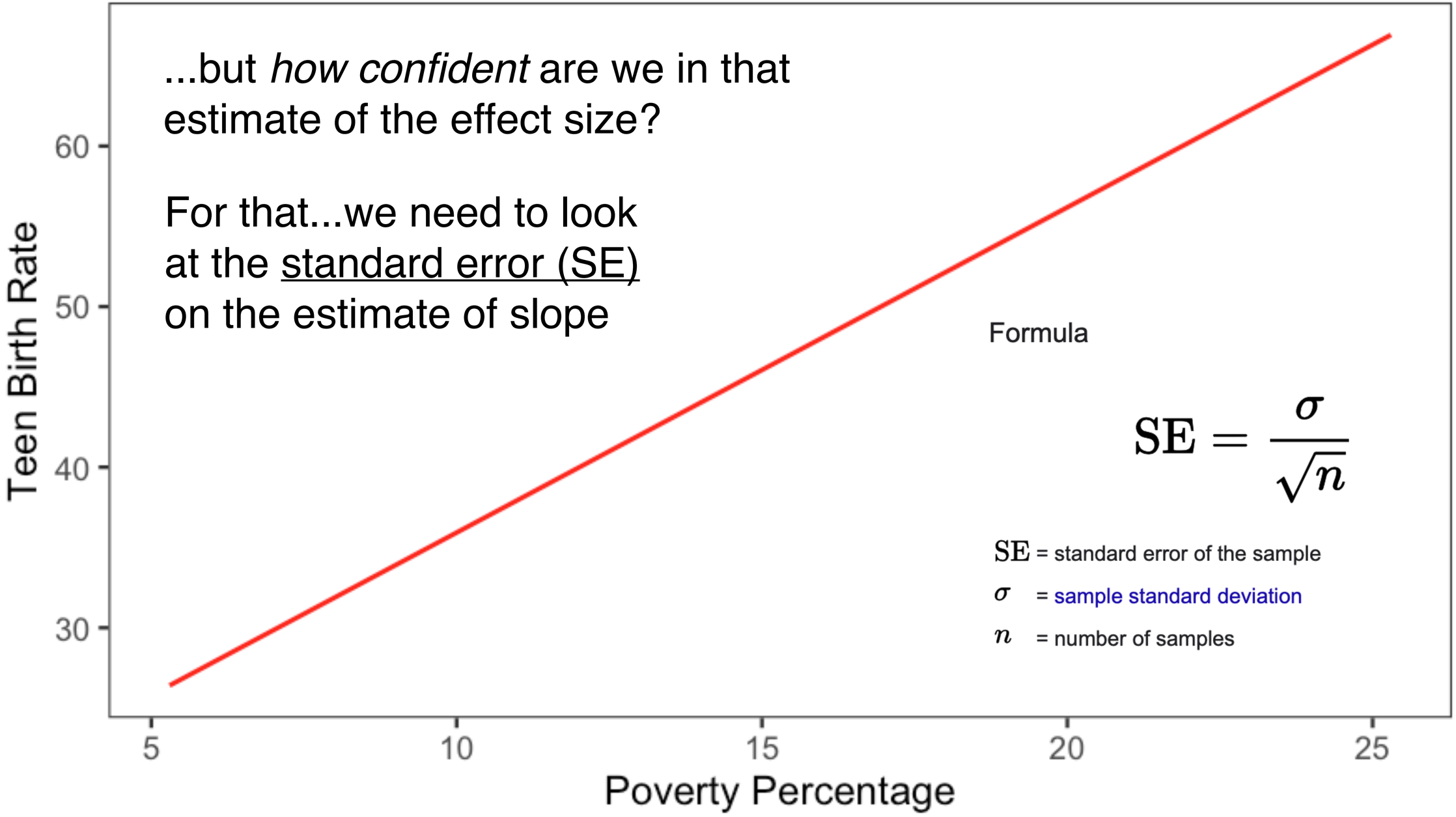

Teen Birth Rate
70
60
50
40
30
20
5
10
15
20
25
Poverty Percentage

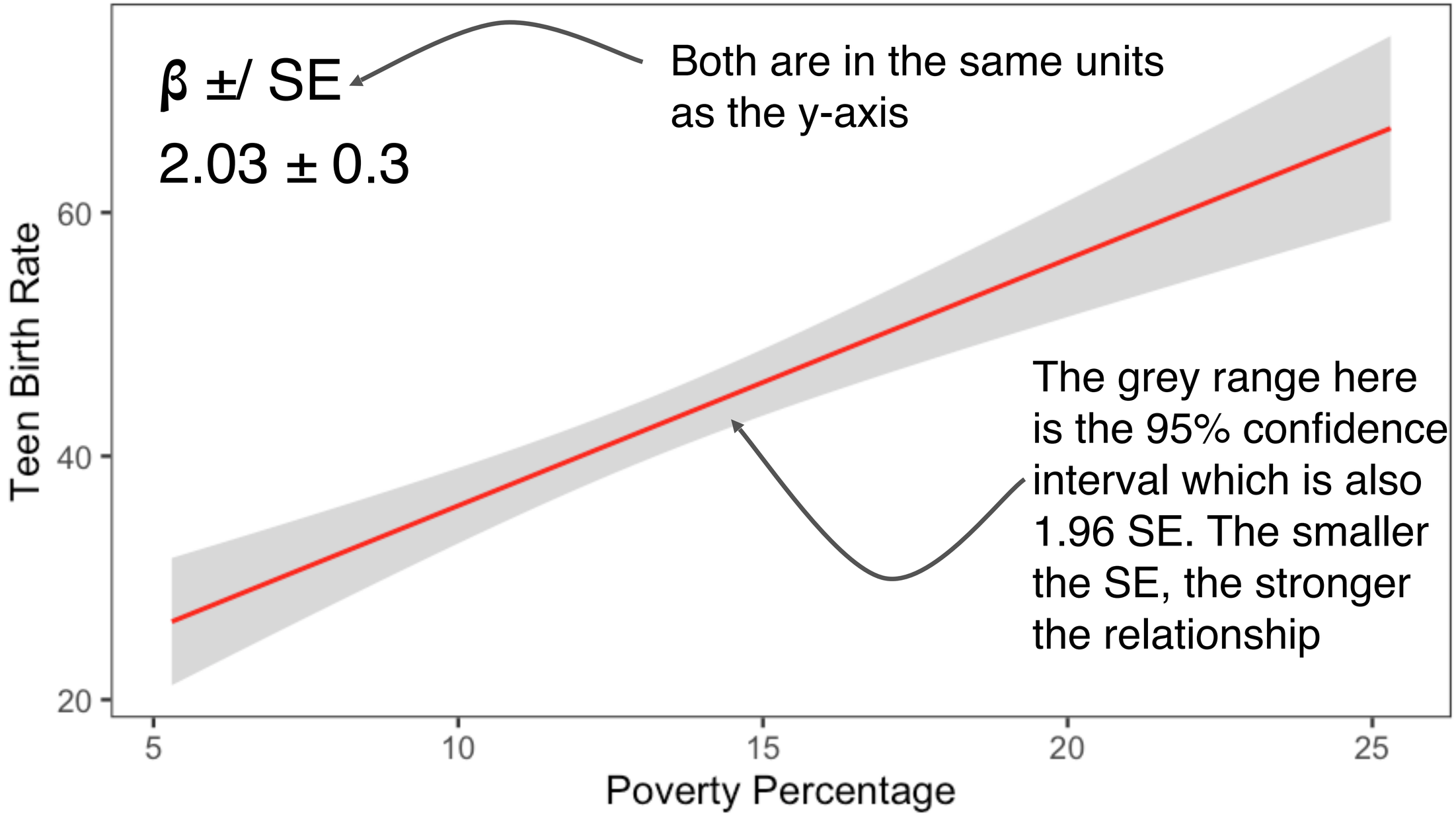

The regression line is the model being used to explain the relationship between Poverty Percentage and Birth Rate
The magnitude of this relationship is measured by the model's effect size (slope of the line, β): 2.03
Teen Birth Rate
60
50
40
30
5
10
15
20
25
Poverty Percentage


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

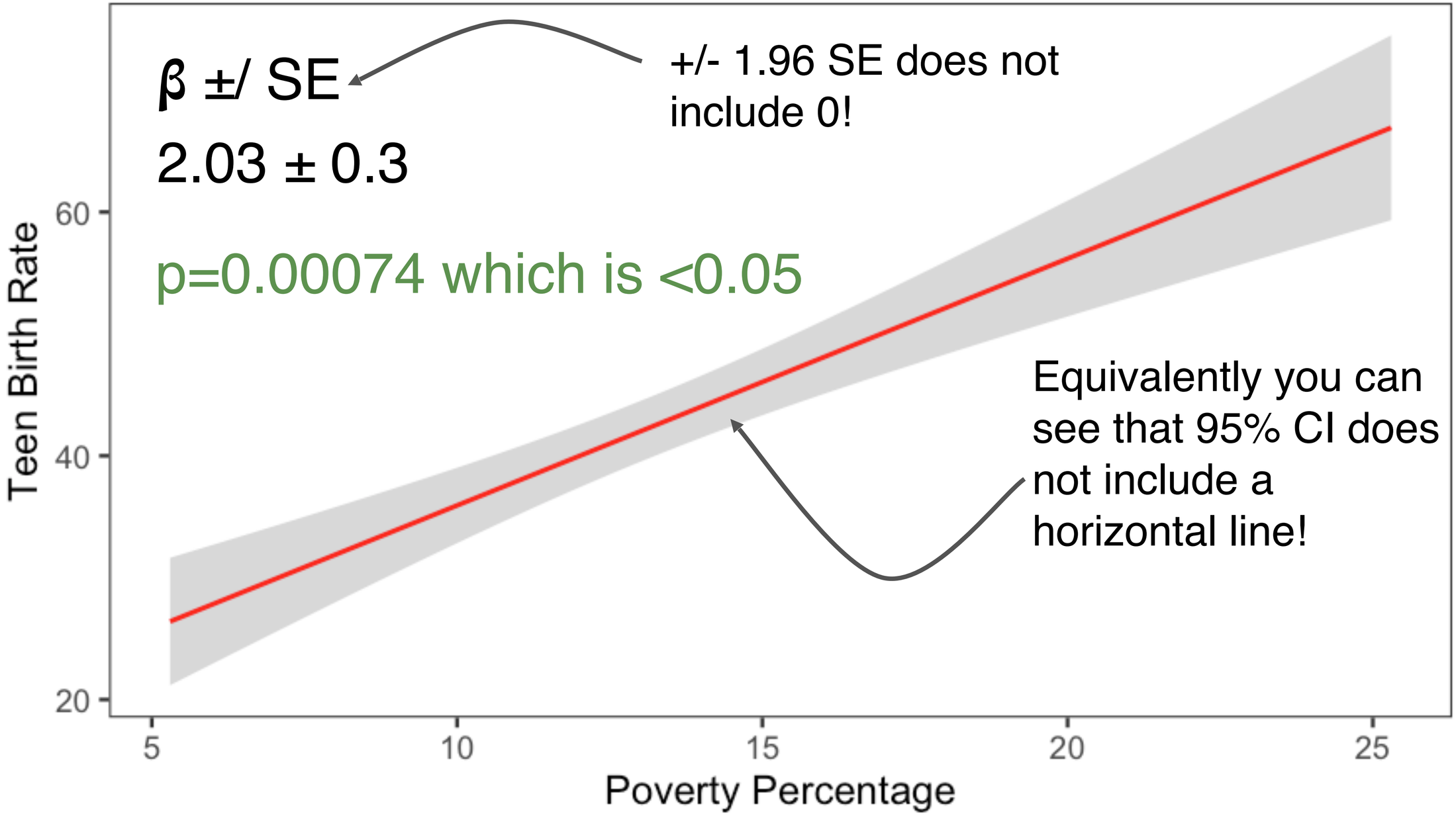



β ±/ SE
2.03 ± 0.3
Both are in the same units as the y-axis
The grey range here is the 95% confidence interval which is also 1.96 SE. The smaller the SE, the stronger the relationship
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage

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

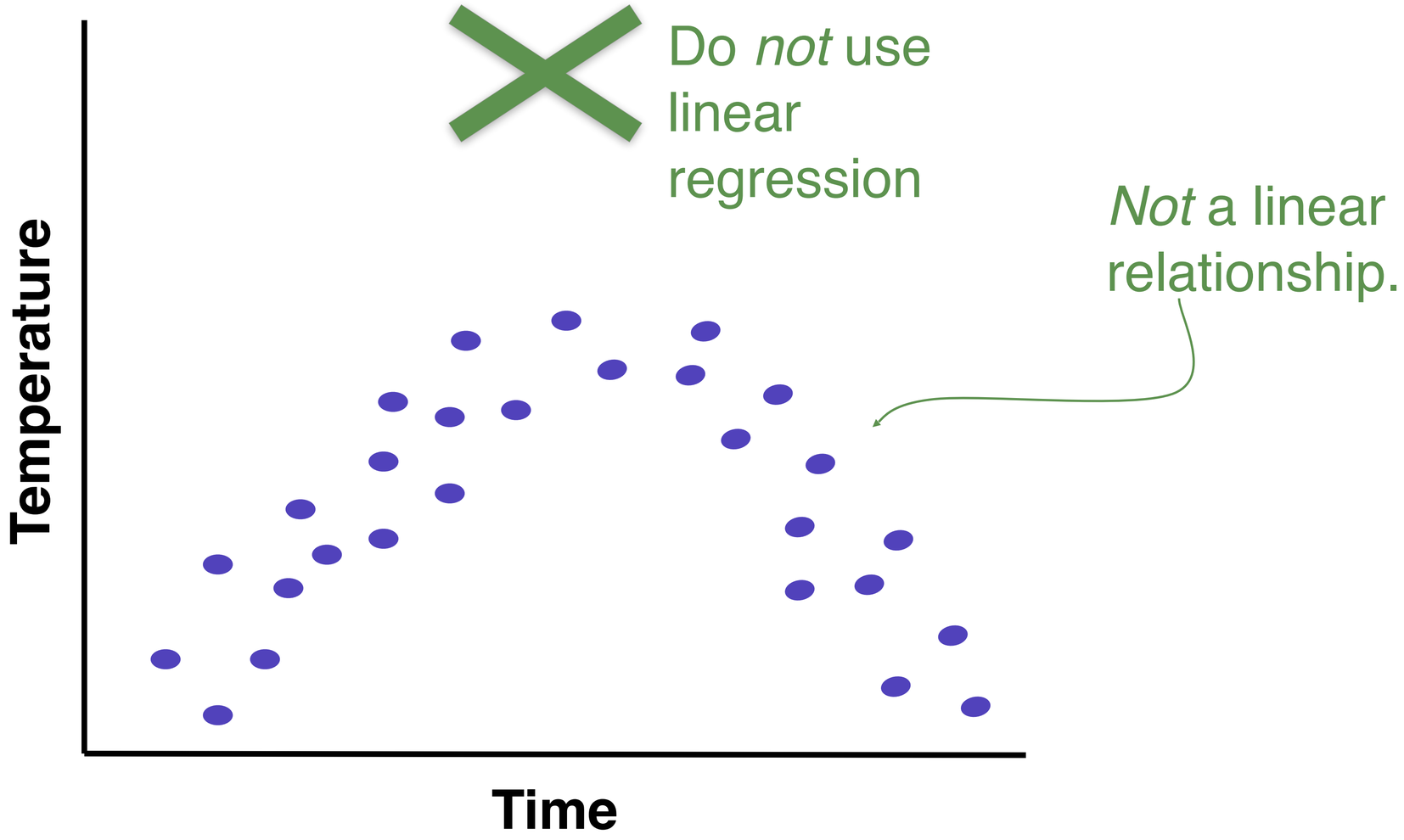
Takes into account the effect size (β) and the SE

Confidence interval : a range of values calculated from a sample statistic, such that there is a specified probability that the value of the true value of the population (parameter) lies within it.

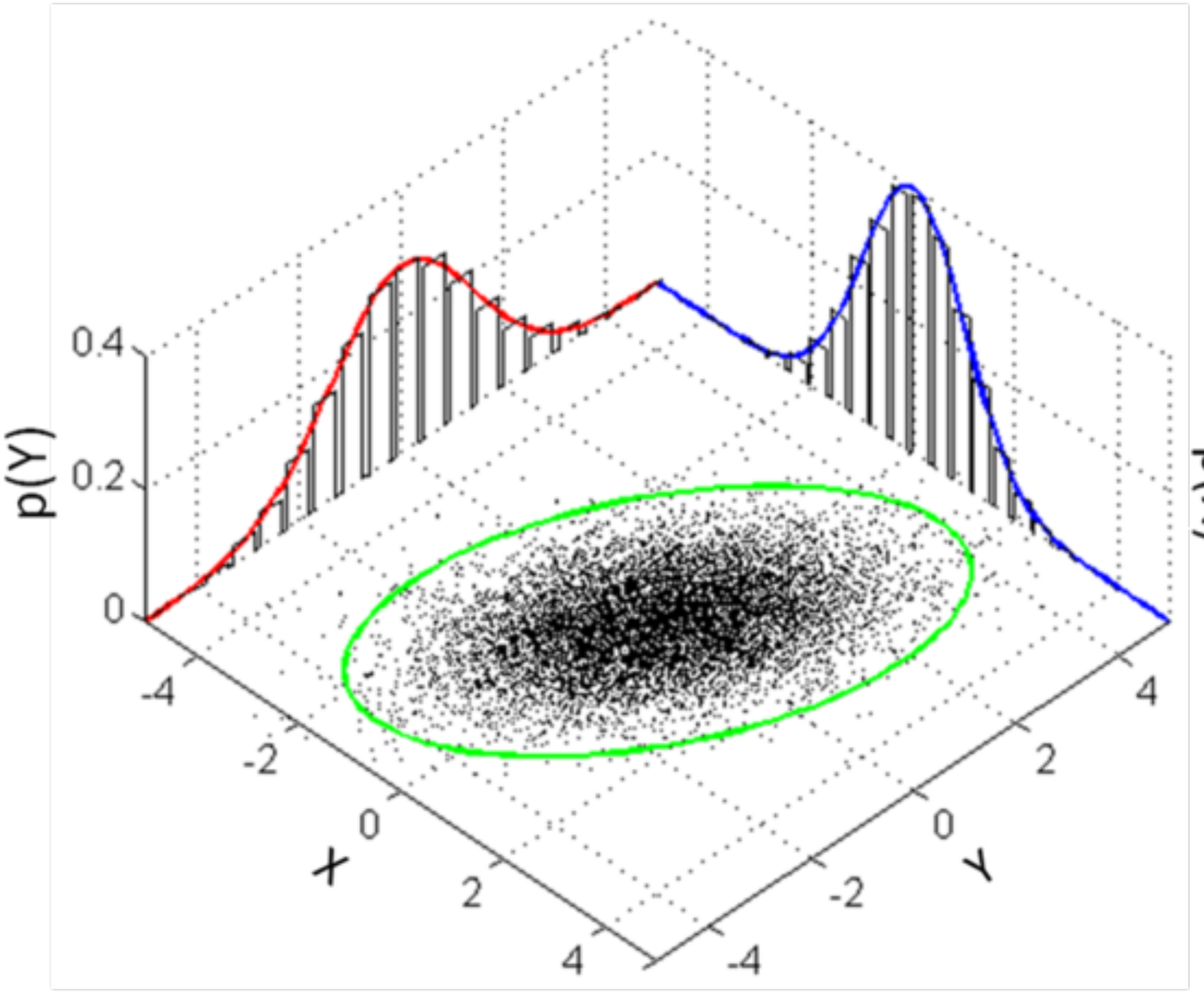

β ±/ SE
2.03 ± 0.3
+/- 1.96 SE does not include 0!
p=0.00074 which is <0.05
Equivalently you can see that 95% CI does not include a horizontal line!
Teen Birth Rate
60
40
20
5
10
15
20
25
Poverty Percentage

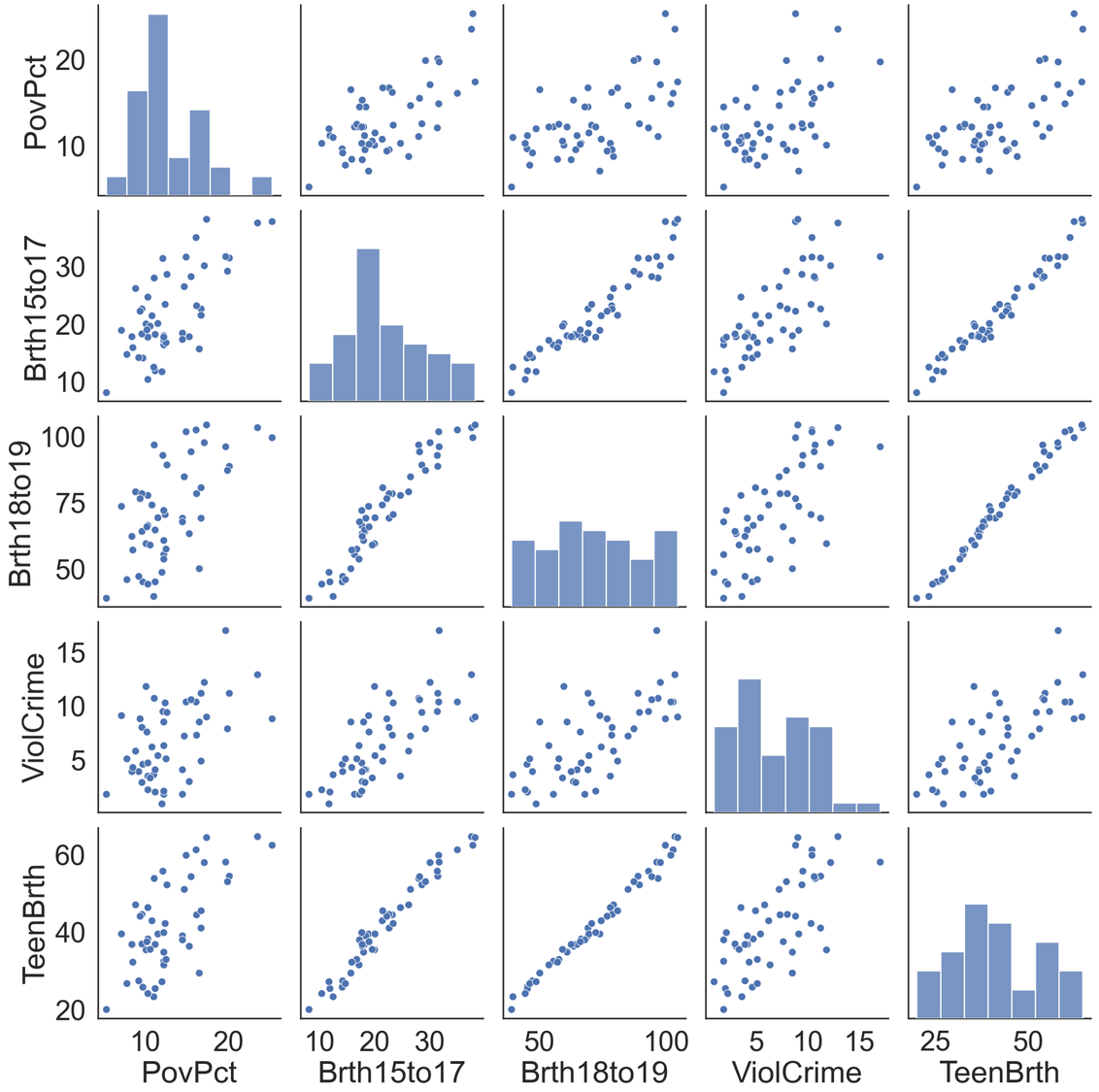
Assumptions of linear regression

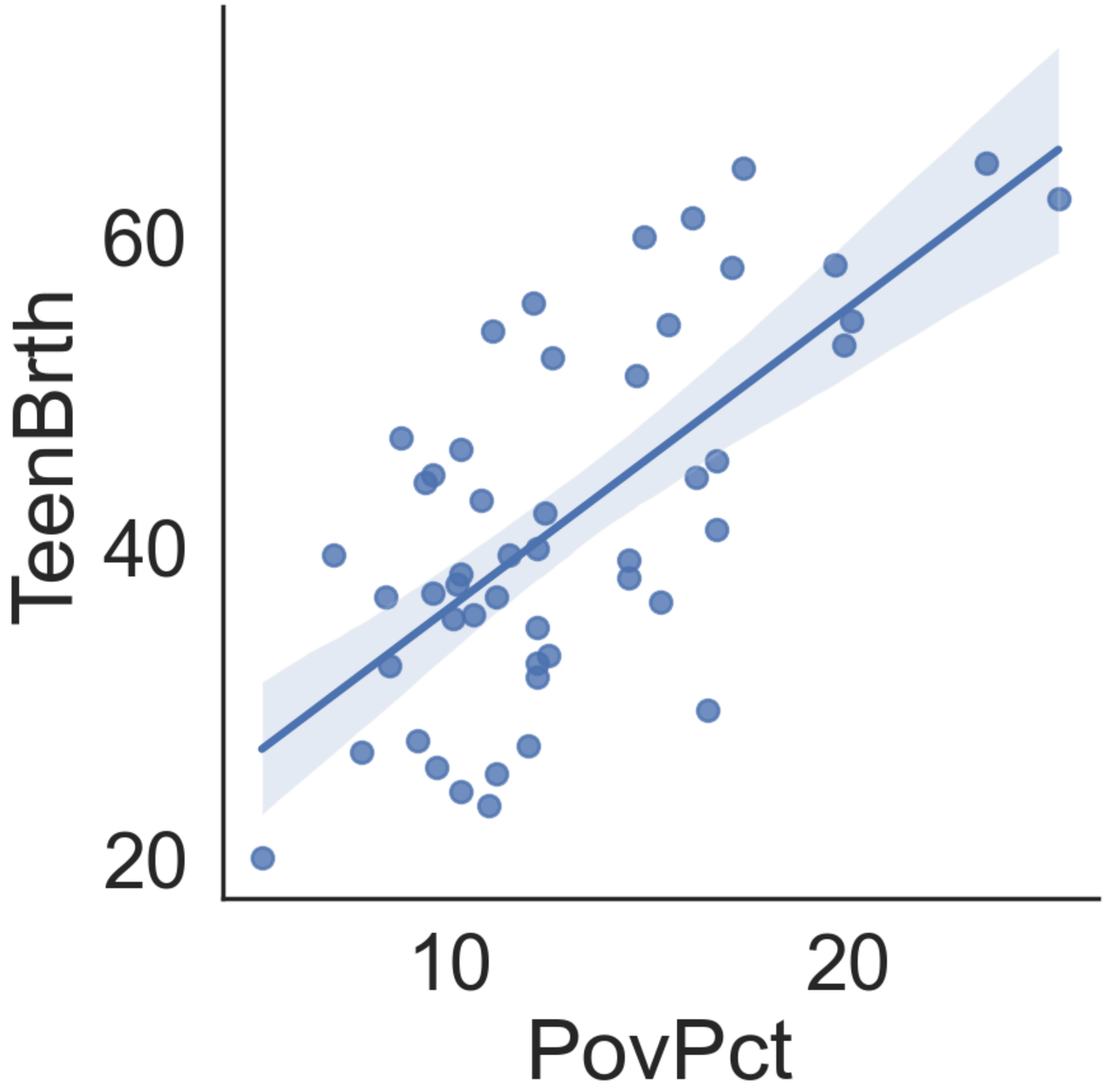
1. Linear relationship
2. Multivariate normality
3. No multicollinearity
4. No autocorrelation
5. Homoscedasticity


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

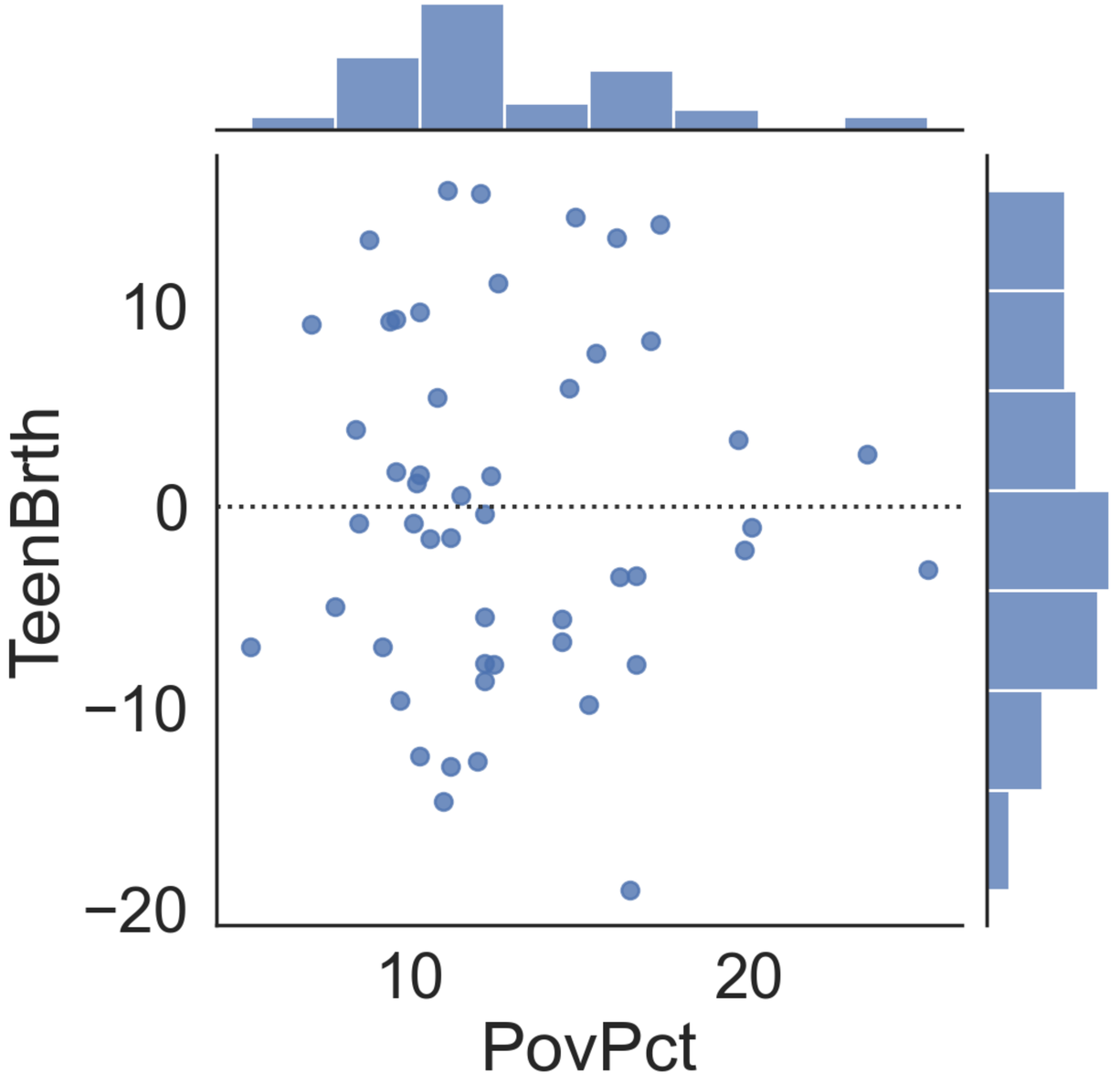
A multivariate normal probability distribution (joint normal)



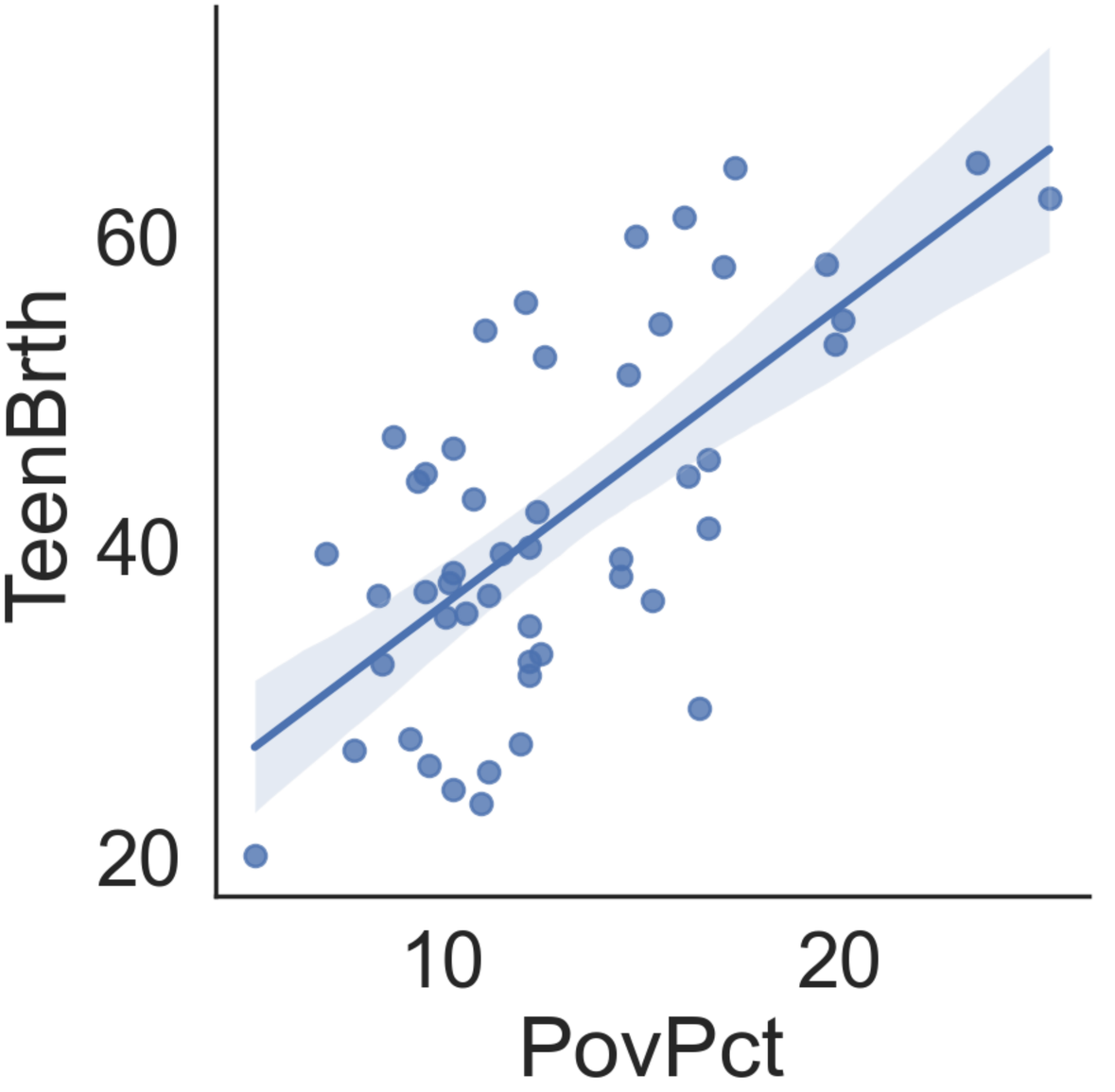

PovPct
Brth15to17
Brth18to19
ViolCrime
TeenBrth
10
20
30
50
100
5
15
25
50
60
40
20


TeenBrth
60
40
20
10
20
PovPct

```
sns.jointplot(df.drop('District_of_Columbia'), x='PovPct', y='TeenBrth', kind='resid');
```

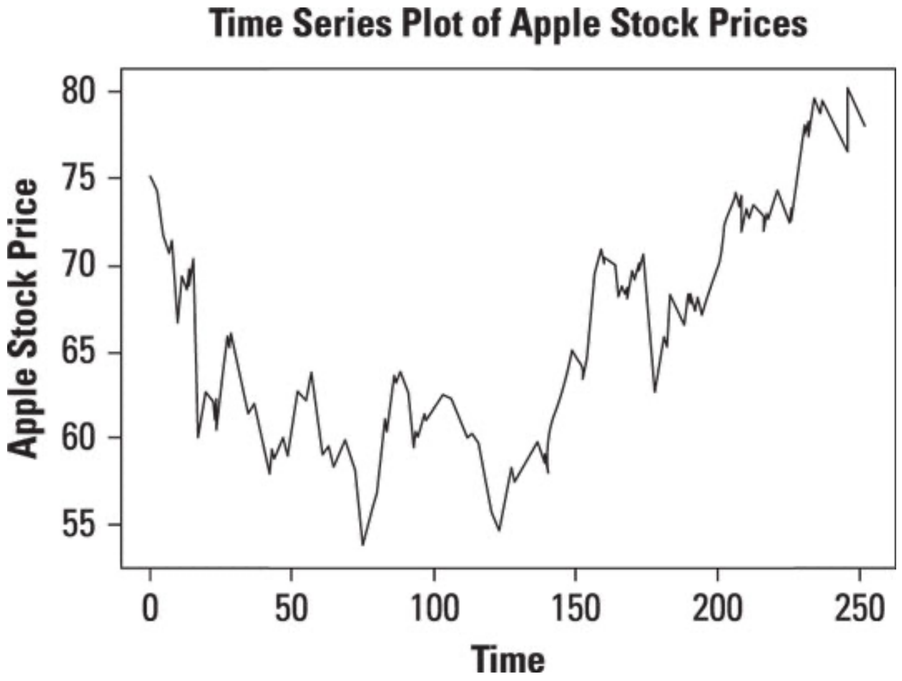


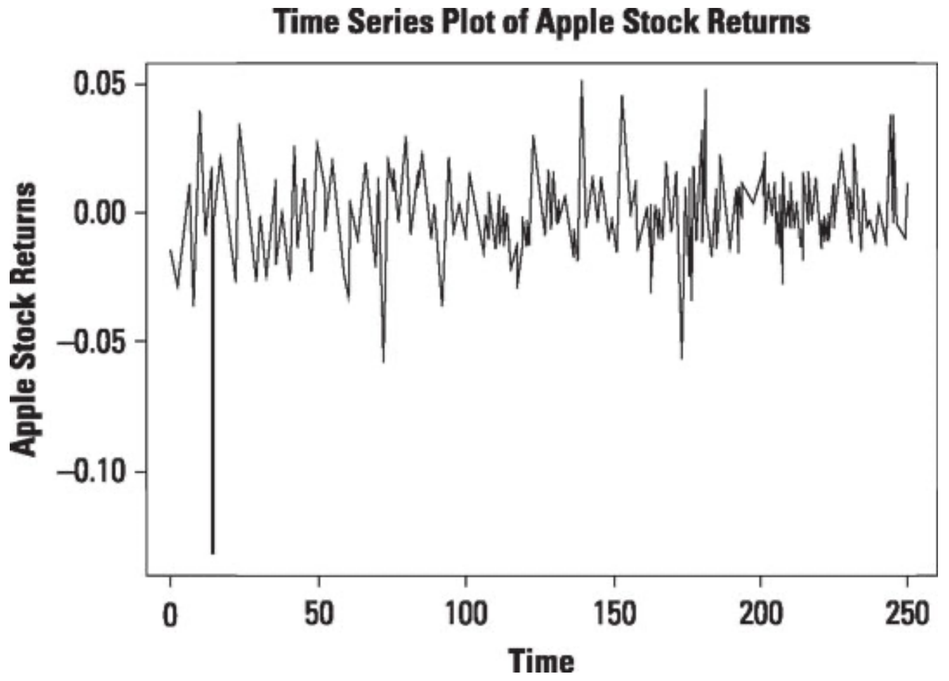

```
sns.lmplot(df.drop('District_of_Columbia'), x='PovPct', y='TeenBrth');
```



Linear regression assumes no multicollinearity. Multicollinearity occurs when the independent variables (in multiple linear regression) are too highly correlated with each other.

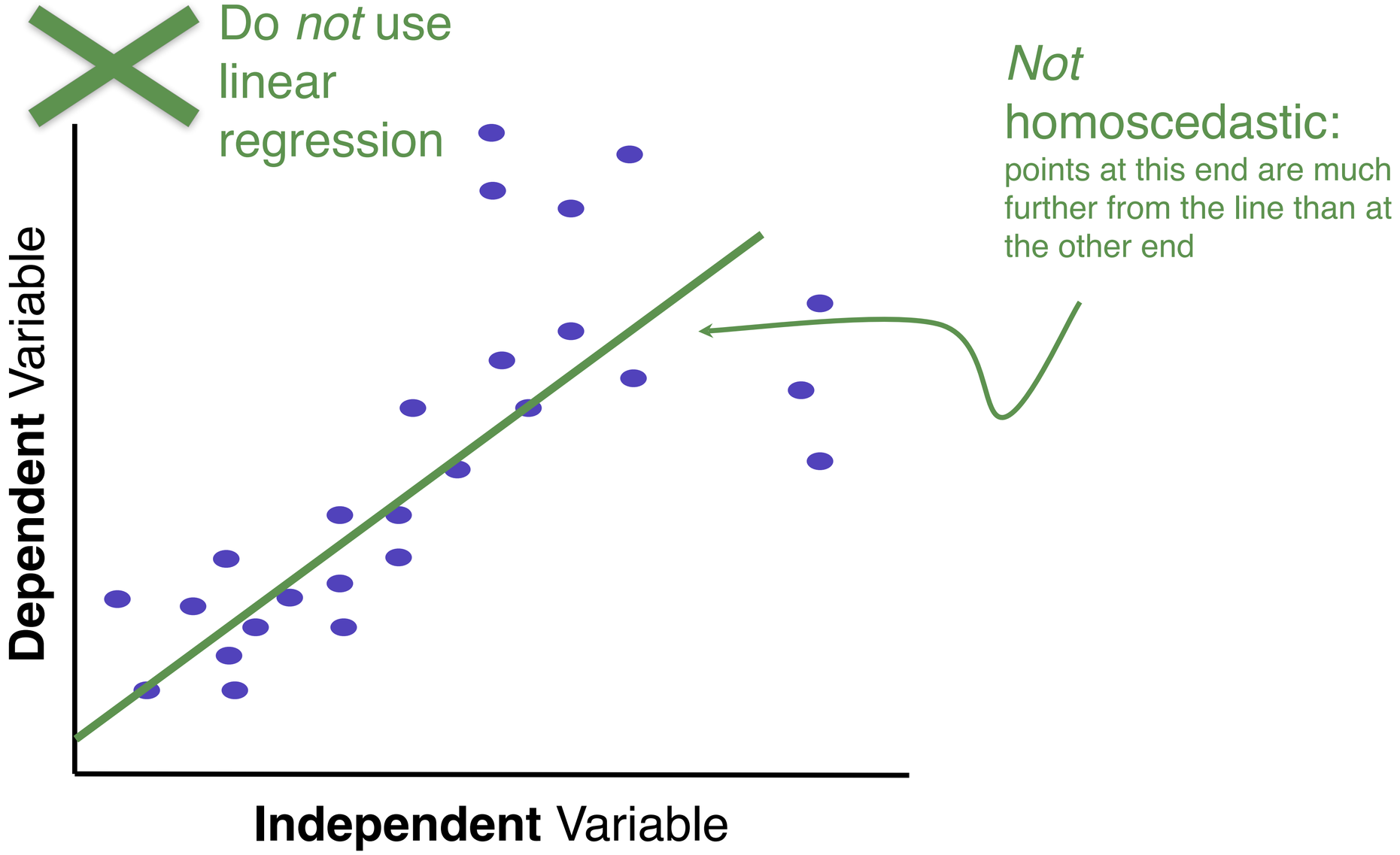
Daily returns are
$\ln(price_t / price_{t-1})$



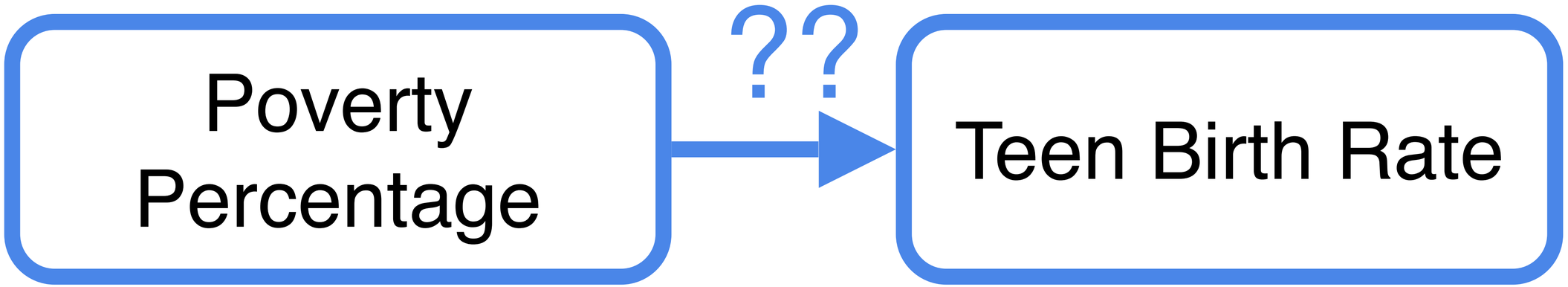


Autocorrelation occurs when the observations are *not* independent of one another (i.e. stock prices)

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

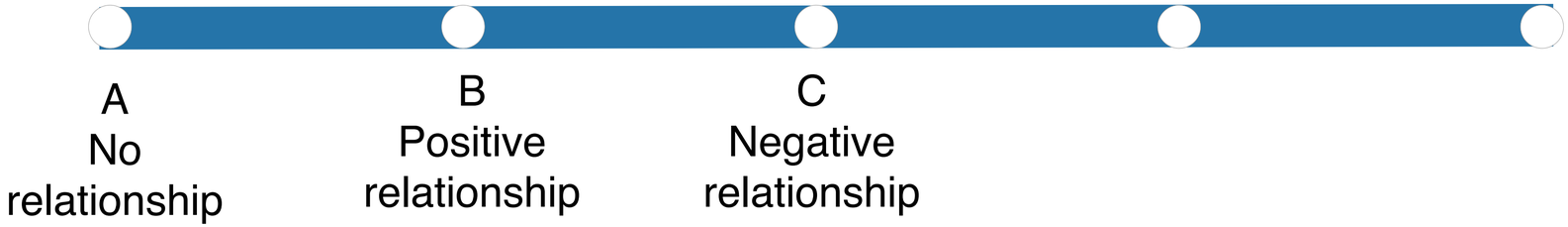

Do *not* use linear regression
Not homoscedastic:
points at this end are much further from the line than at the other end
Dependent Variable
Independent Variable

Does Poverty Percentage affect Teen Birth Rate?


Poverty Percentage
??
Teen Birth Rate

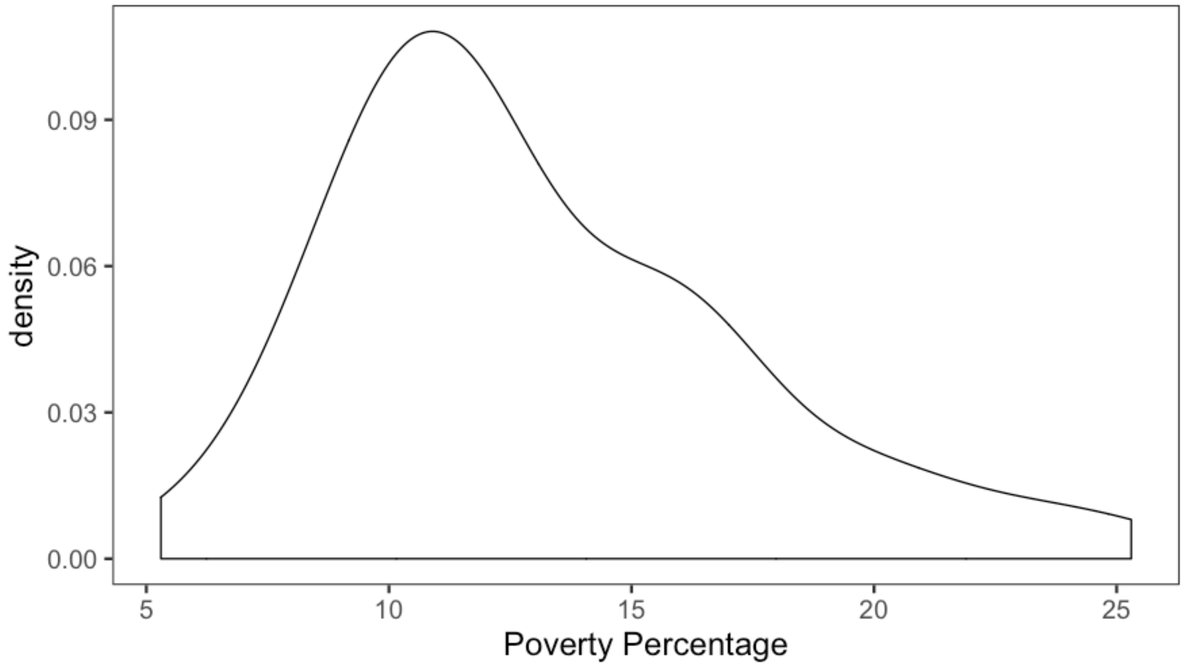
What is the relationship between Poverty Percentage & Teen Birth Rate?

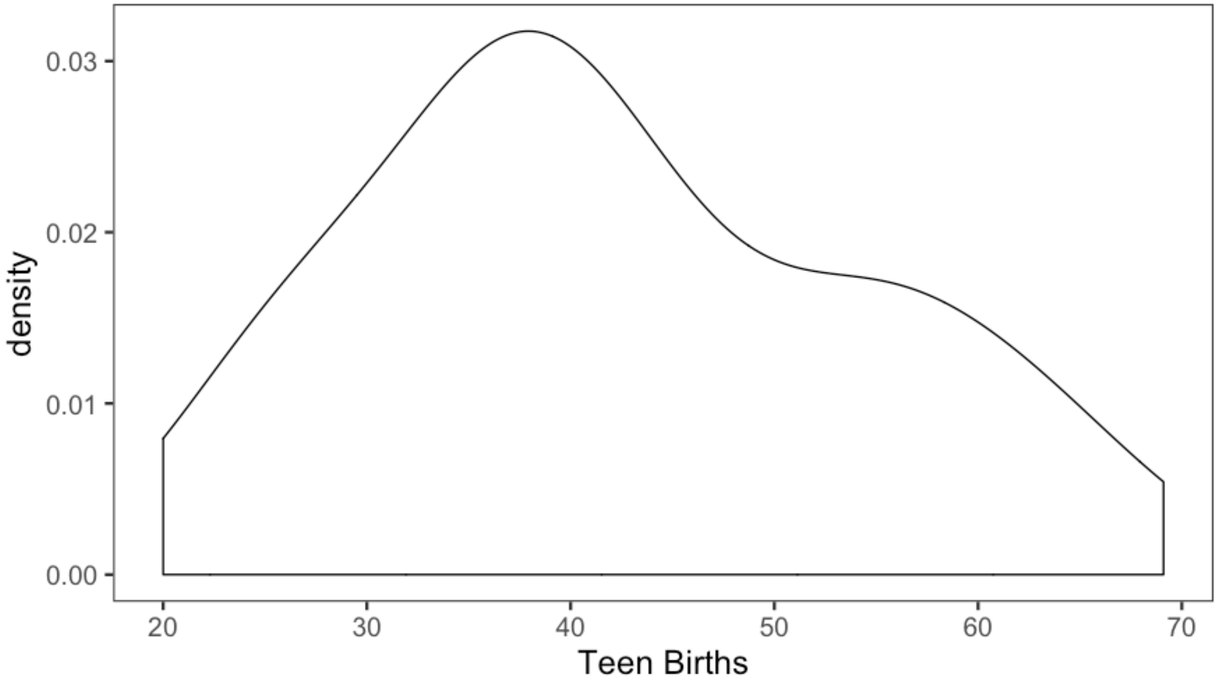
What's your hypothesis?

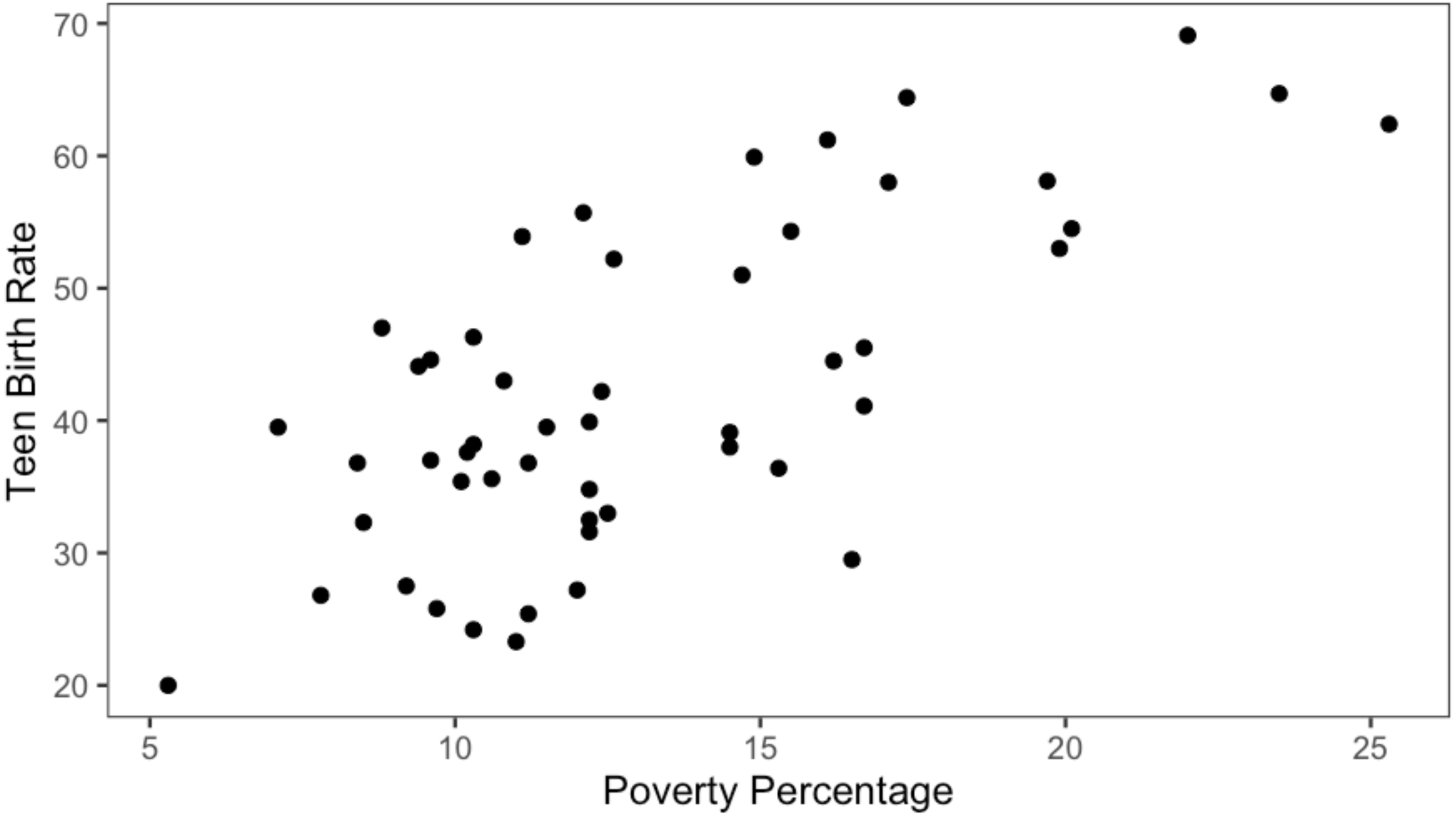


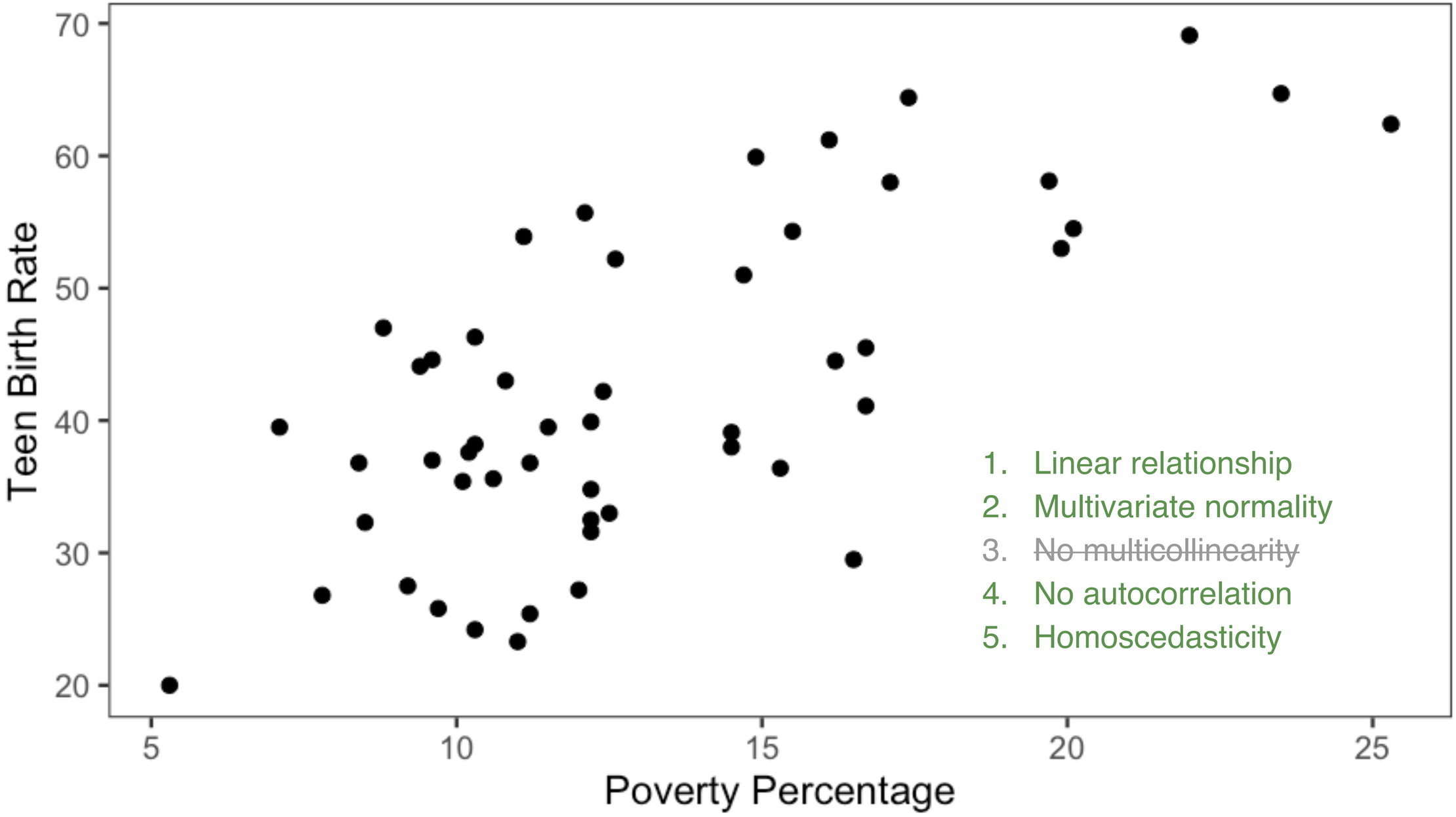
	Location	PovPct	Brth15to17	Brth18to19	ViolCrime	TeenBrth
1	Alabama	20.1	31.5	88.7	11.2	54.5
2	Alaska	7.1	18.9	73.7	9.1	39.5
3	Arizona	16.1	35.0	102.5	10.4	61.2
4	Arkansas	14.9	31.6	101.7	10.4	59.9
5	California	16.7	22.6	69.1	11.2	41.1
6	Colorado	8.8	26.2	79.1	5.8	47.0
7	Connecticut	9.7	14.1	45.1	4.6	25.8
8	Delaware	10.3	24.7	77.8	3.5	46.3
9	District_of_Columbia	22.0	44.8	101.5	65.0	69.1
10	Florida	16.2	23.2	78.4	7.3	44.5
11	Georgia	12.1	31.4	92.8	9.5	55.7
12	Hawaii	10.3	17.7	66.4	4.7	38.2
13	Idaho	14.5	18.4	69.1	4.1	39.1
14	Illinois	12.4	23.4	70.5	10.3	42.2
15	Indiana	9.6	22.6	78.5	8.0	44.6
16	Iowa	12.2	16.4	55.4	1.8	32.5
17	Kansas	10.8	21.4	74.2	6.2	43.0

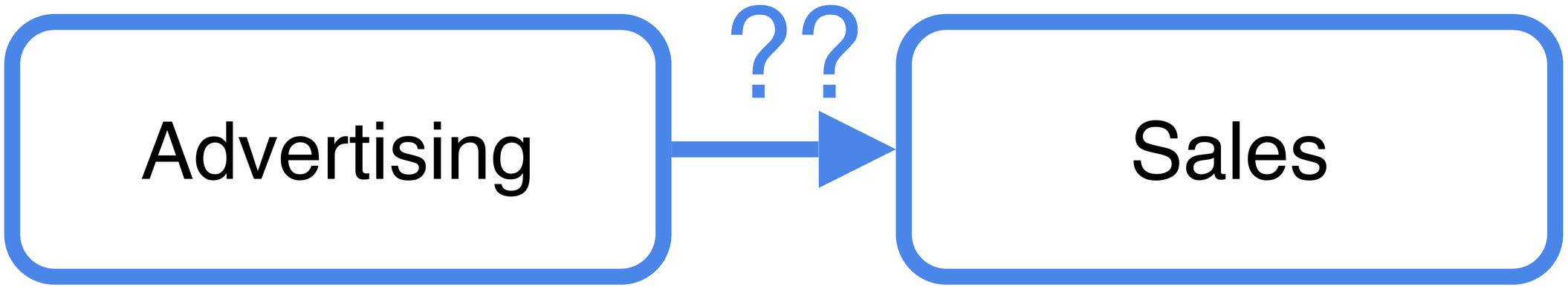
Normal(ish) distributions





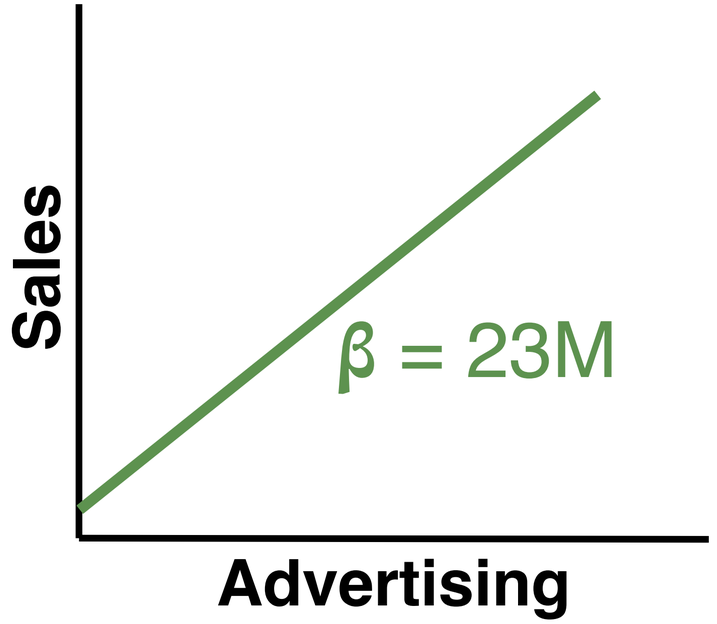

Teen Birth Rate
70
60
50
40
30
20
5
10
15
20
25
Poverty Percentage


Teen Birth Rate
70
60
50
40
30
20
5
10
15
20
25
Poverty Percentage
1. Linear relationship
2. Multivariate normality
3. ~~No multicollinearity~~
4. No autocorrelation
5. Homoscedasticity


Advertising
??
Sales

Effect size interpretation

Sales (Million Euro)	Advertising (Million Euro)
651	23
762	26
856	30
1,063	34
1,190	43
1,298	48
1,421	52
1,440	57
1,518	58



The effect size (β) between the advertising and sales is 23M. What does this mean?

A
For every 1M Euro spent on advertising, the company sees 23M more in sales

B
For every 1M Euro spent in sales, the company spends 23M more in advertising

C
For every 1M Euro spent on advertising, the company sees 24M less in sales

D
For every 1M Euro spent in sales, the company spends 23M less in advertising

Model selection problems

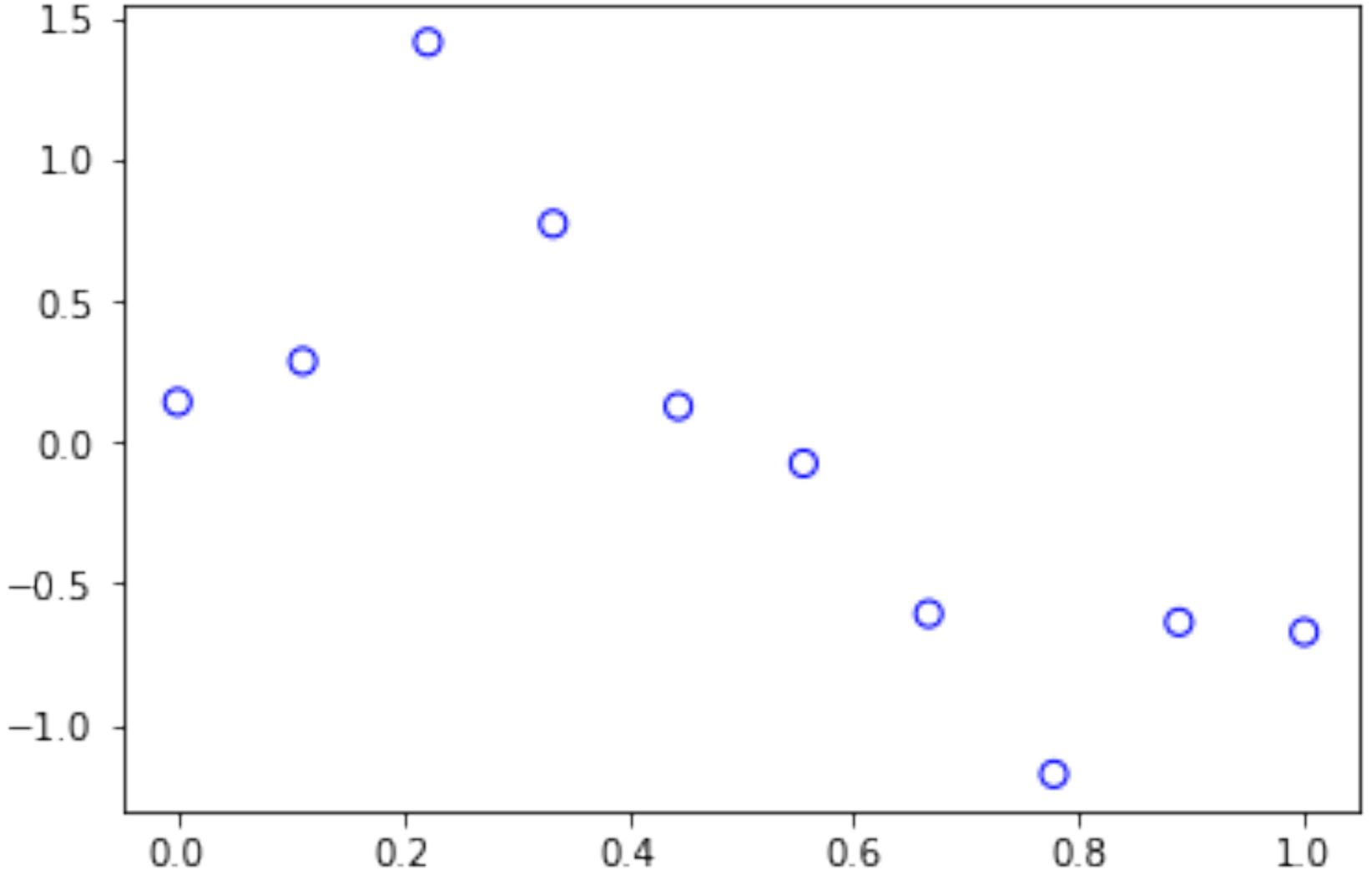
2.3 Parsimony

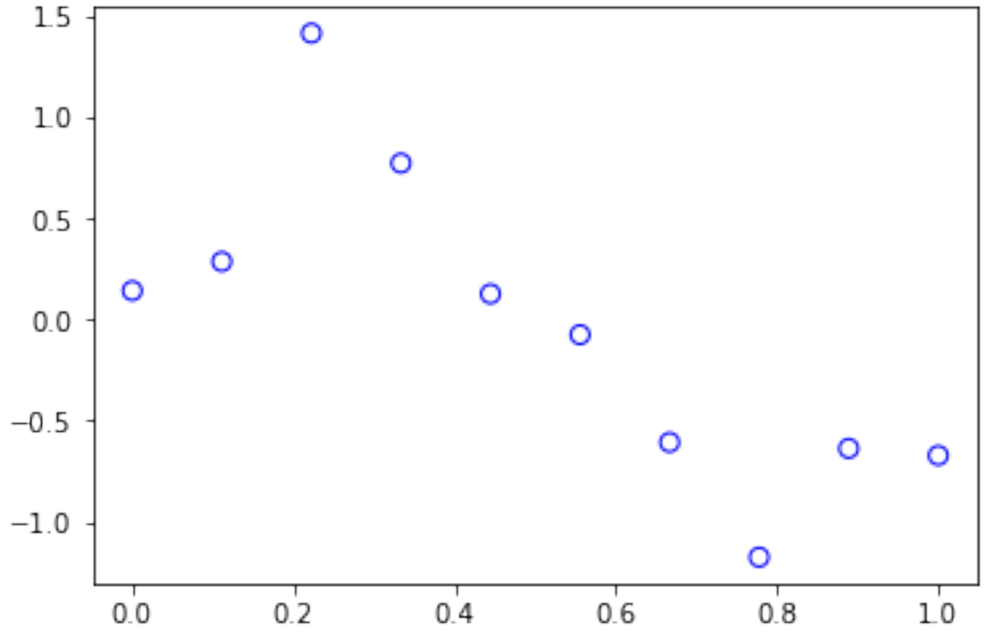
Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist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

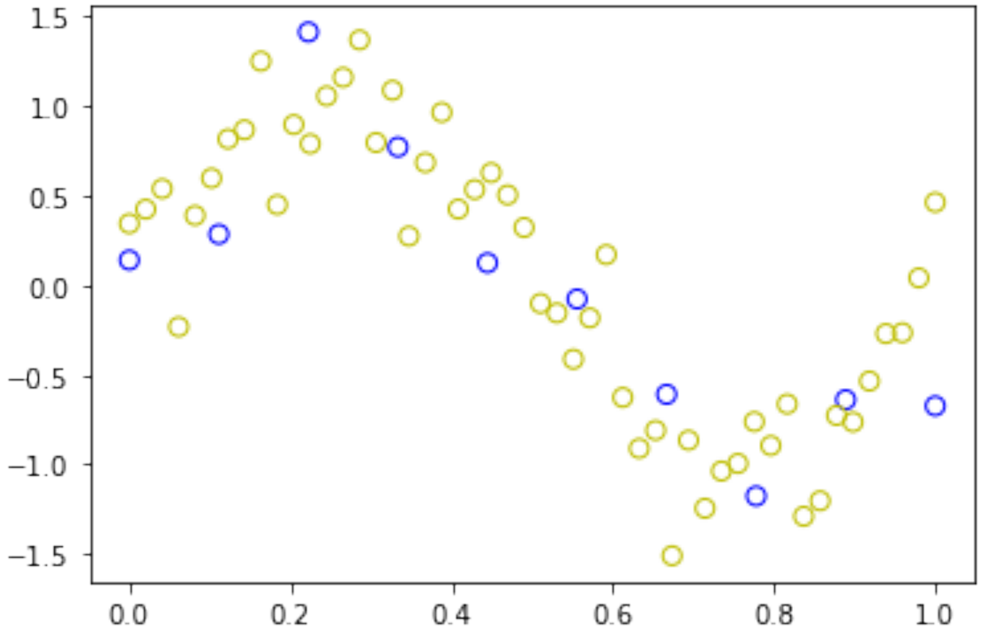
2.4 Worrying Selectively

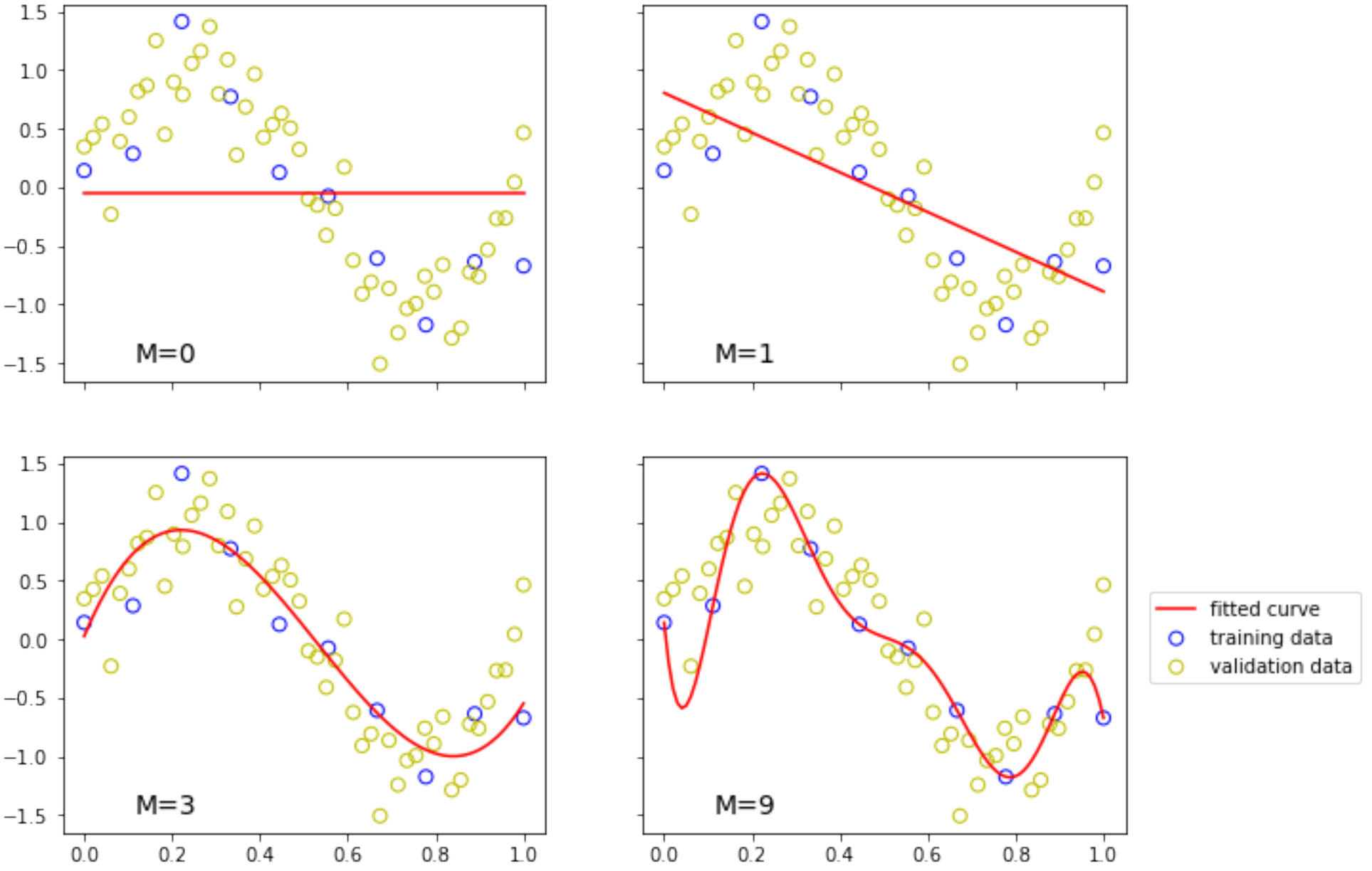
Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

George Box

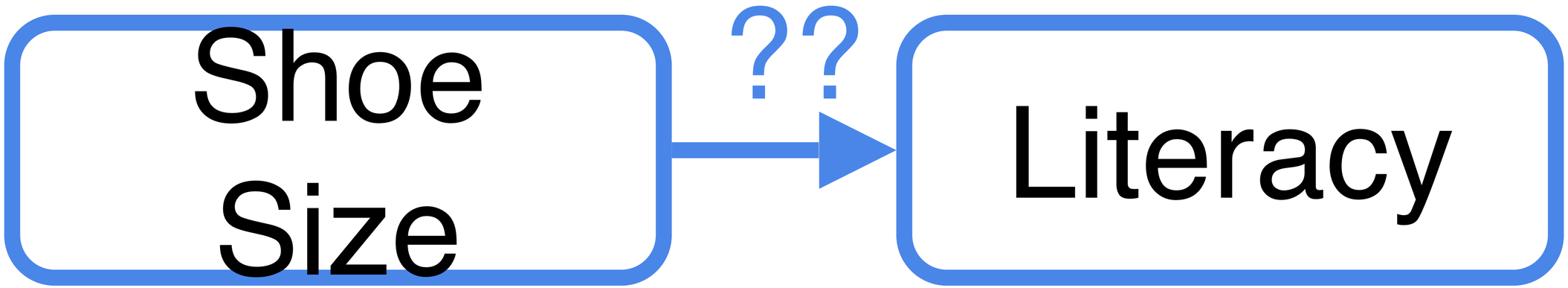

1.5
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0


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

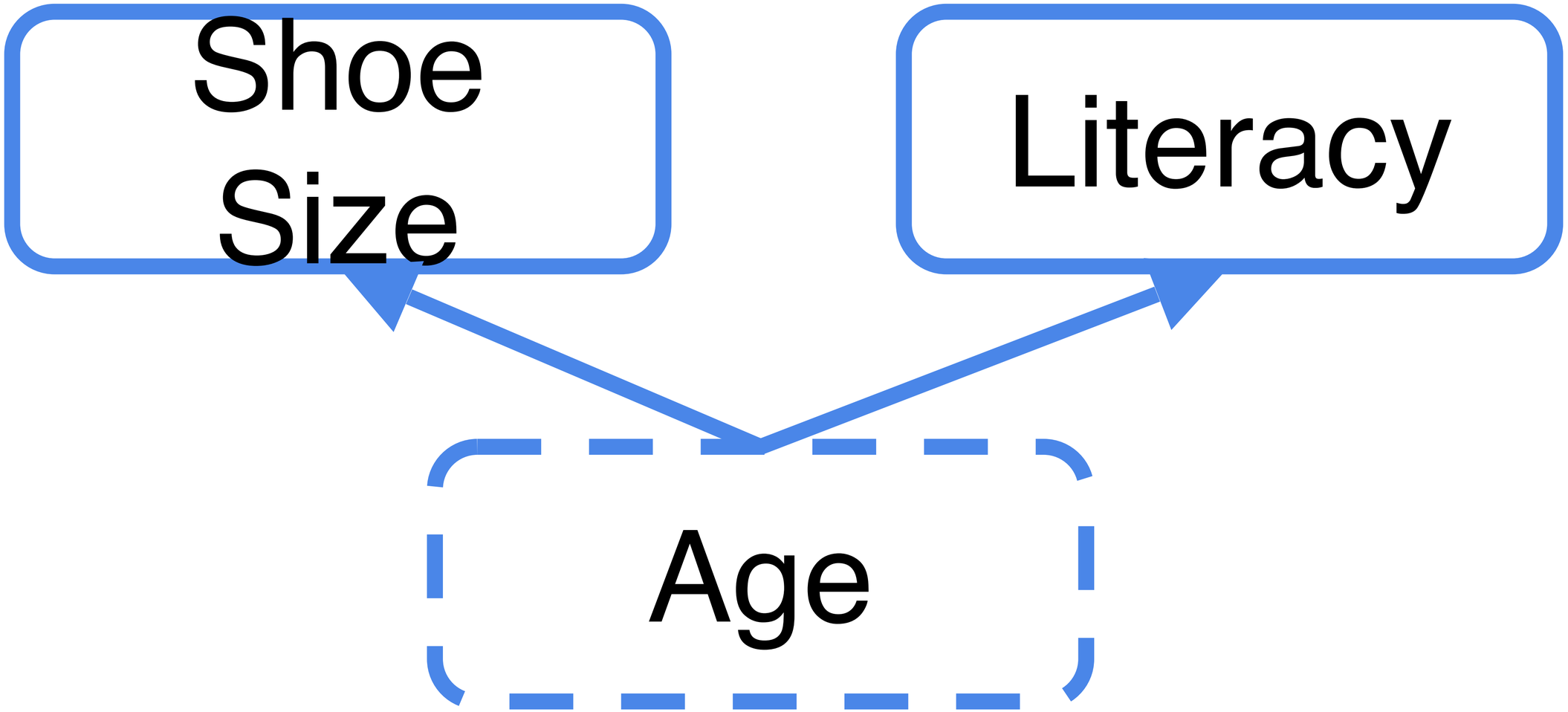

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

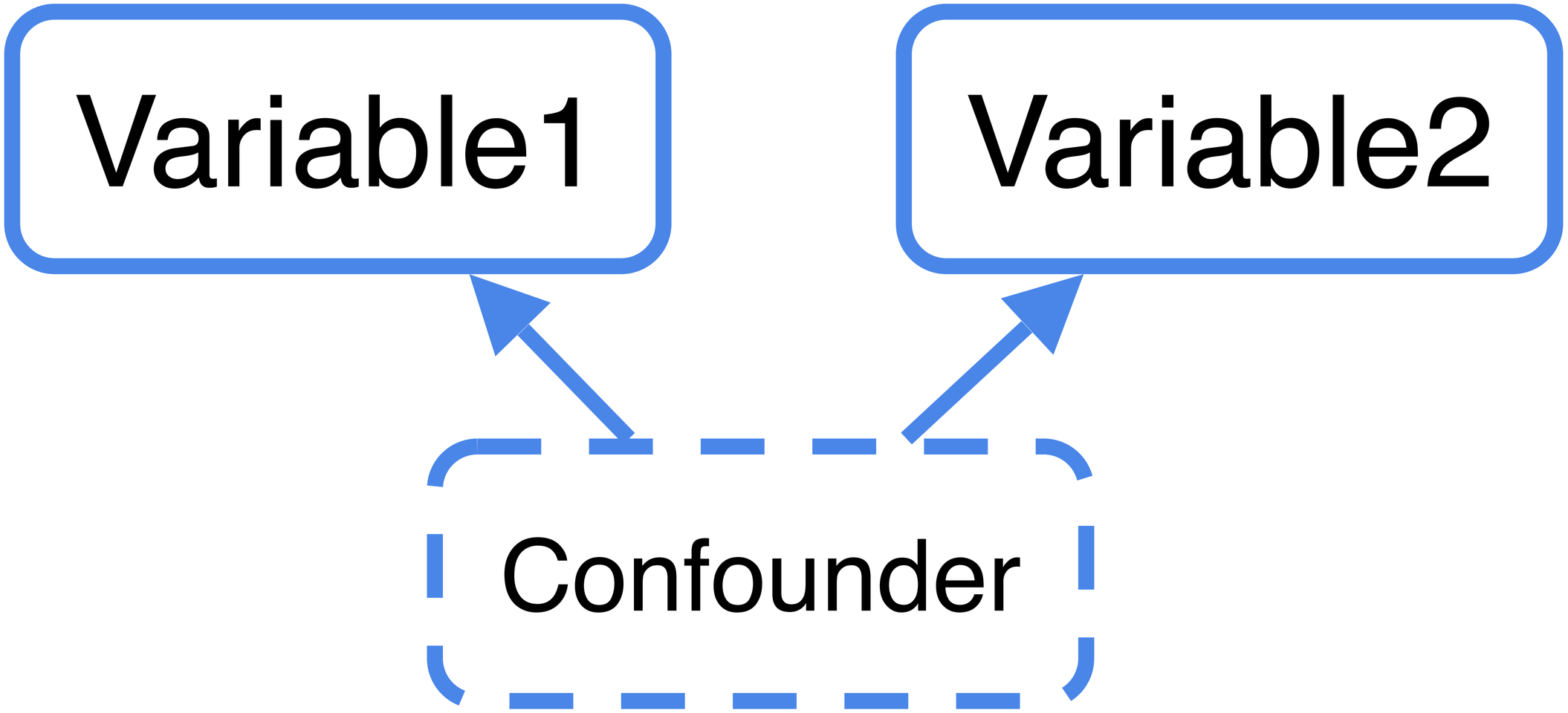

M=0
M=1
M=3
M=9
fitted curve
training data
validation data

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/