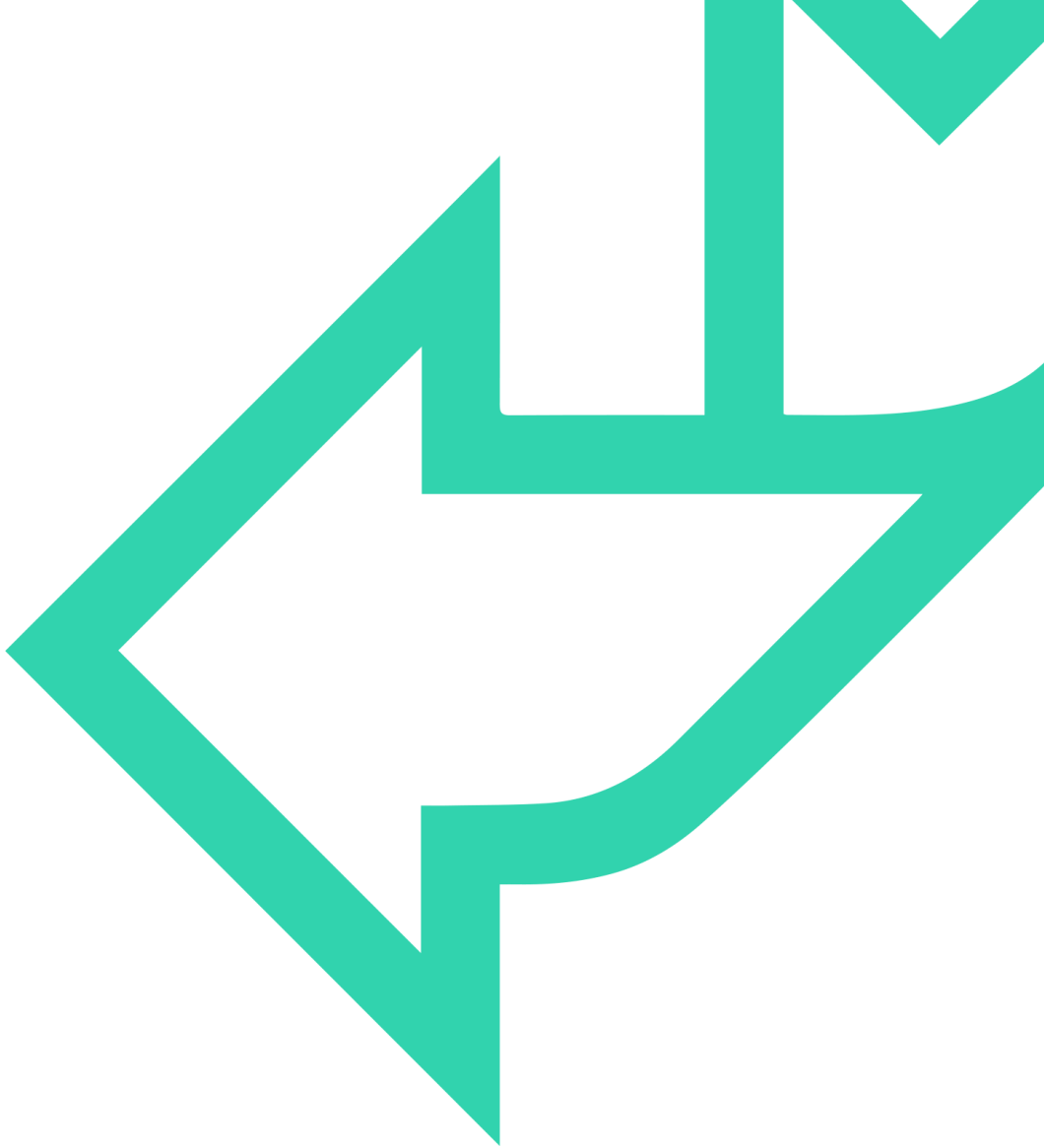
POWERING POTENTIAL

Linear Regression



Line Fitting

Modeling numerical variables

- quantify the relationship between two numerical variables
- modeling numerical response variables using a numerical or categorical explanatory variable

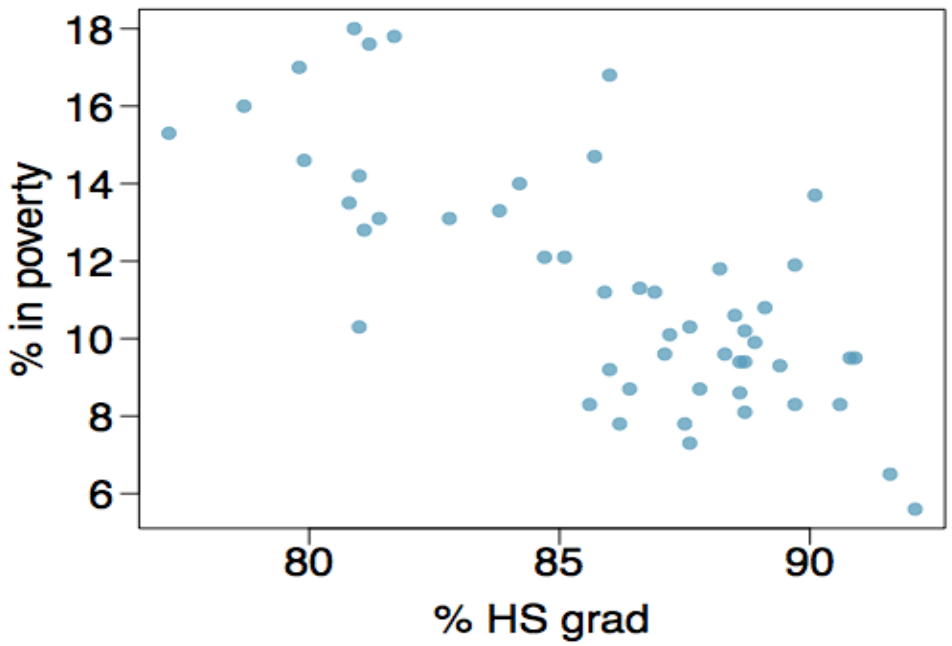
Scatterplot

The scatterplot below shows the relationship between

- HS graduate rate
  in all 50 US states and DC

and

- the percent of residents
  who live below the poverty line
  (income below $23,050
  for a family of 4 in 2012).

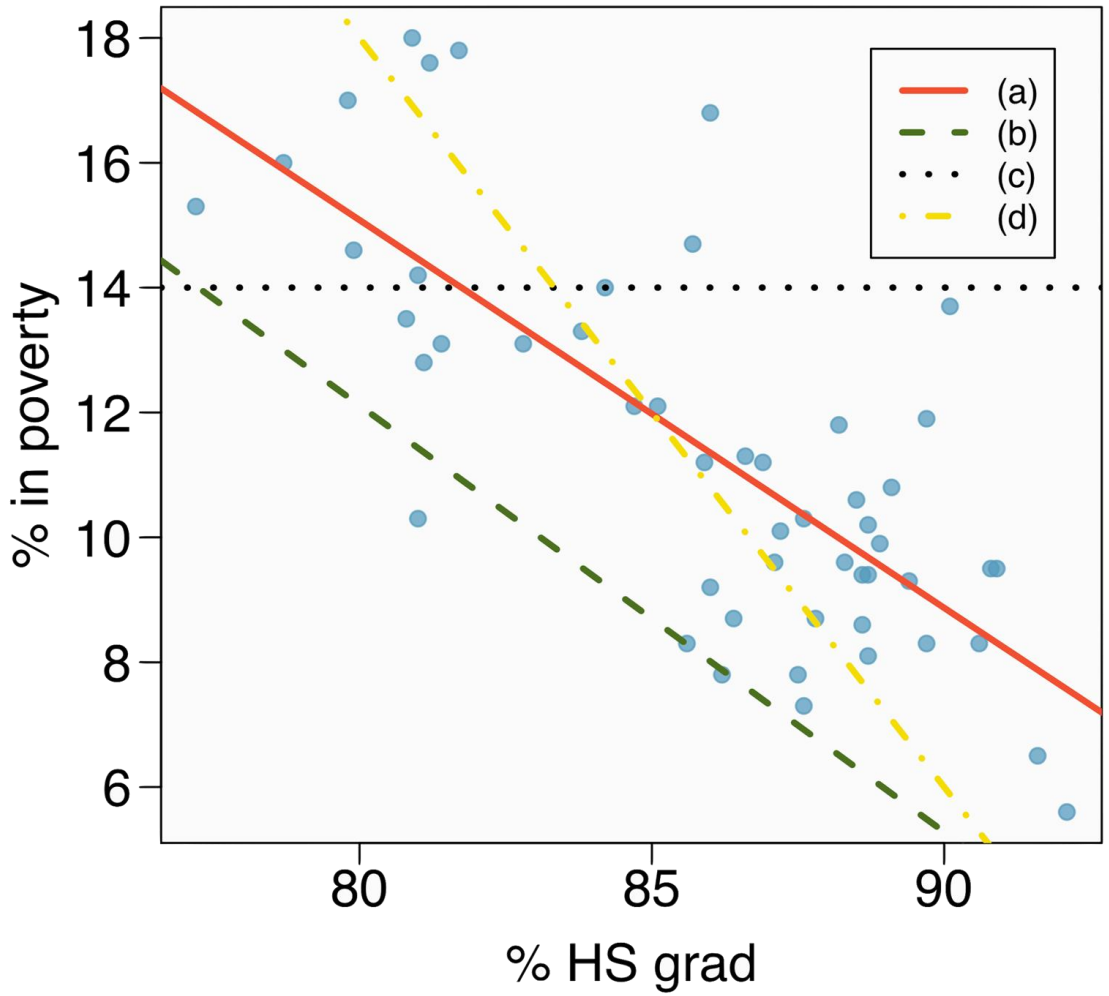


Response variable?
% in poverty
Explanatory variable?
% HS grad
Relationship?
linear, negative, moderately strong

Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?

(a)

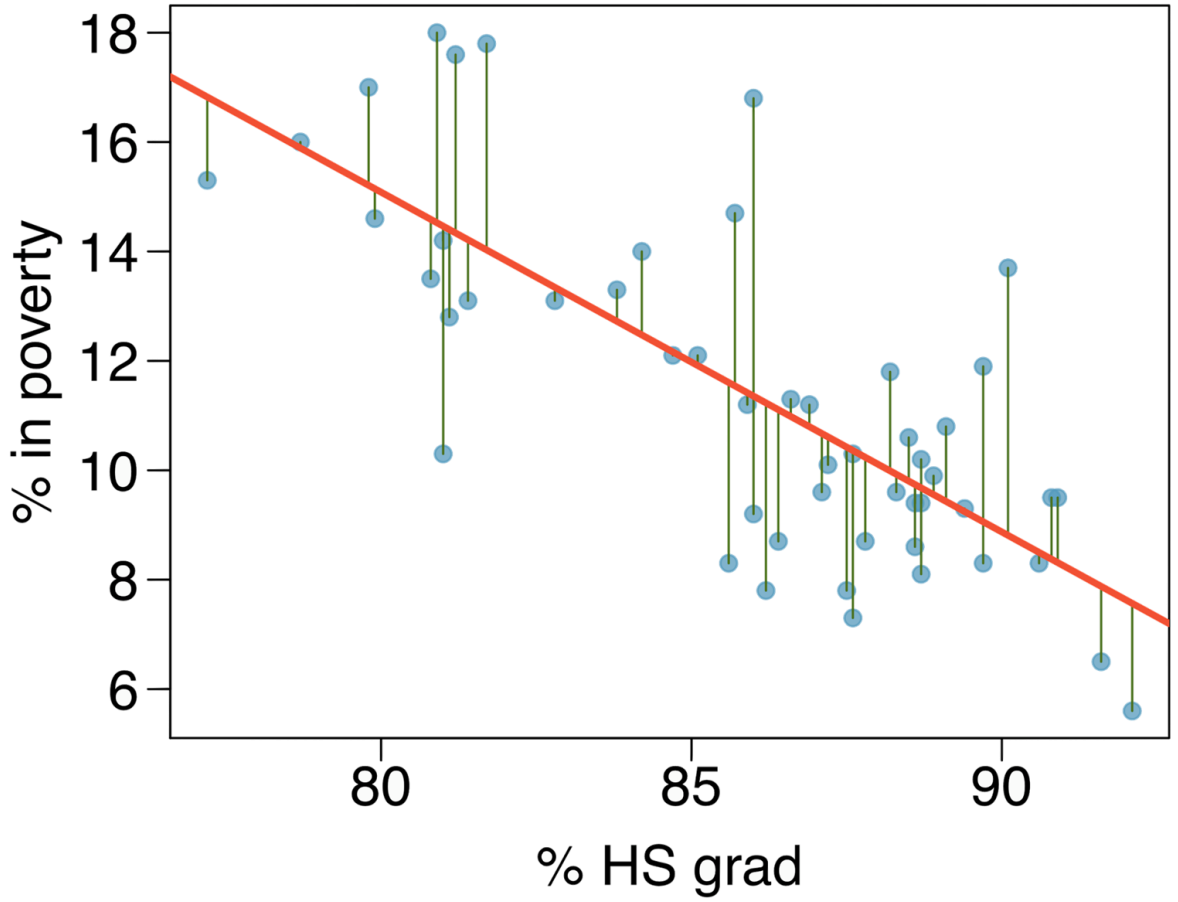


Residuals

Residuals

leftovers from the model fit:

Data = Fit + Residual

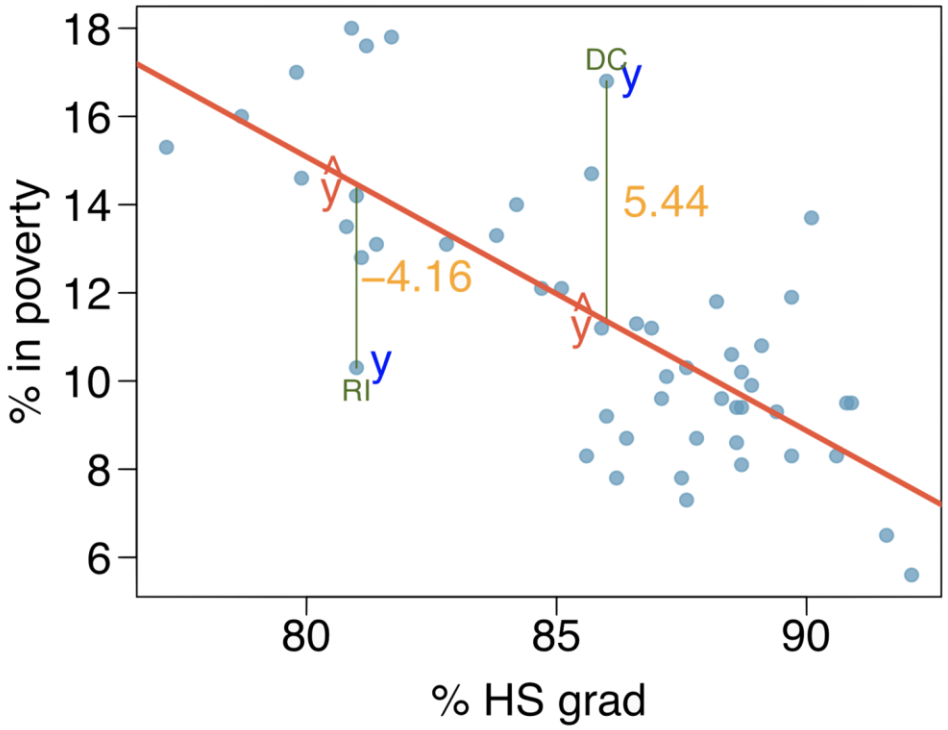


Residuals

$$e_i = y_i - \hat{y}_i$$

difference between the observed (y_i) and predicted $\hat{y}_i$

% living in poverty in RI is 4.16% less than predicted



% living in poverty in DC is 5.44% more than predicted

Correlation

Quantifying the relationship

- **Correlation** describes
  the strength of the **linear** association between two variables.

- It takes values between -1 (perfect negative) and +1 (perfect positive).

- A value of 0 indicates no linear association.

Quantifying the relationship

Covariance

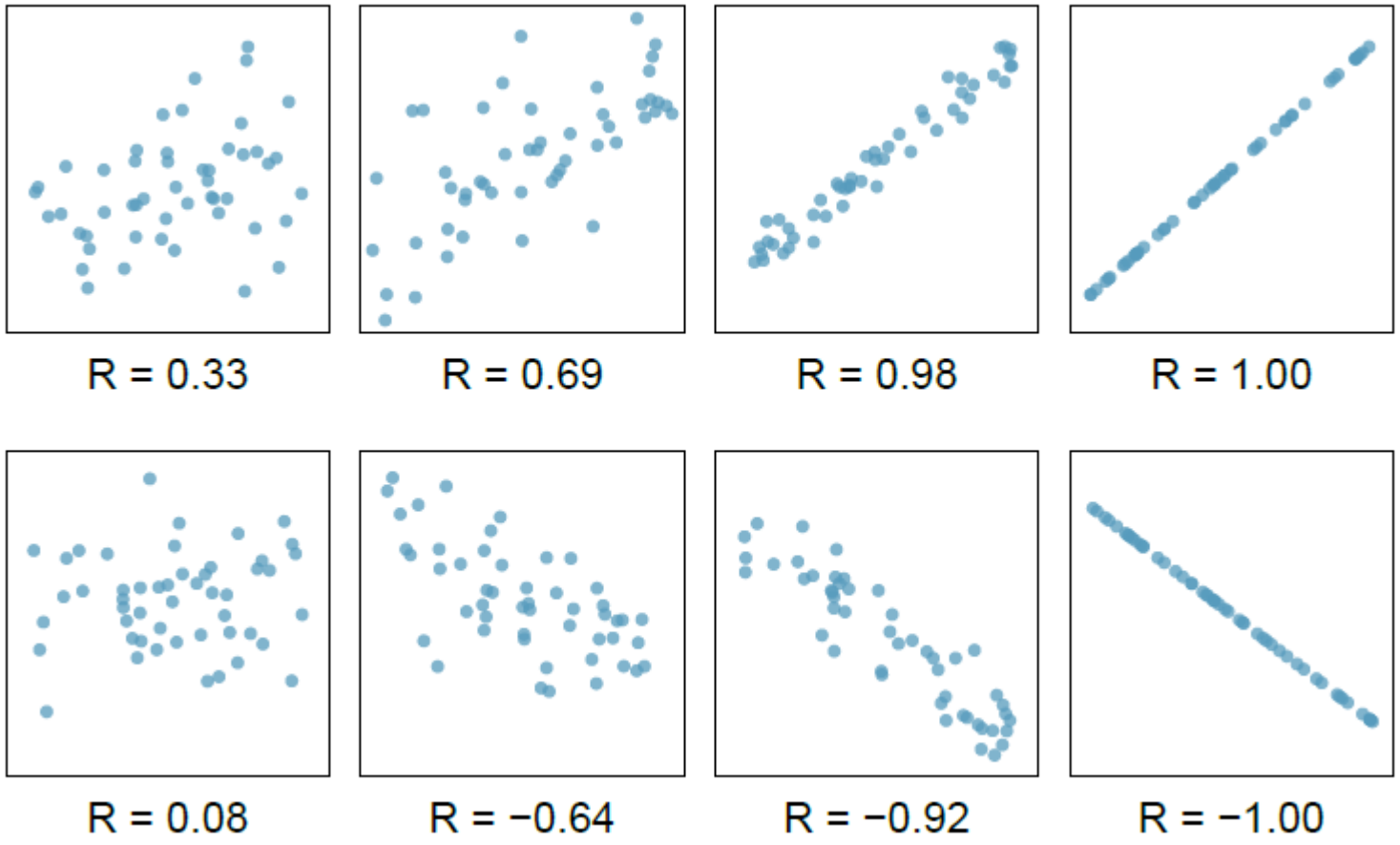
$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Quantifying the relationship

Correlation

$$\rho_{xy} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}}{\sigma_x \sigma_y}$$

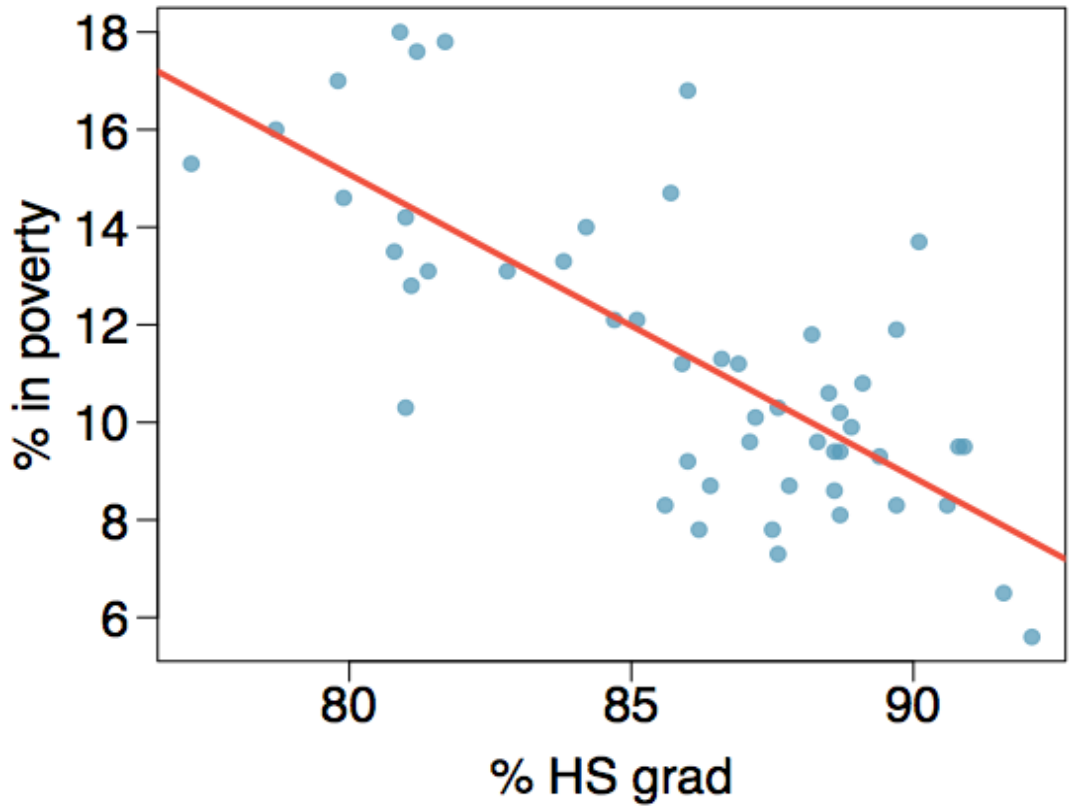
Quantifying the relationship



Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent HS grad?

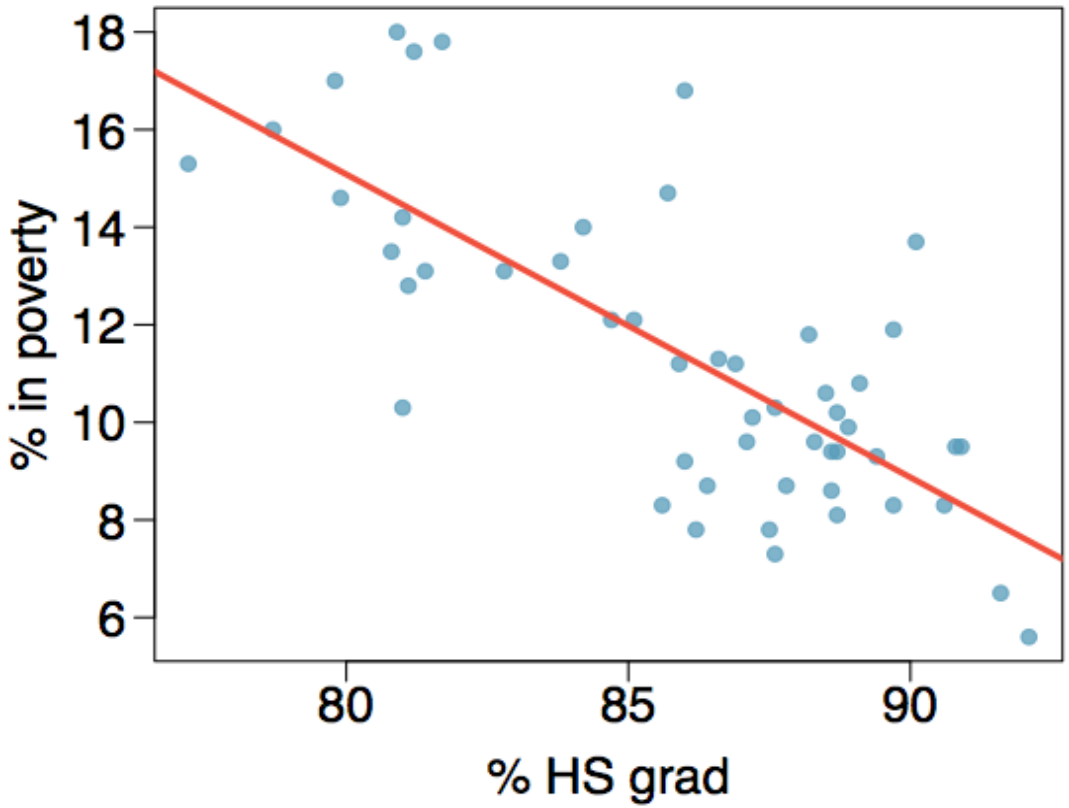
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent HS grad?

(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent female householder?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

Guessing the correlation

Which of the following is the best guess for the correlation between percent in poverty and percent female householder?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

Possums: True/False?

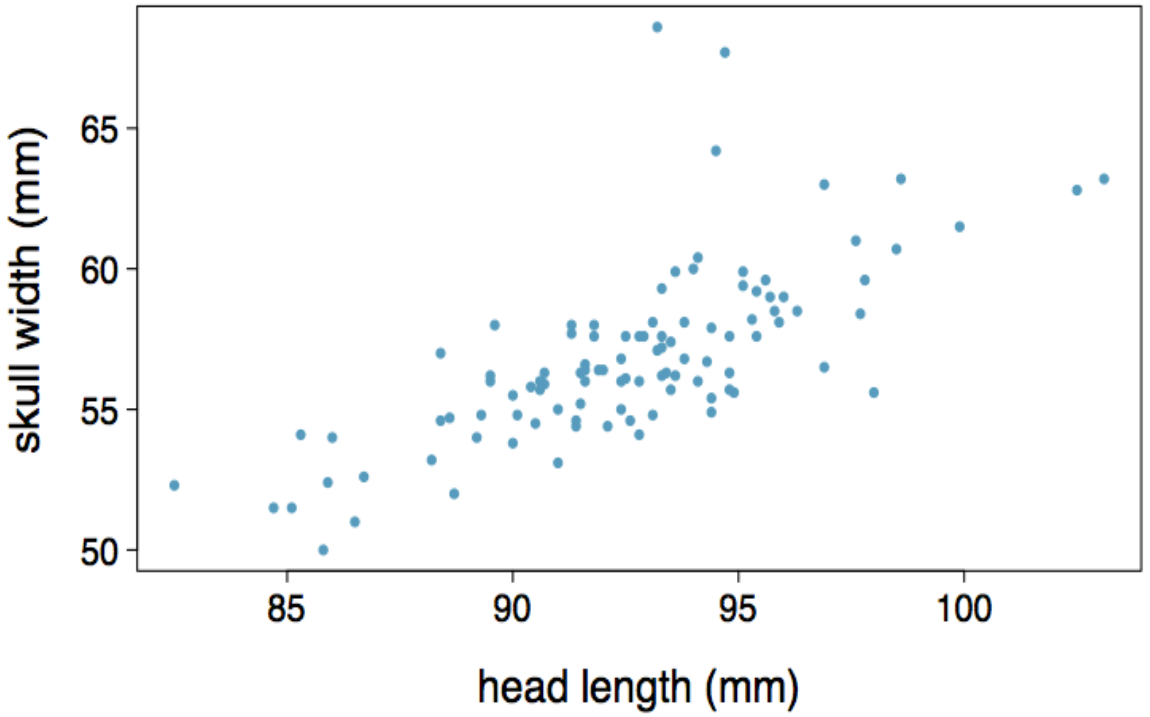
(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.



Possums: True/False?

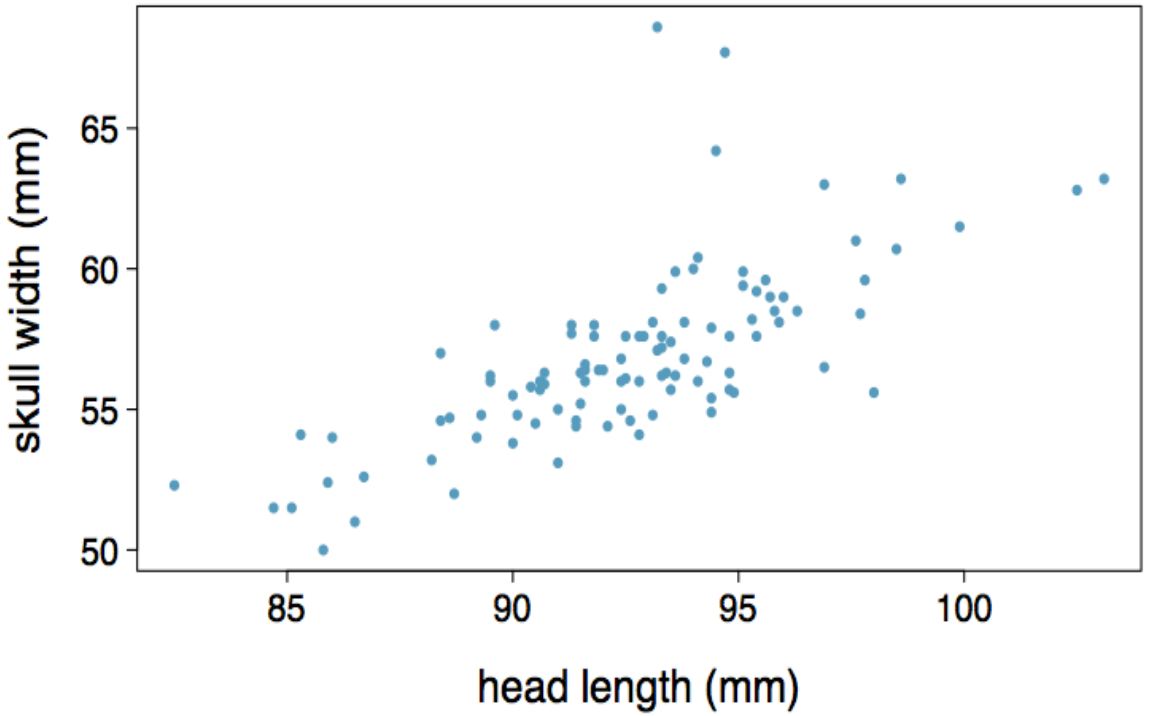
(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.



Assessing the correlation

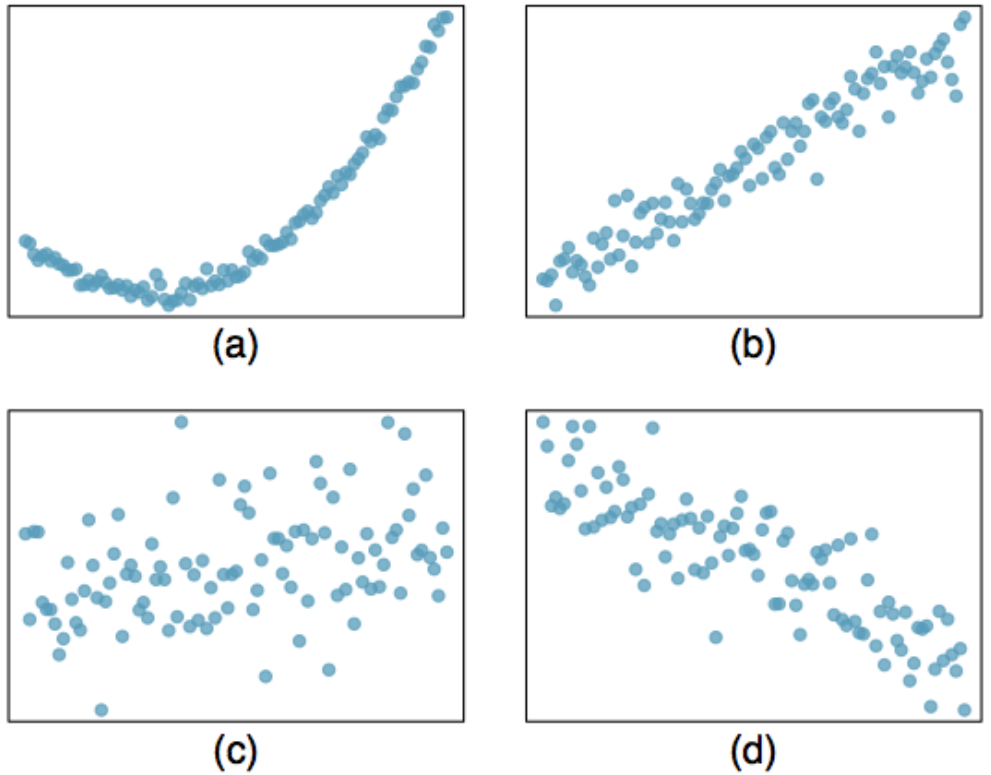
Which of the following is has the strongest correlation?

i.e. correlation coefficient closest to +1 or -1?

(b)

Why?

Correlation means LINEAR association



Fitting a line by least squares regression

A measure for the best line

- We want a line that has **small residuals**

  Option 1: Minimize the sum of magnitudes (absolute values) of residuals

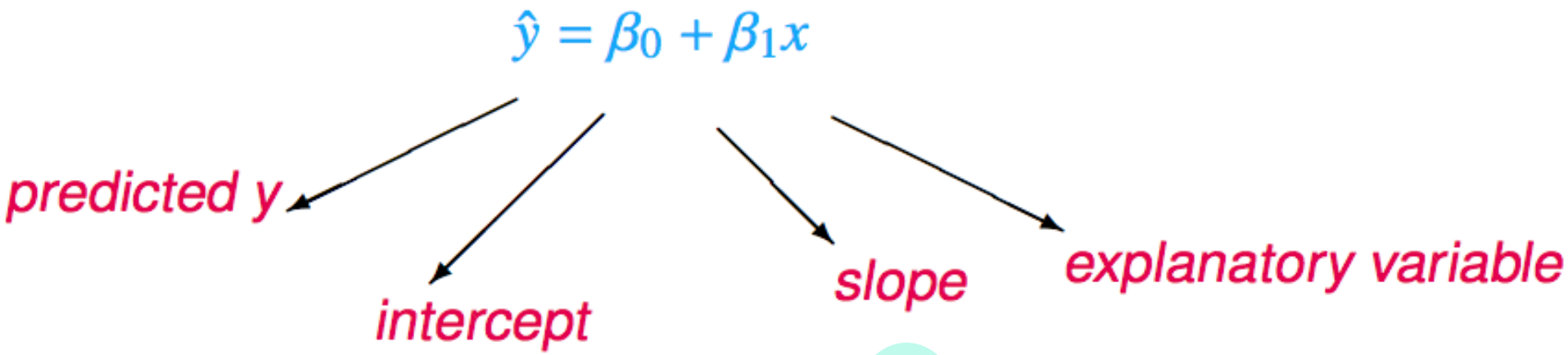
  $$|e_1| + |e_2| + \ldots + |e_n|$$

  Option 2: Minimize the sum of squared residuals – **"least squares"**

  $$e_1^2 + e_2^2 + \ldots + e_n^2$$

- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software
  3. In many applications, a residual twice as large as another is usually more than twice as bad

least squares line

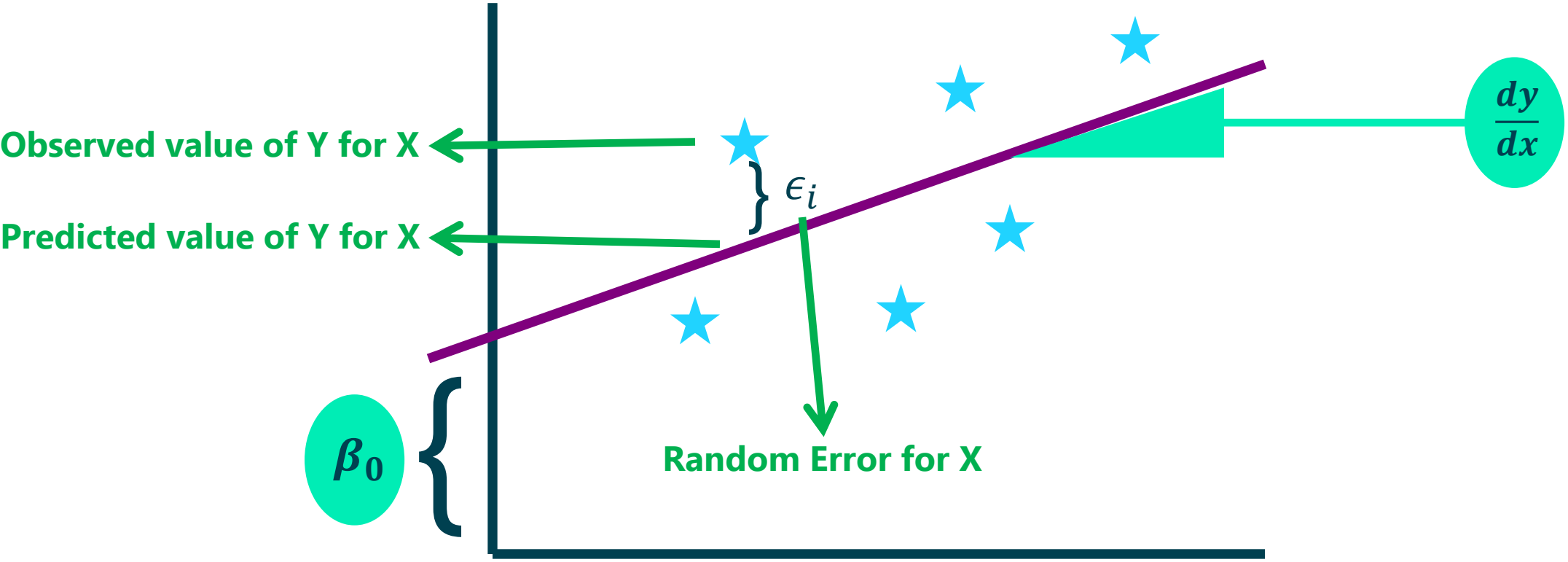


Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

regression **model**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



least squares line: Conditions

1. Linearity
2. Nearly normal residuals
3. Constant variability

LS line: Condition: Linearity

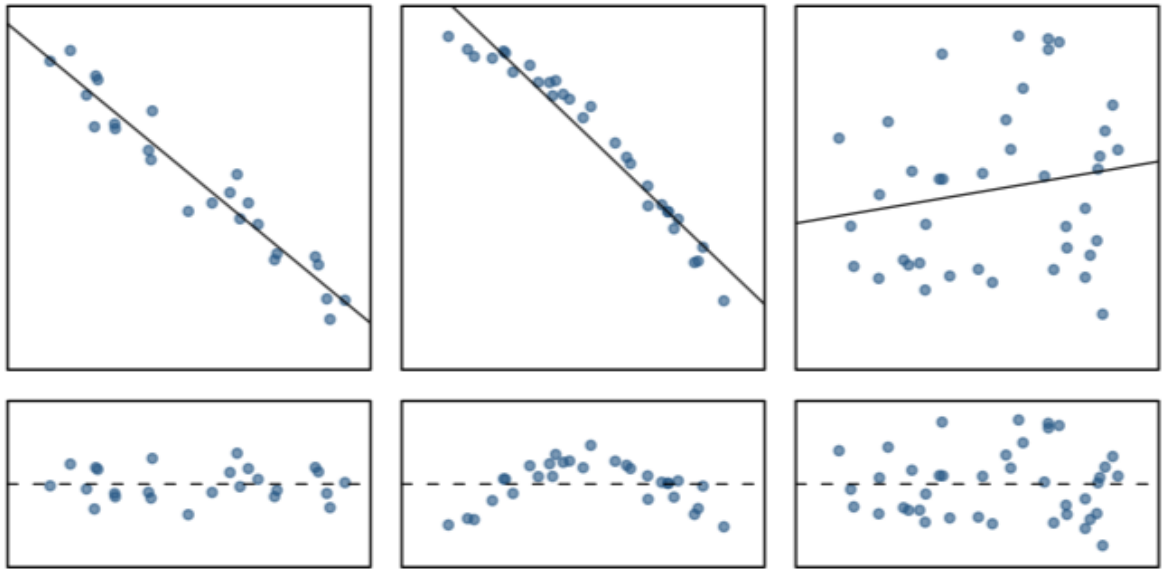
The relationship between
the explanatory and
the response variable
should be **linear**.

How to check?

a scatterplot

a residuals *plot*.

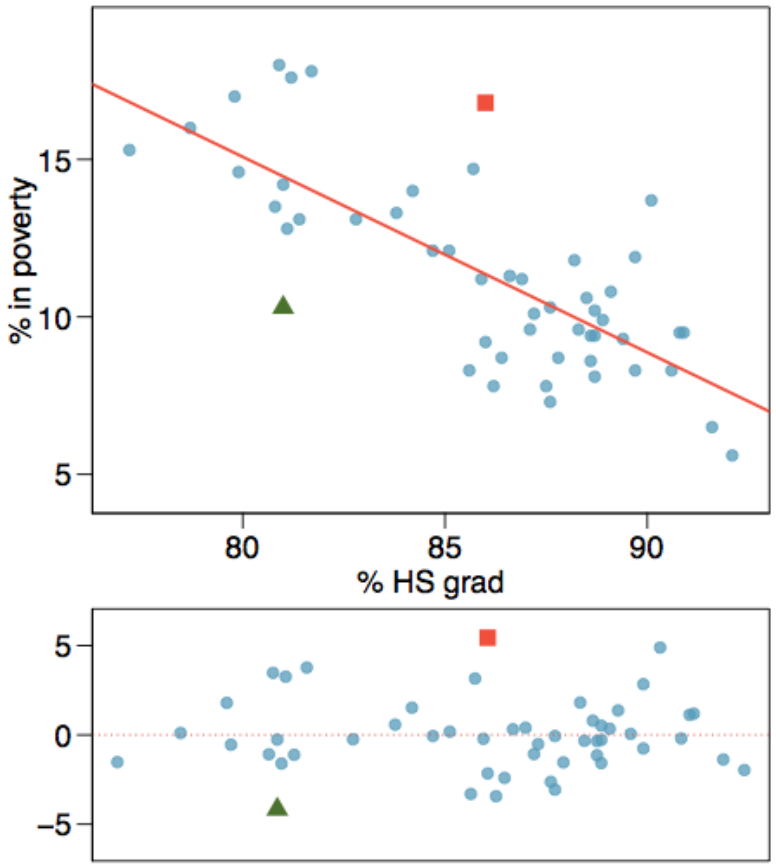
- *anatomy explained on next slide*



Anatomy of a residuals plot

▲ *RI:*

$$\% \text{ HS grad} = 81 \qquad \% \text{ in poverty} = 10.3$$
$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 81 = 14.46$$
$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$
$$= 10.3 - 14.46 = -4.16$$



■ *DC:*

$$\% \text{ HS grad} = 86 \qquad \% \text{ in poverty} = 16.8$$
$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 * 86 = 11.36$$
$$e = \% \text{ in poverty} - \widehat{\% \text{ in poverty}}$$
$$= 16.8 - 11.36 = 5.44$$

LS line: Condition: normal ε_i

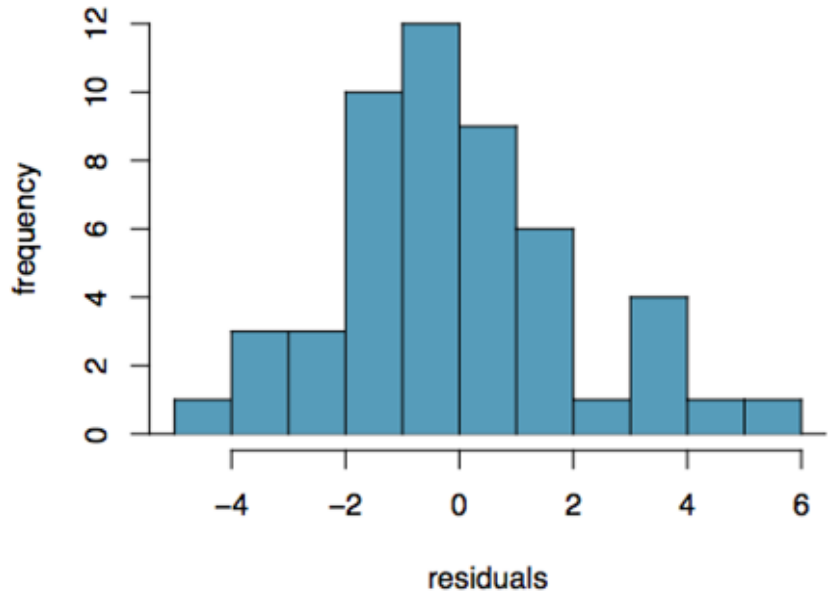
The residuals should be nearly **normal**.

(this condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data)

How to check?

using a **histogram of residuals**

[or using a normal probability plot:
qqnorm(model$residuals)
scipy.stats.probplot((z, plot=plt))]



LS line: Condition: constant σ^2

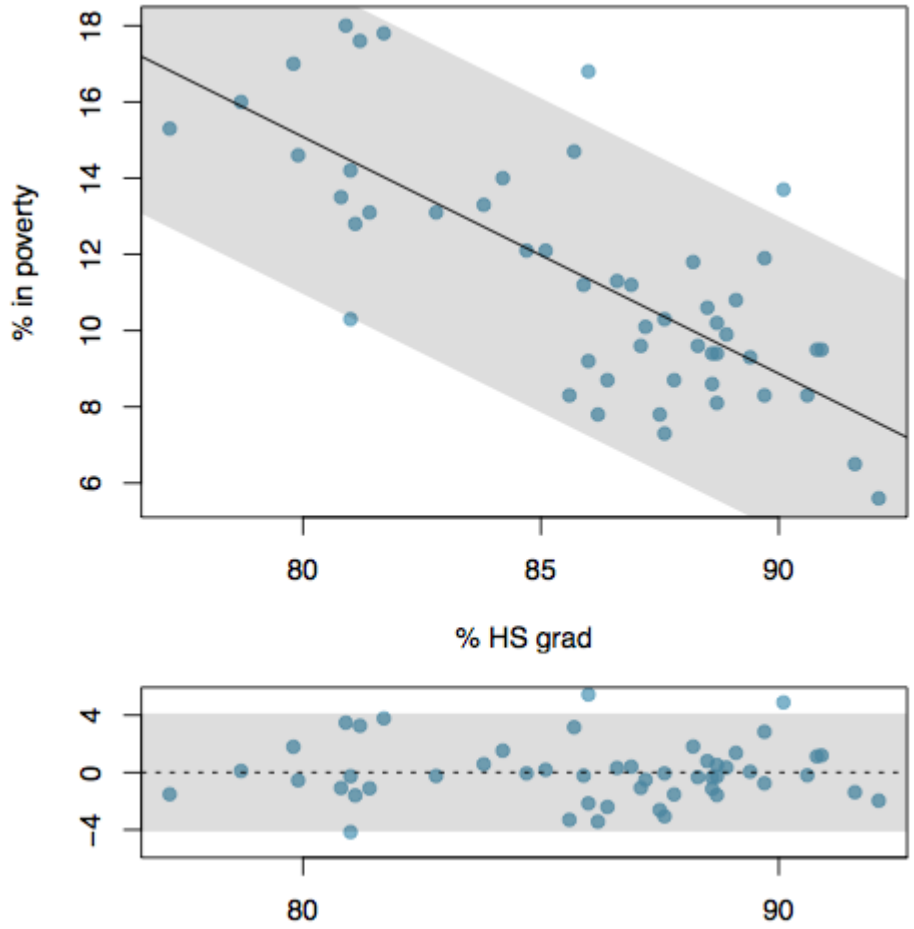
The **variability** of points around the least squares line should be roughly **constant**

The variability of **residuals around the 0** line should be roughly constant

a.k.a. “homoscedasticity”

How to check?

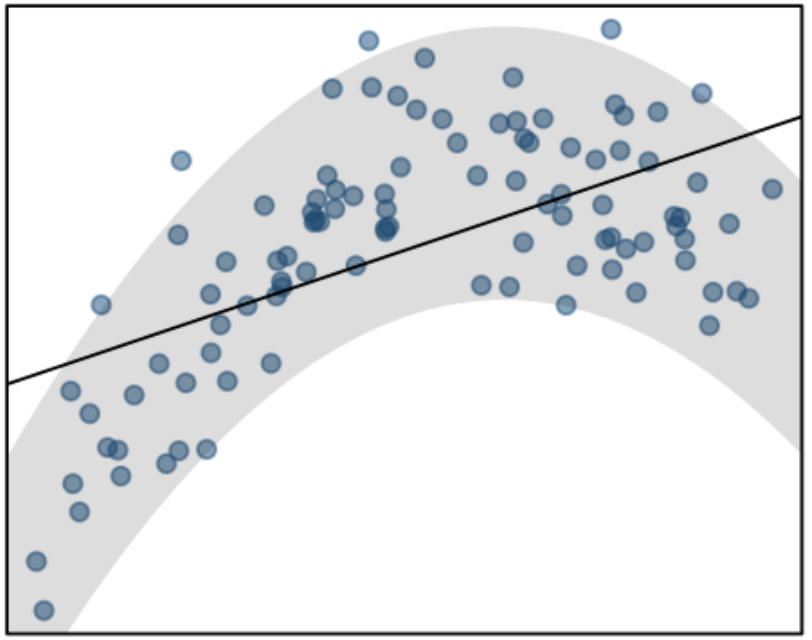
using a residual plot

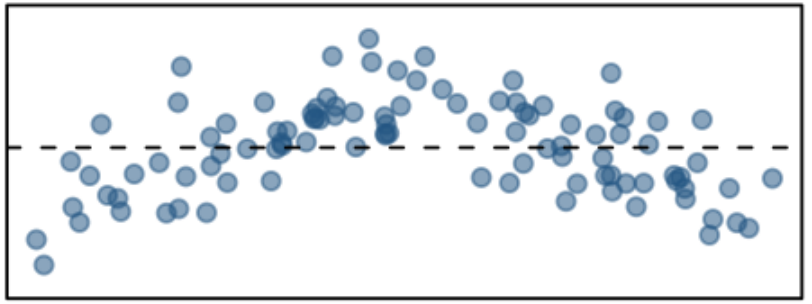


LS line: Condition: Check

What condition is this linear model obviously violating?

(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers

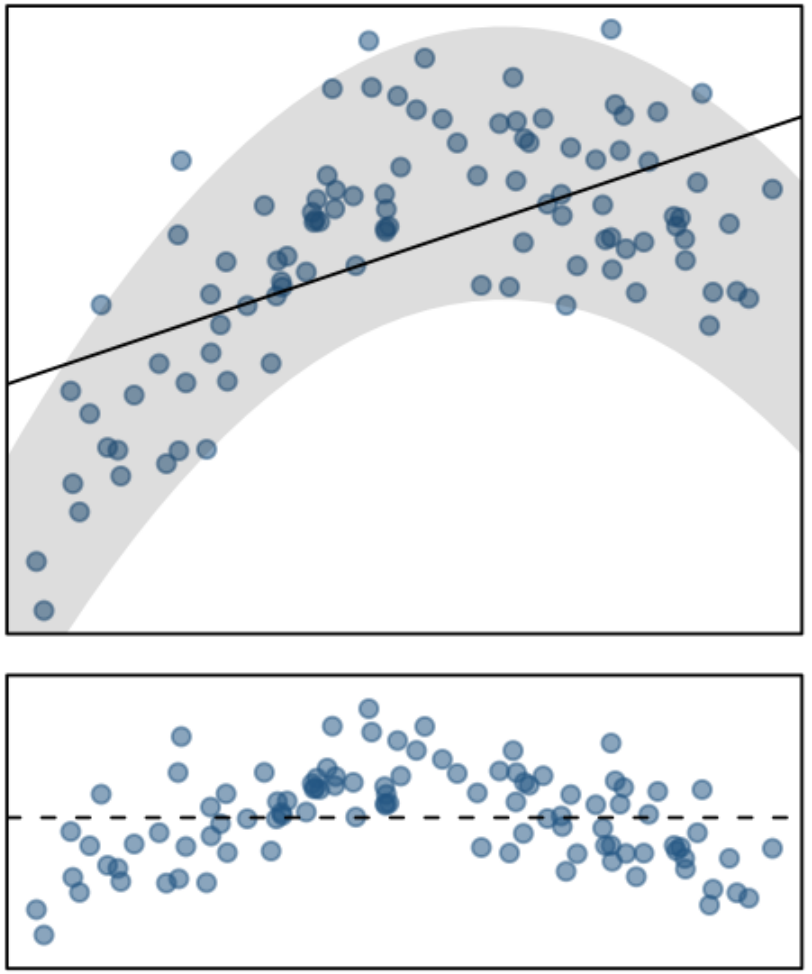




LS line: Condition: Check

What condition is this linear model obviously violating?

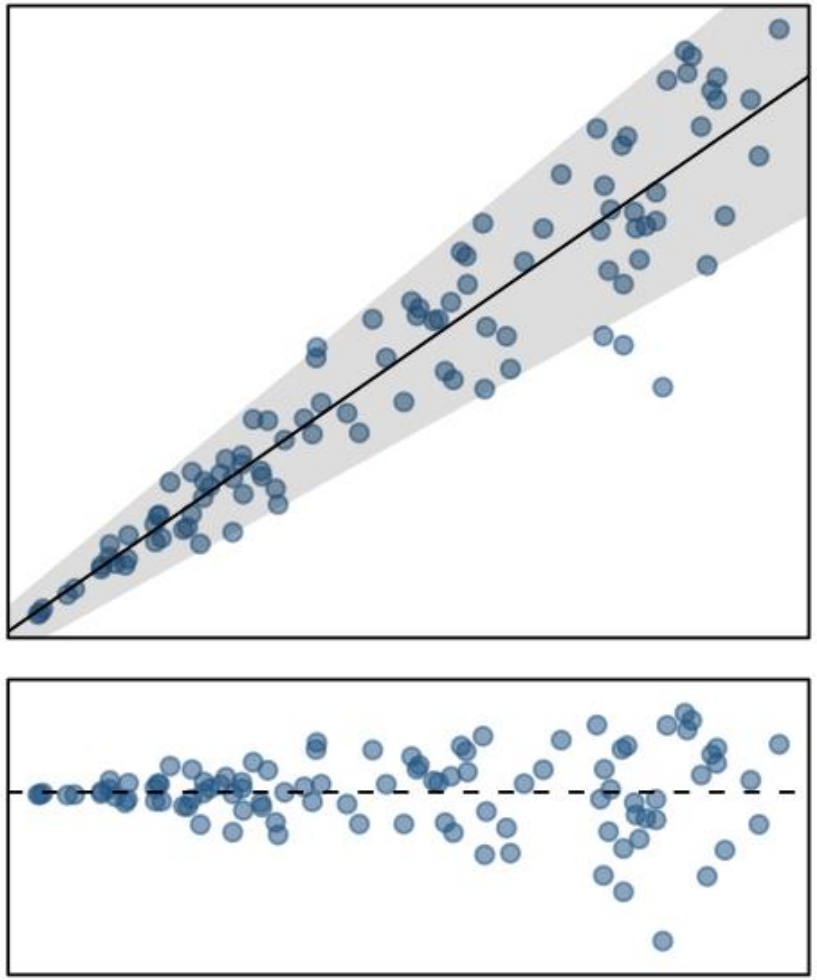
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



LS line: Condition: Check

What condition is this linear model obviously violating?

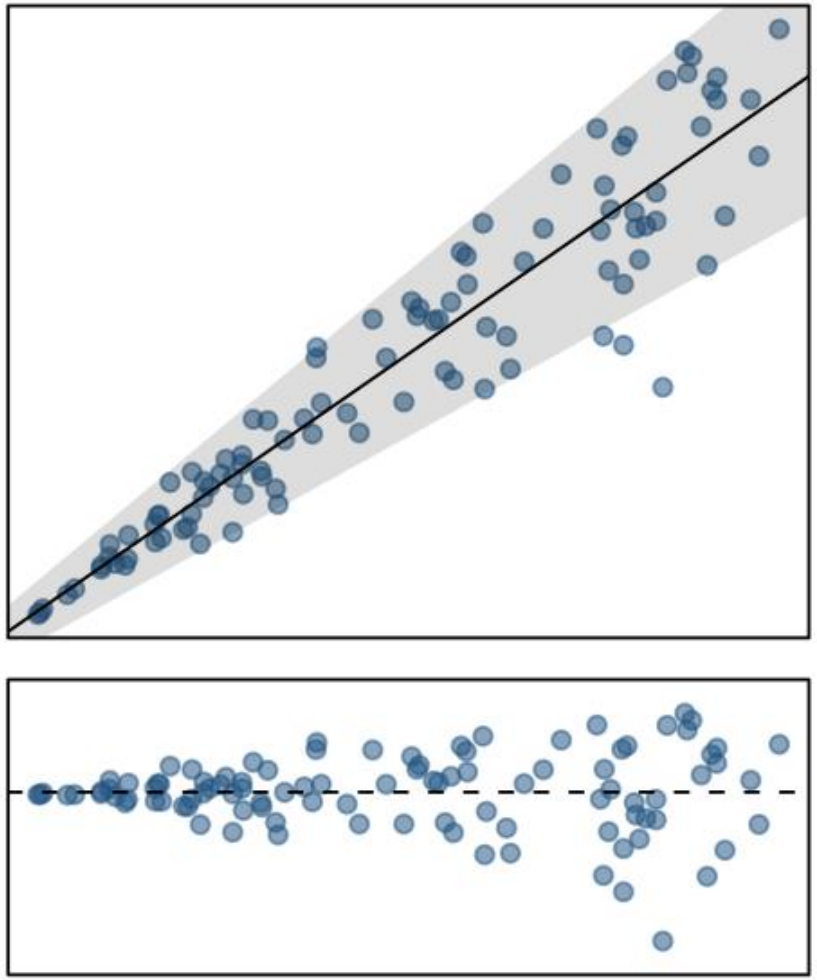
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers

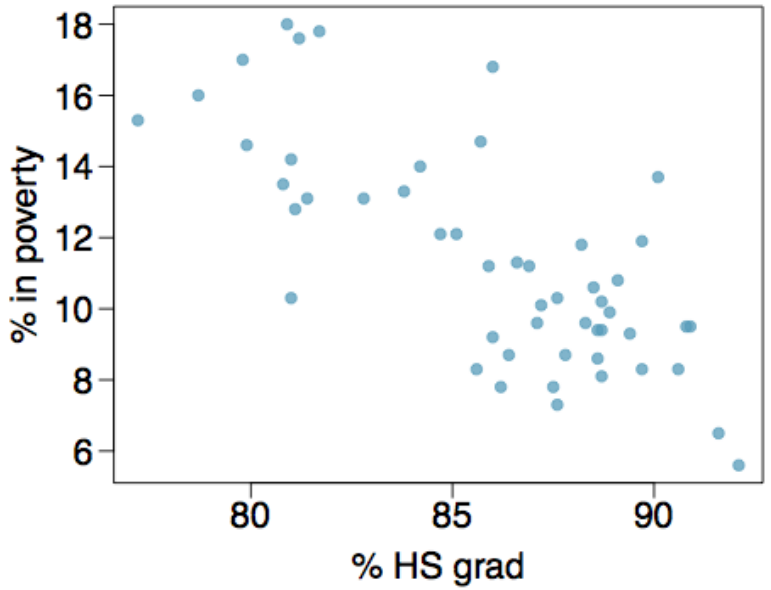


LS line: Condition: Check

What condition is this linear model obviously violating?

(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers





	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

slope

$$b_1 = \frac{s_y}{s_x} R$$

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

Interpretation:
For each additional % point in HS graduate rate, we would expect the % living in poverty to be **lower on average by 0.62**% points

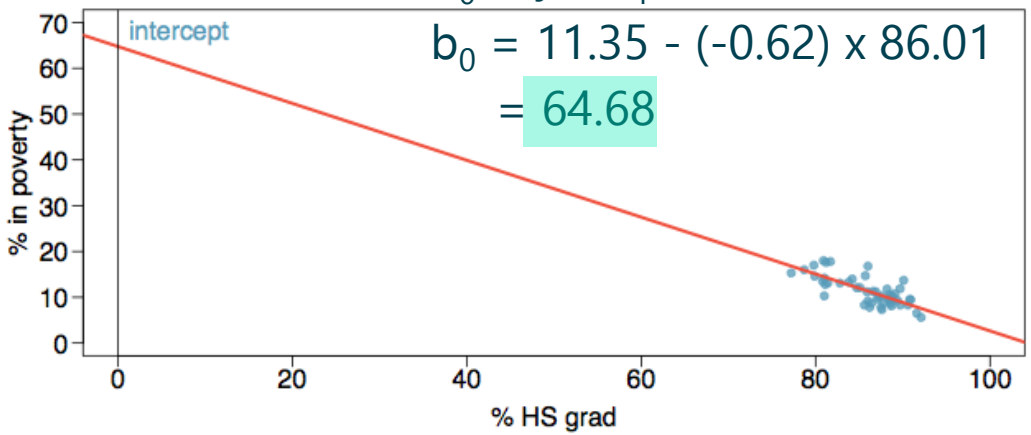
intercept

regression line always passes through $(\bar{x}, \bar{y})$

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = 11.35 - (-0.62) \times 86.01$

$= 64.68$

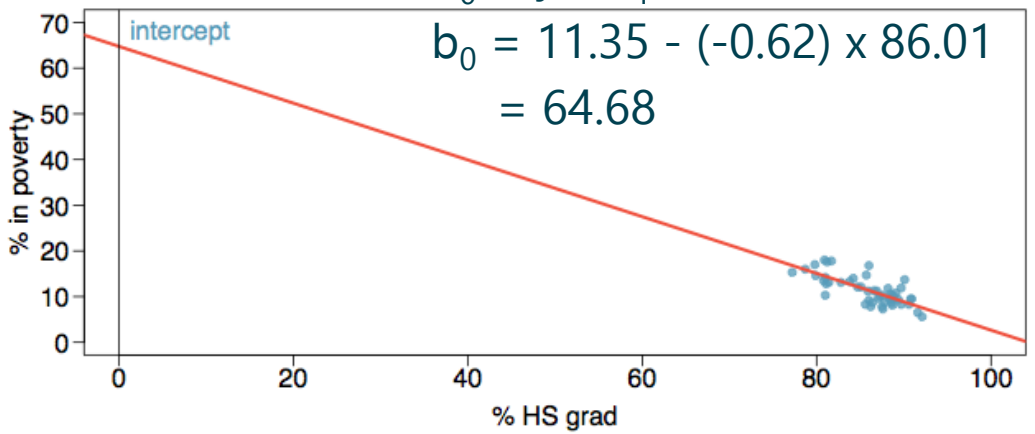


Intercept?

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = 11.35 - (-0.62) \times 86.01$

$= 64.68$



Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line

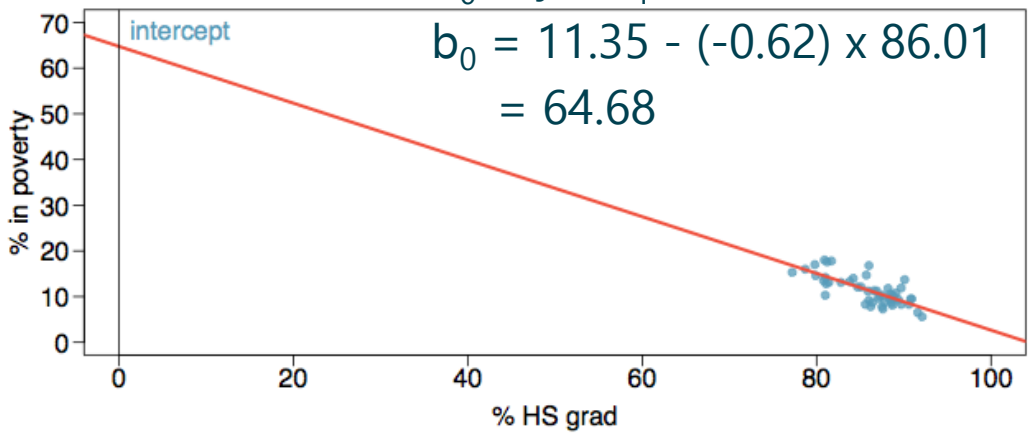
(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

Intercept?

$b_0 = \bar{y} - b_1 \bar{x}$

$b_0 = 11.35 - (-0.62) \times 86.01$

$= 64.68$



Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

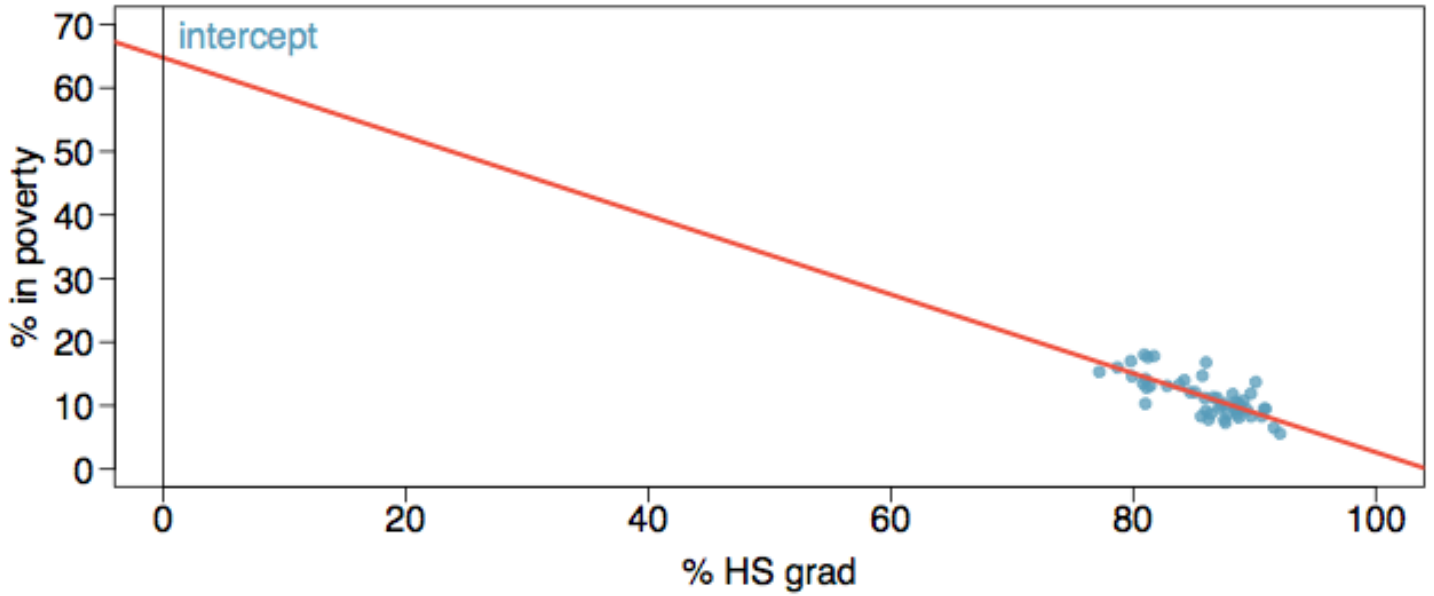
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

Intercept: aside

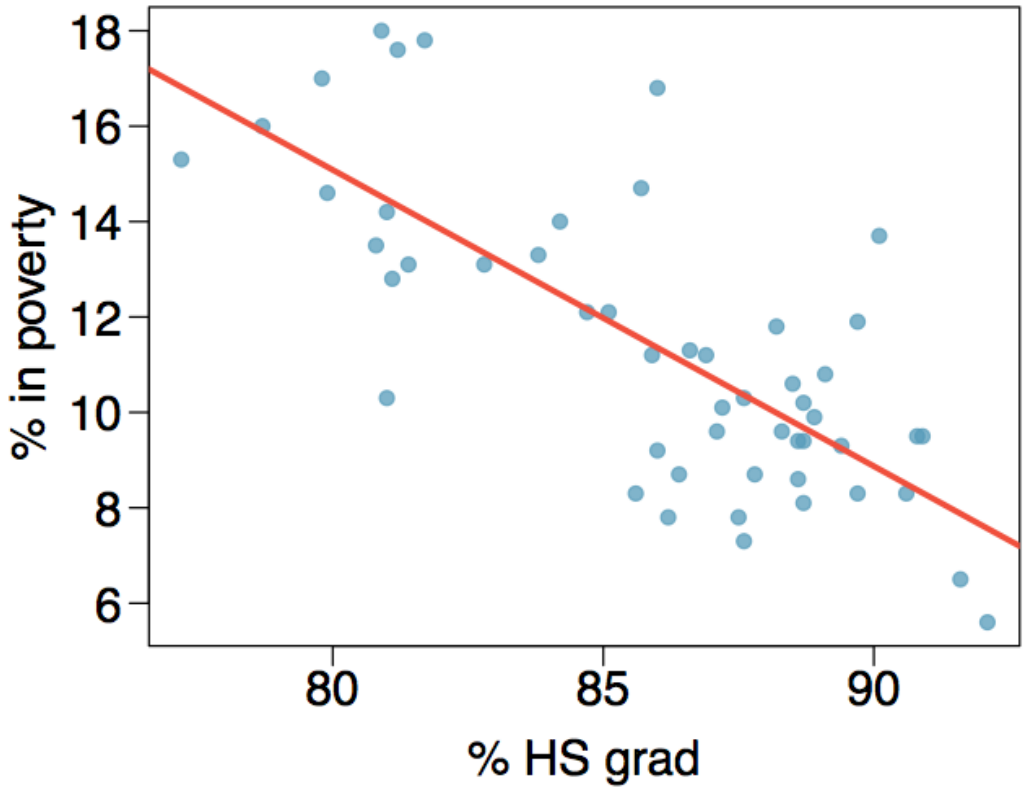
there are no states in the dataset with no HS graduates => intercept is not very useful & not reliable



especially because the predicted value of the intercept is so far from the bulk of the data

Regression line

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$

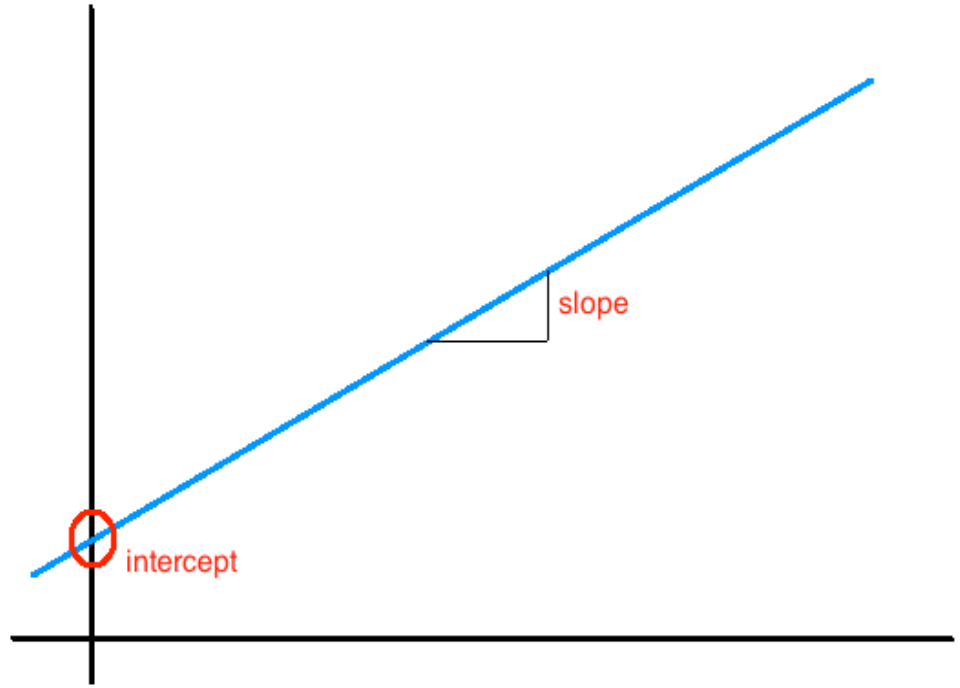


interpretation

intercept
when $x = 0$,
y is expected to
equal the intercept

slope
for each unit in x,
y is expected to
increase / decrease on average by the slope

NB:
These statements are not **causal**, unless the study is a randomized controlled experiment.



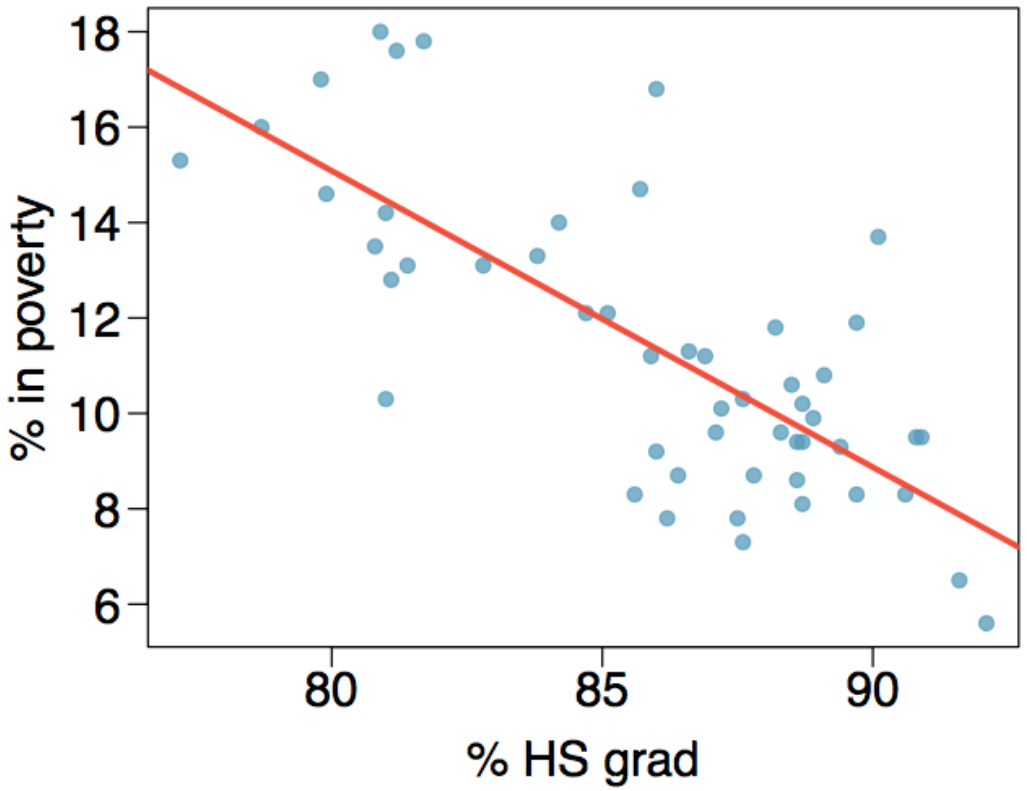
Prediction

The linear model is used to predict
the value of the response variable
for a given value of the explanatory variable

i.e. plug in the value of x
into the linear model equation

NB:
there will be some uncertainty associated
with the predicted value

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



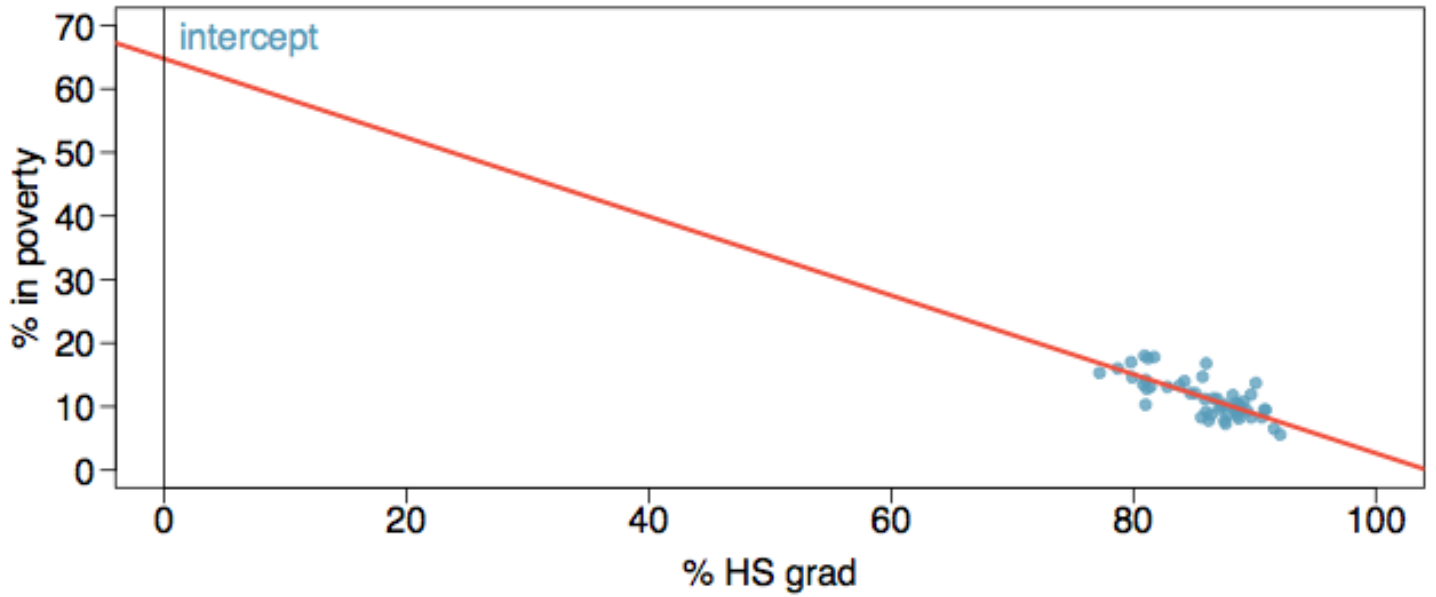
Extrapolation

applying a model estimate to values outside of the realm of the original data

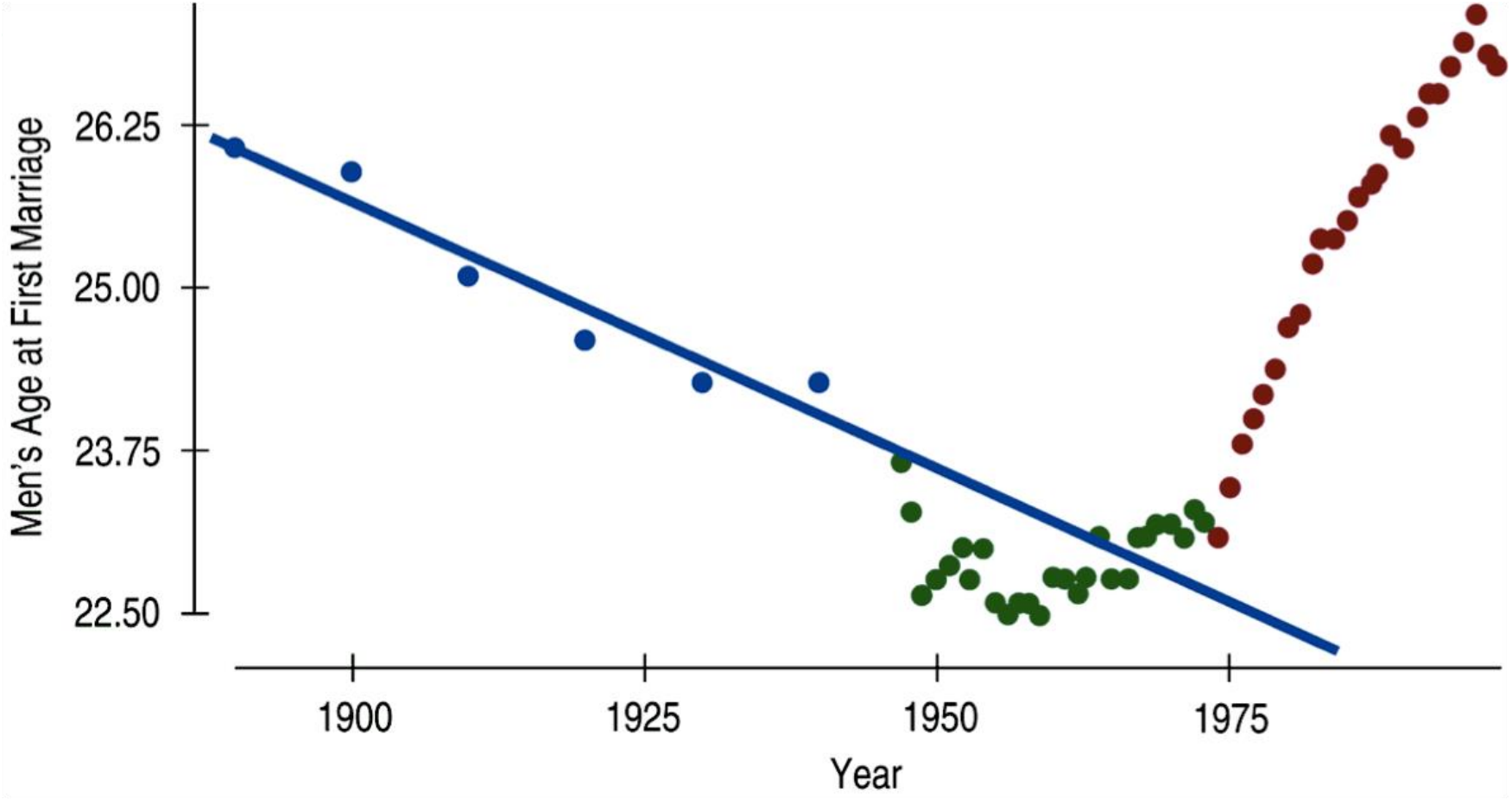
Extrapolation: e.g.

applying a model estimate to values outside of the realm of the original data

The intercept might be an extrapolation:



Extrapolation: e.g.



BBC NEWS

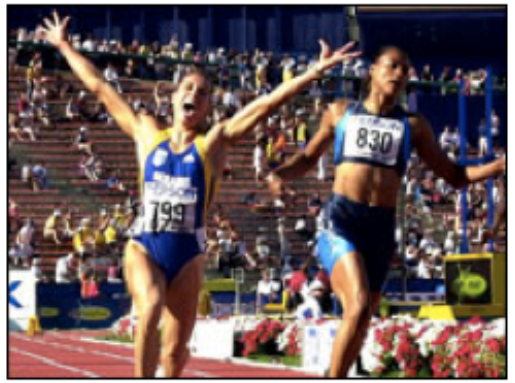
Watch One-Minute World News

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.



Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

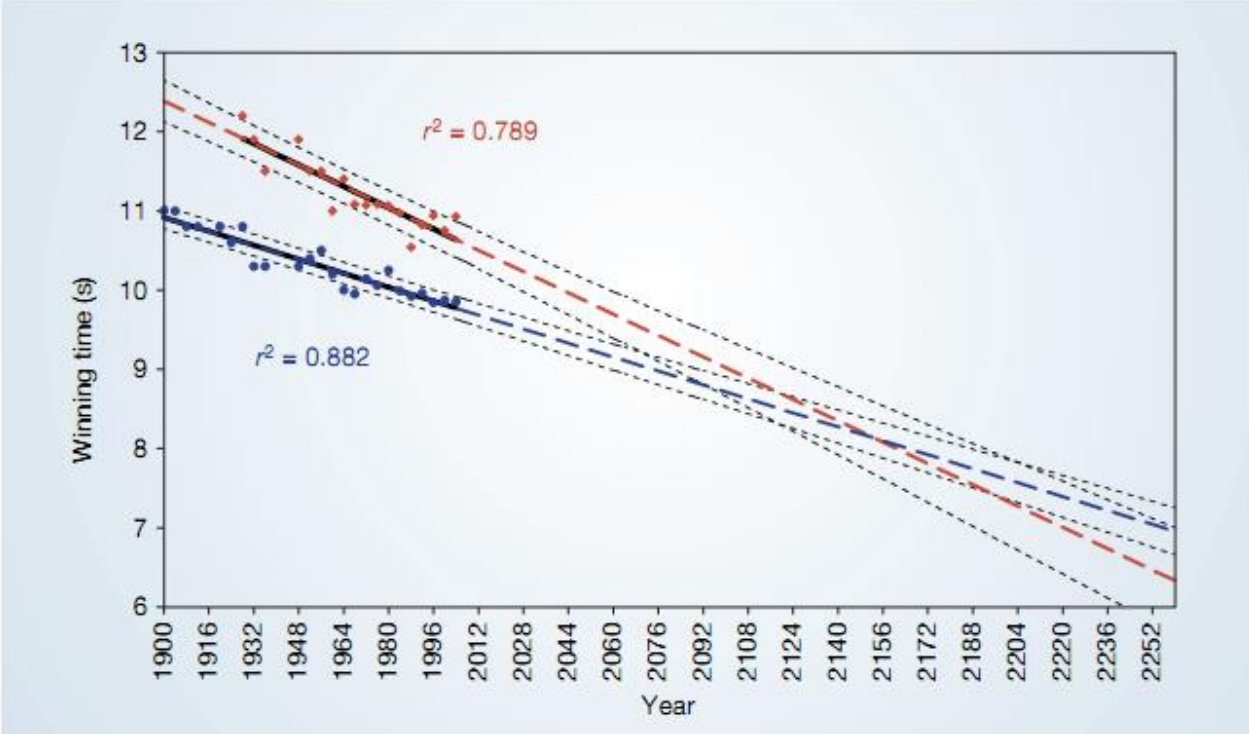
The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.



The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The **regression lines are extrapolated** (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The **projections intersect just before the 2156 Olympics**, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s

Regression Performance: R^2

What is R^2 ?

R^2 evaluates the strength of the fit of a linear model

R^2 = square of the correlation coefficient

R^2 = "coefficient of determination"

What does R^2 tell us?

what percent of variability in the response variable is **explained** by the model

What about the remainder of the variability?

remainder is explained by variables NOT included in the model

or by inherent randomness in the data

for the model (% in poverty vs % HS grad), $R^2 = (- 0.62)^2 = 0.38$

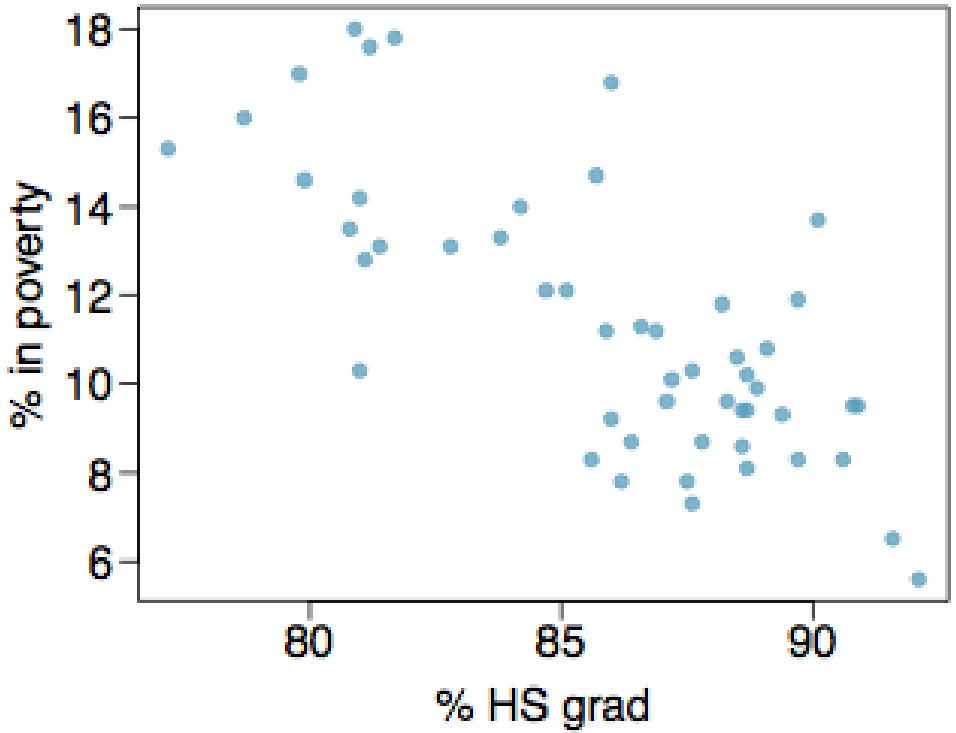
$R^2 = 0.38$ interpretation?

a) 38% of the variability in the % of HS graduates among the 51 states is explained by the model.

b) 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

c) 38% of the time % HS graduates predict % living in poverty correctly.

d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model



$R^2 = 0.38$ interpretation?

a) 38% of the variability in the % of HS graduates among the 51 states is explained by the model.

b) 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

c) 38% of the time % HS graduates predict % living in poverty correctly.

d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model