Linear Regression

Line Fitting

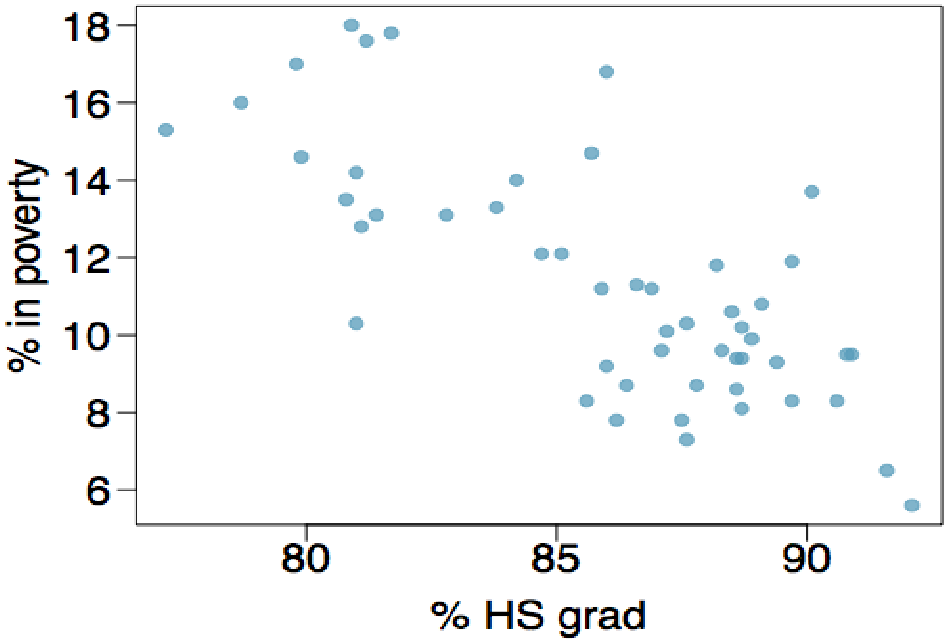
MODELING NUMERICAL VARIABLES

Modeling numerical variables

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

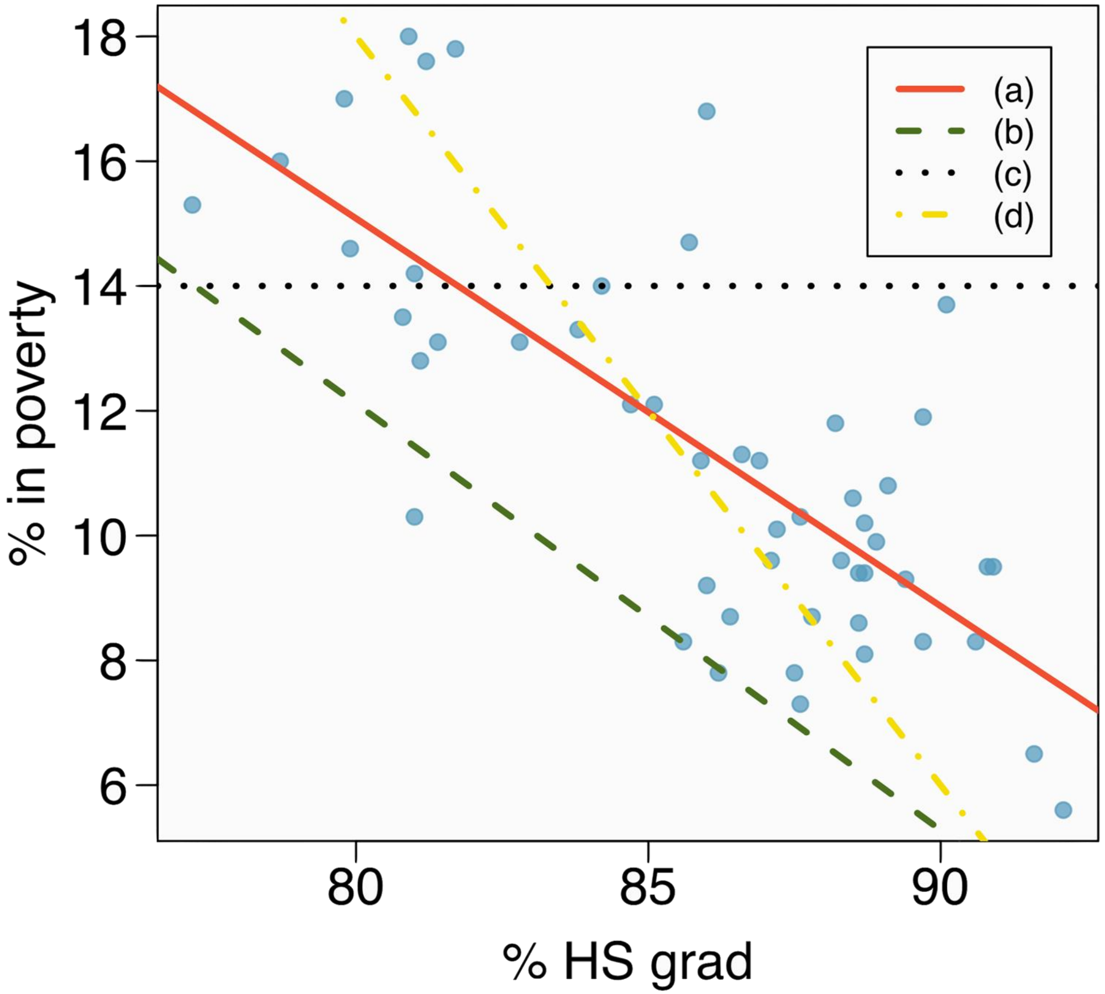
SCATTERPLOT

Poverty vs. HS graduate rate



EYEBALLING THE LINE

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

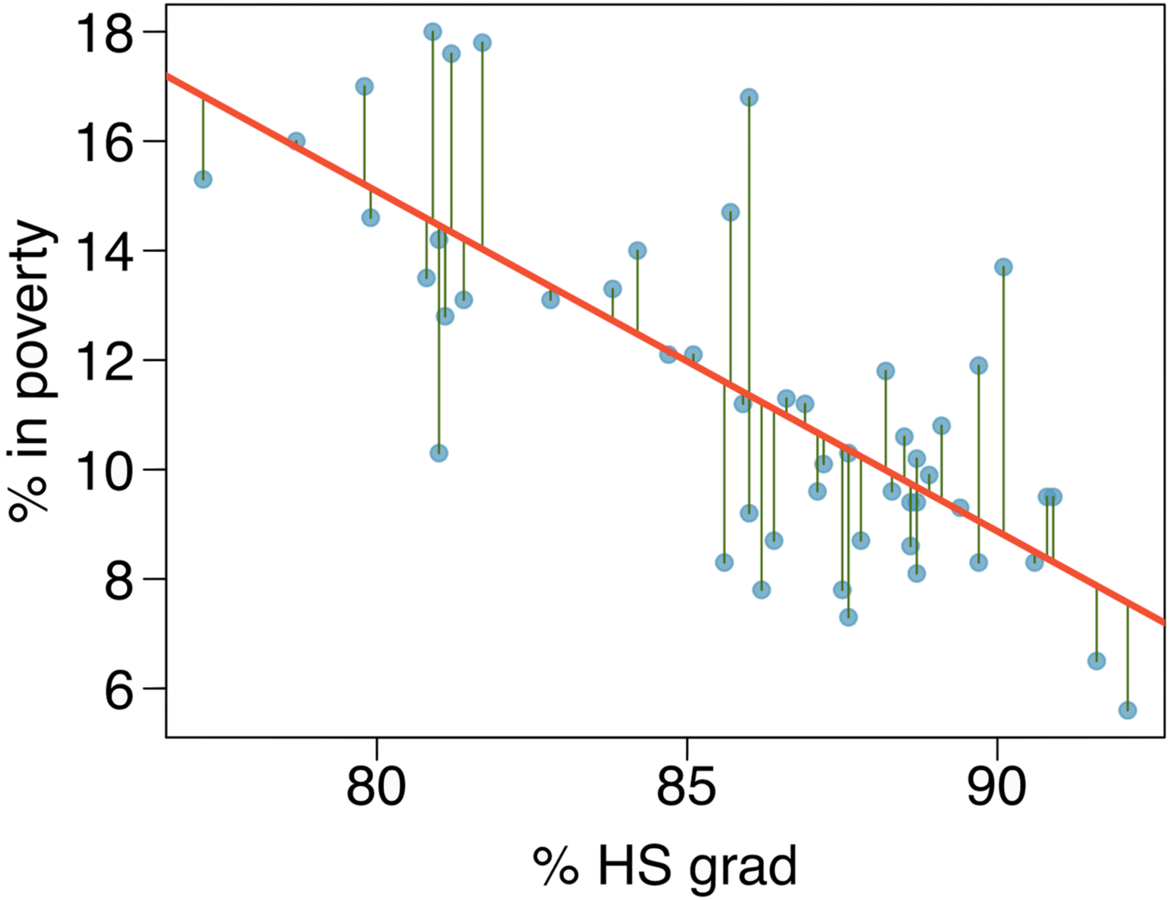


Residuals

RESIDUALS

Leftovers from the model fit:

Data = Fit + Residual

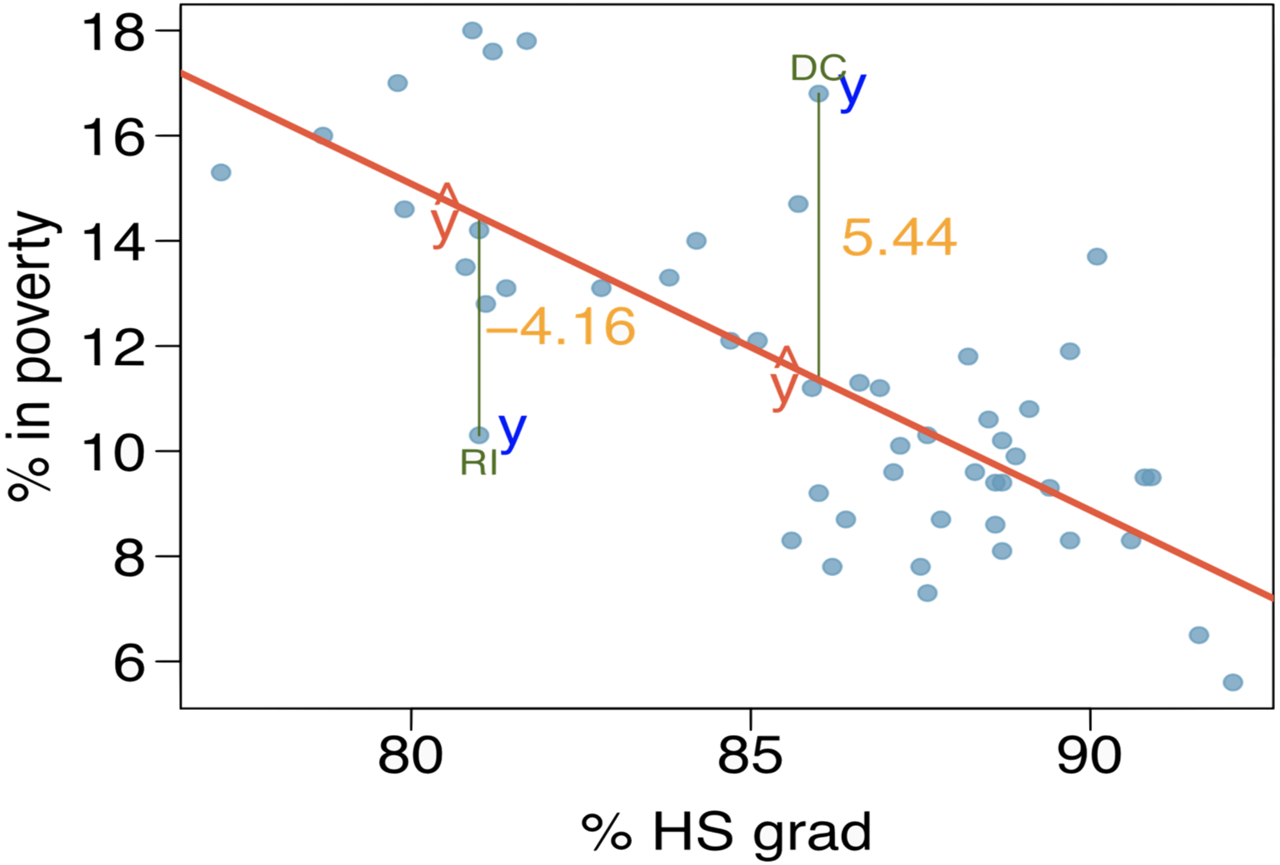


RESIDUALS

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

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

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



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.

COVARIANCE

Quantifying the relationship

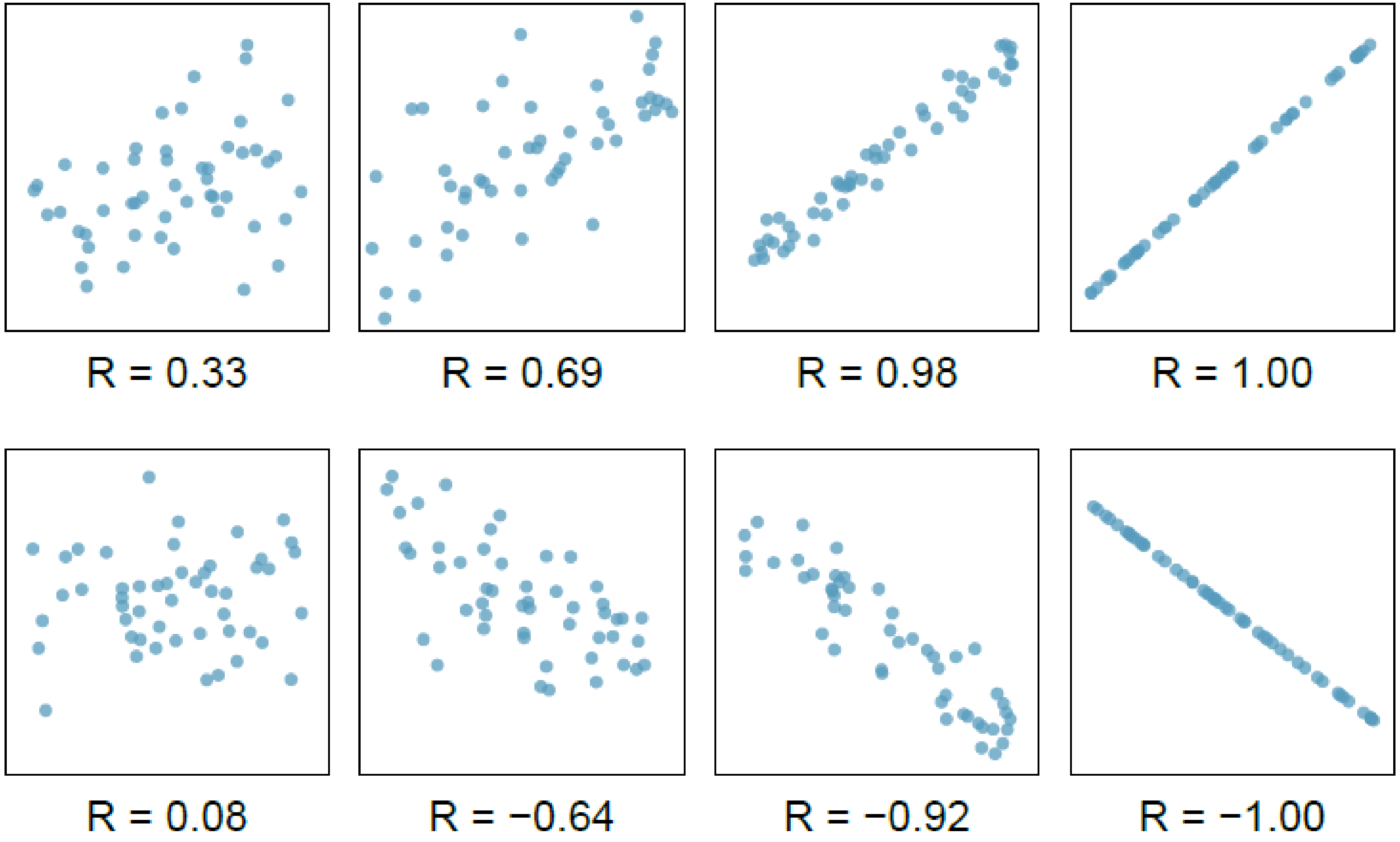
$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

CORRELATION

Quantifying the relationship

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

QUANTIFYING THE RELATIONSHIP



R = 0.33

R = 0.69

R = 0.98

R = 1.00

R = 0.08

R = −0.64

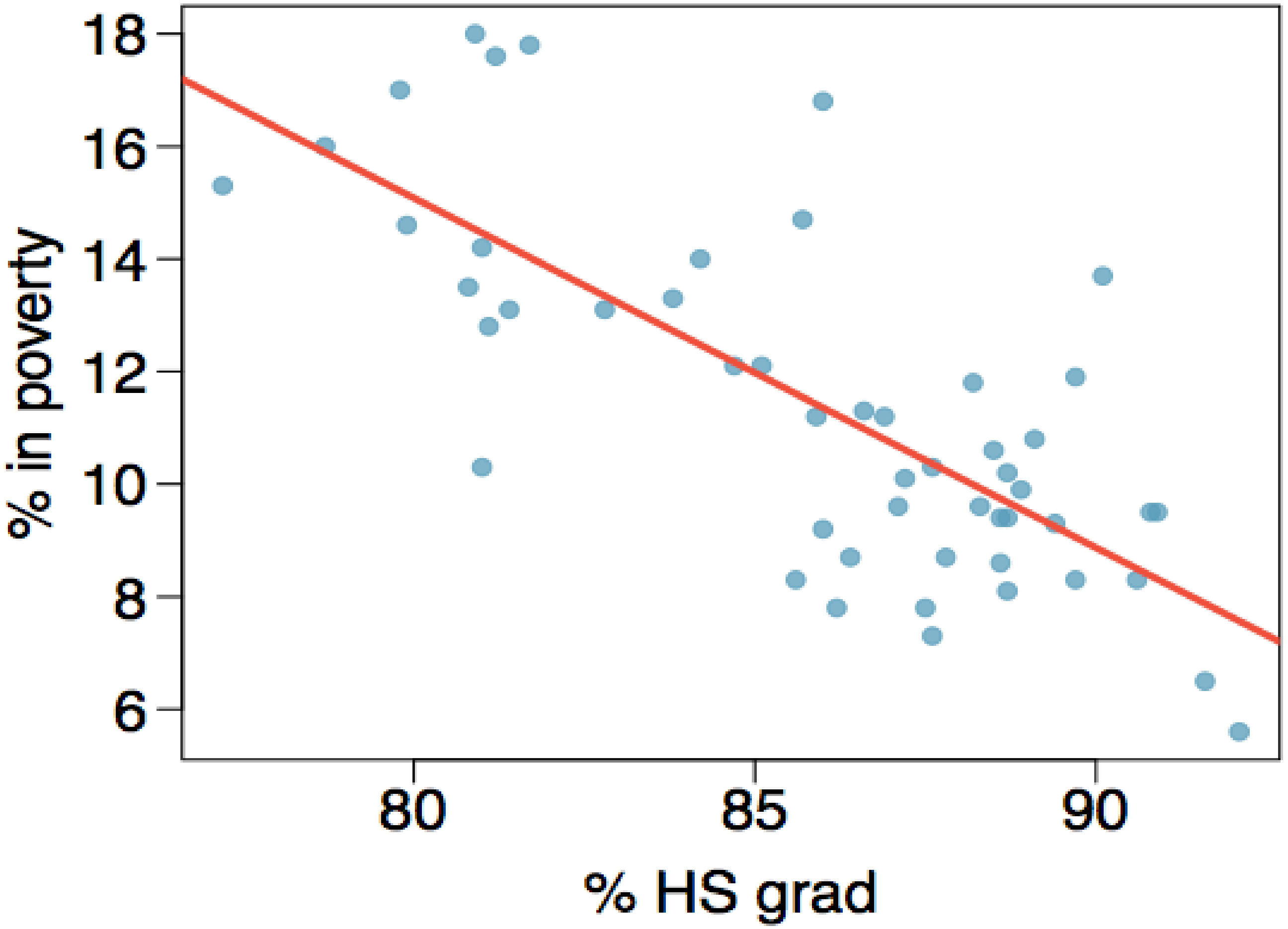
R = −0.92

R = −1.00

GUESSING THE CORRELATION

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

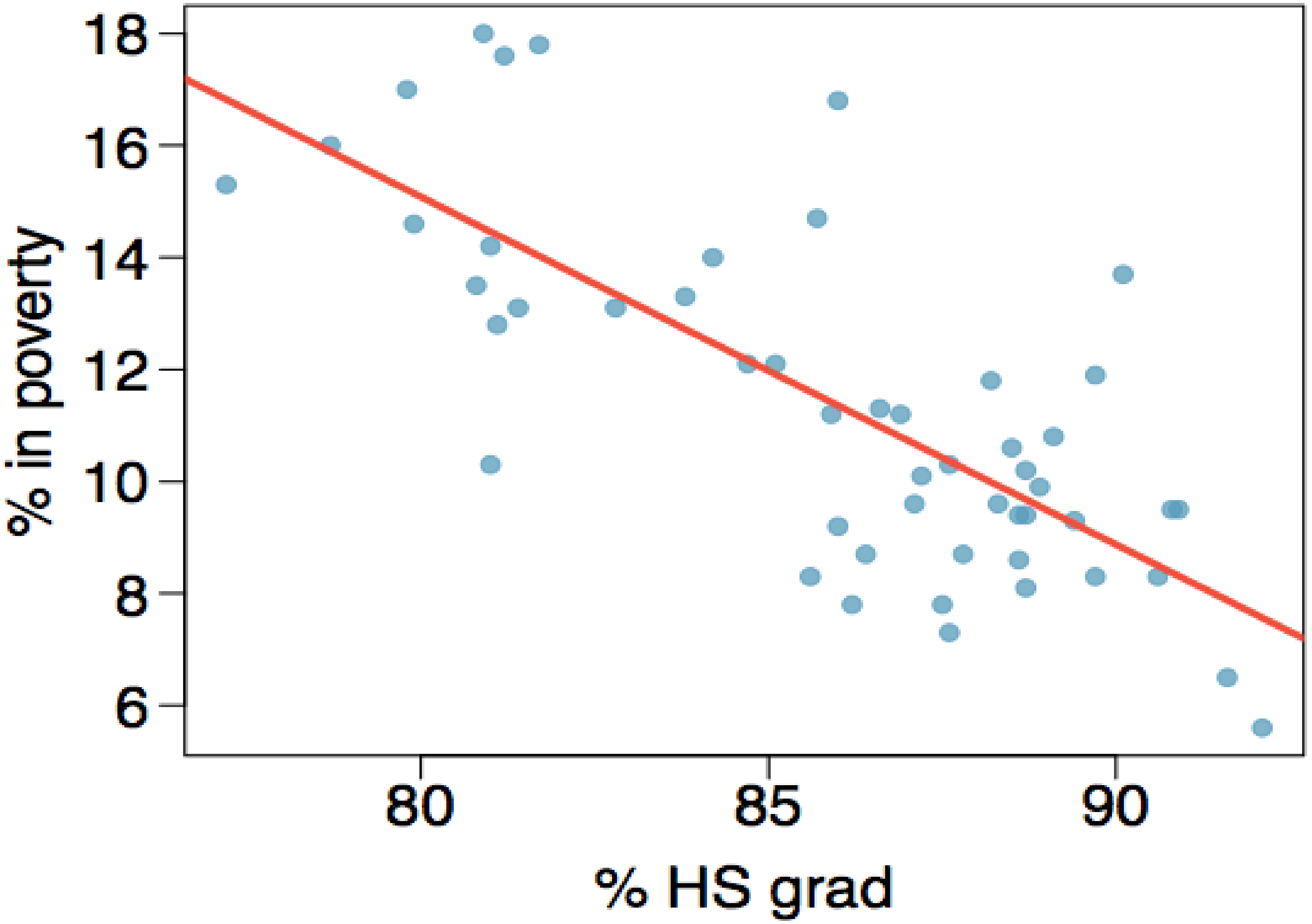
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



GUESSING THE CORRELATION

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

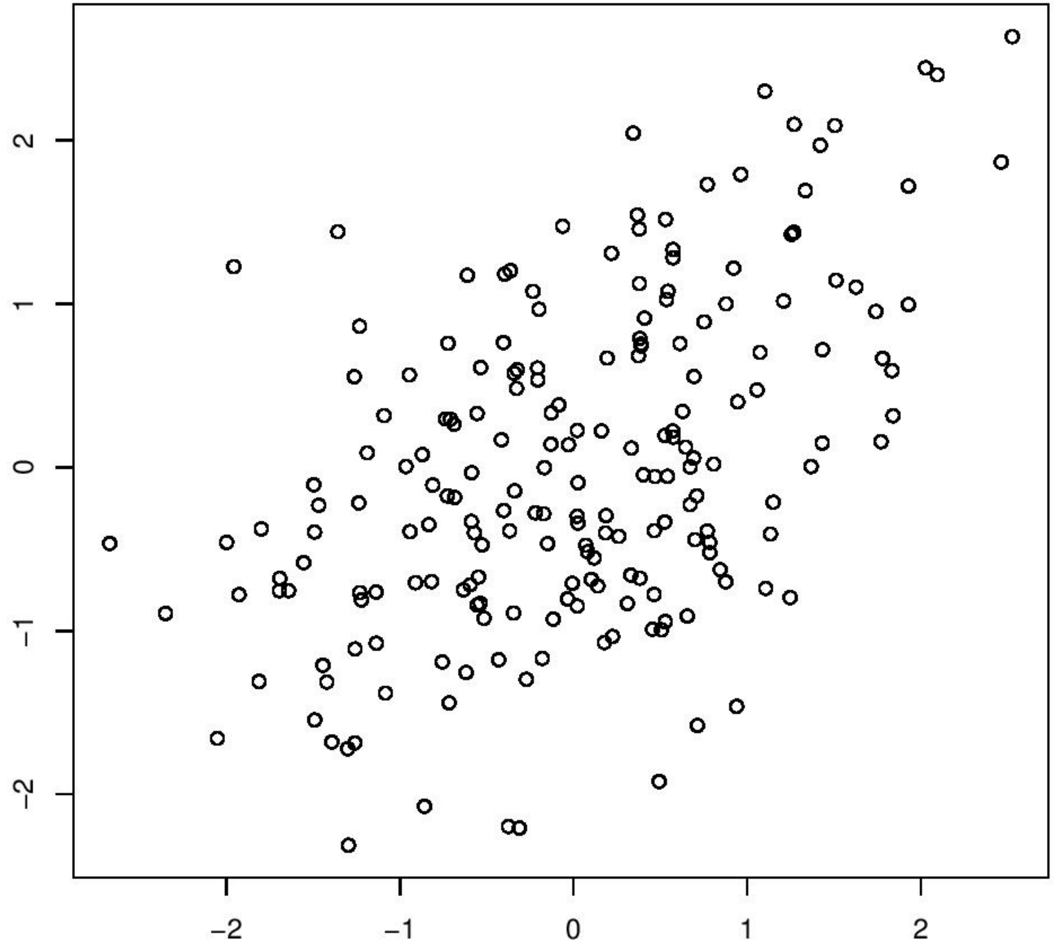
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



GUESSING THE CORRELATION

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

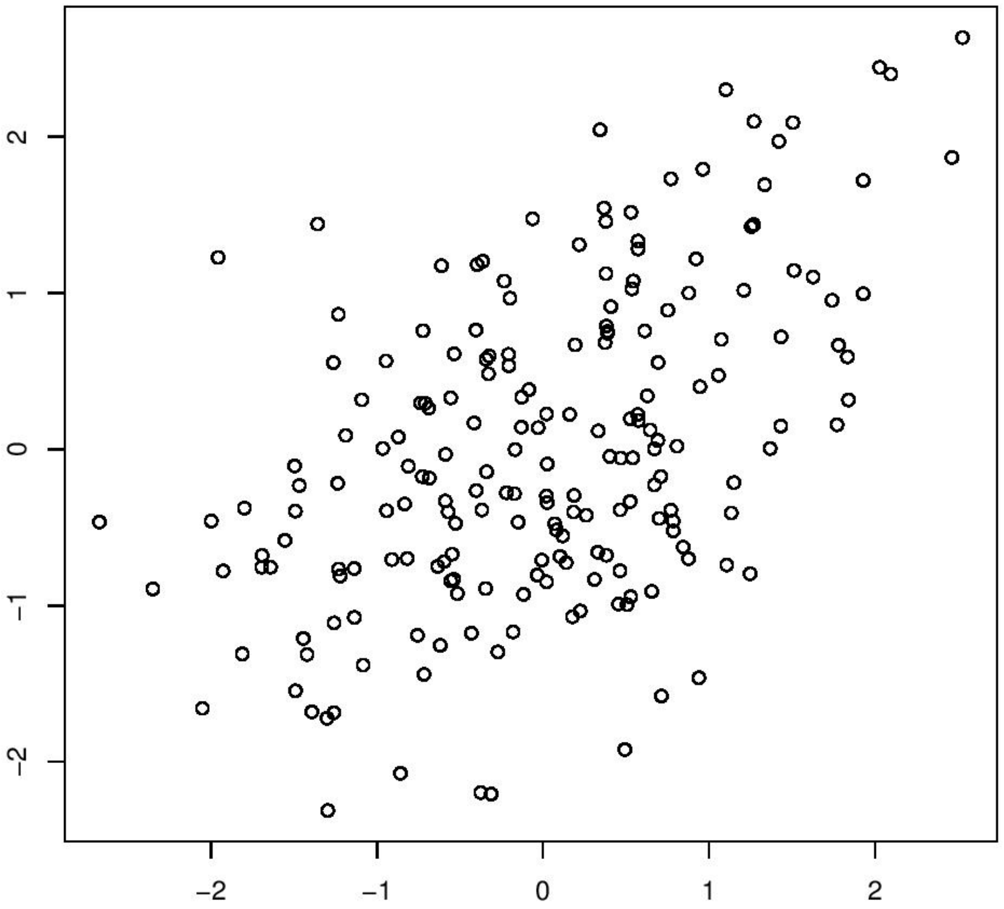
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



GUESSING THE CORRELATION

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

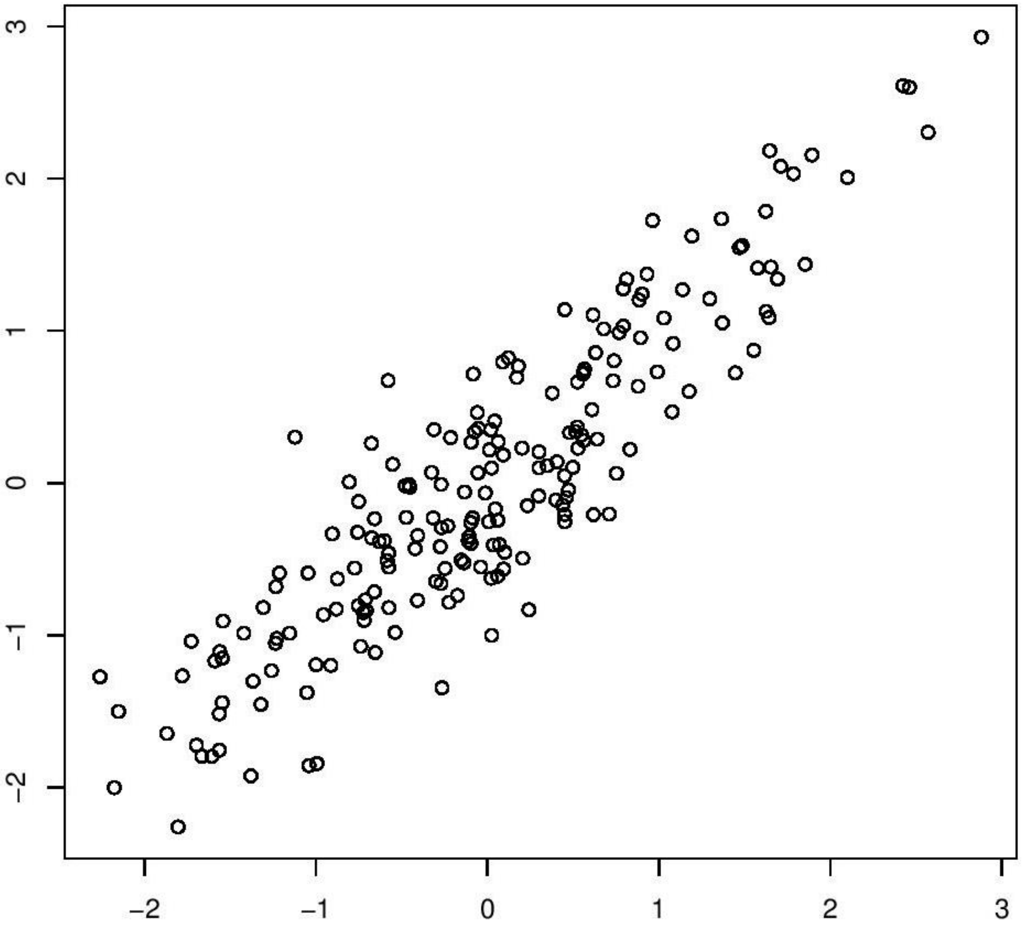
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



GUESSING THE CORRELATION

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

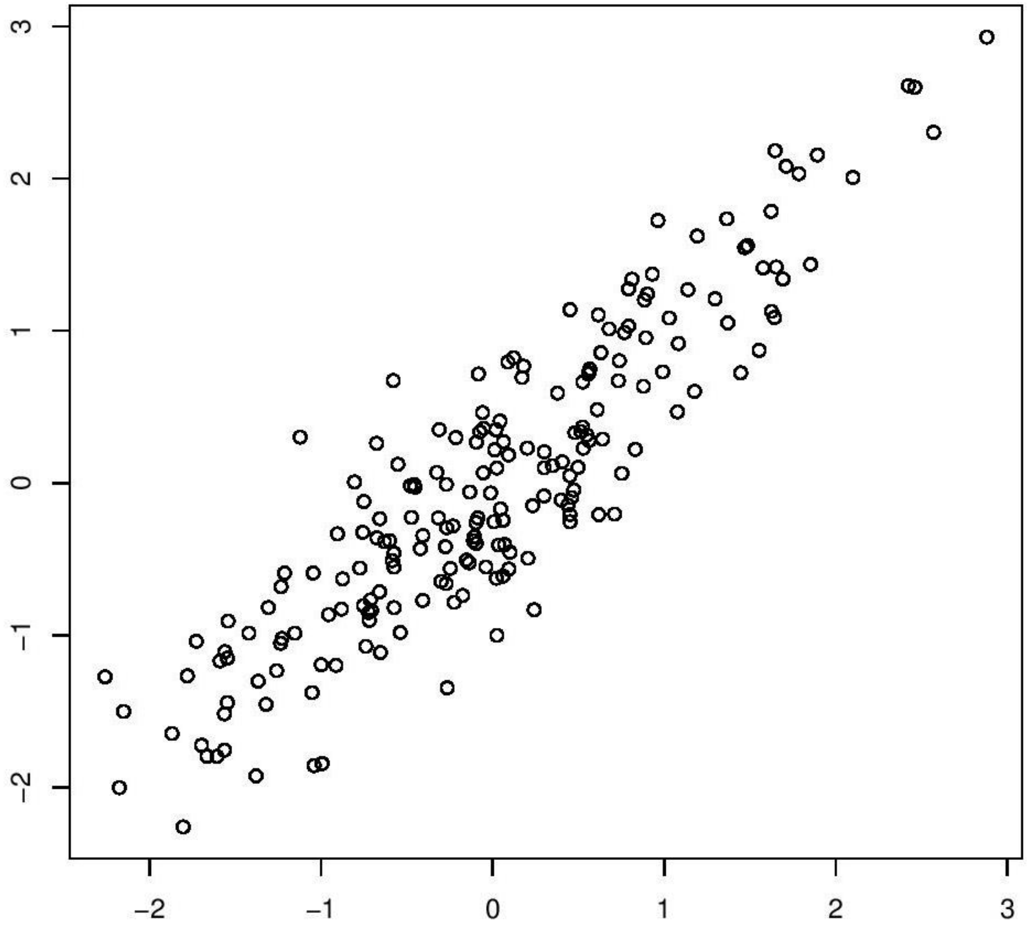
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



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?

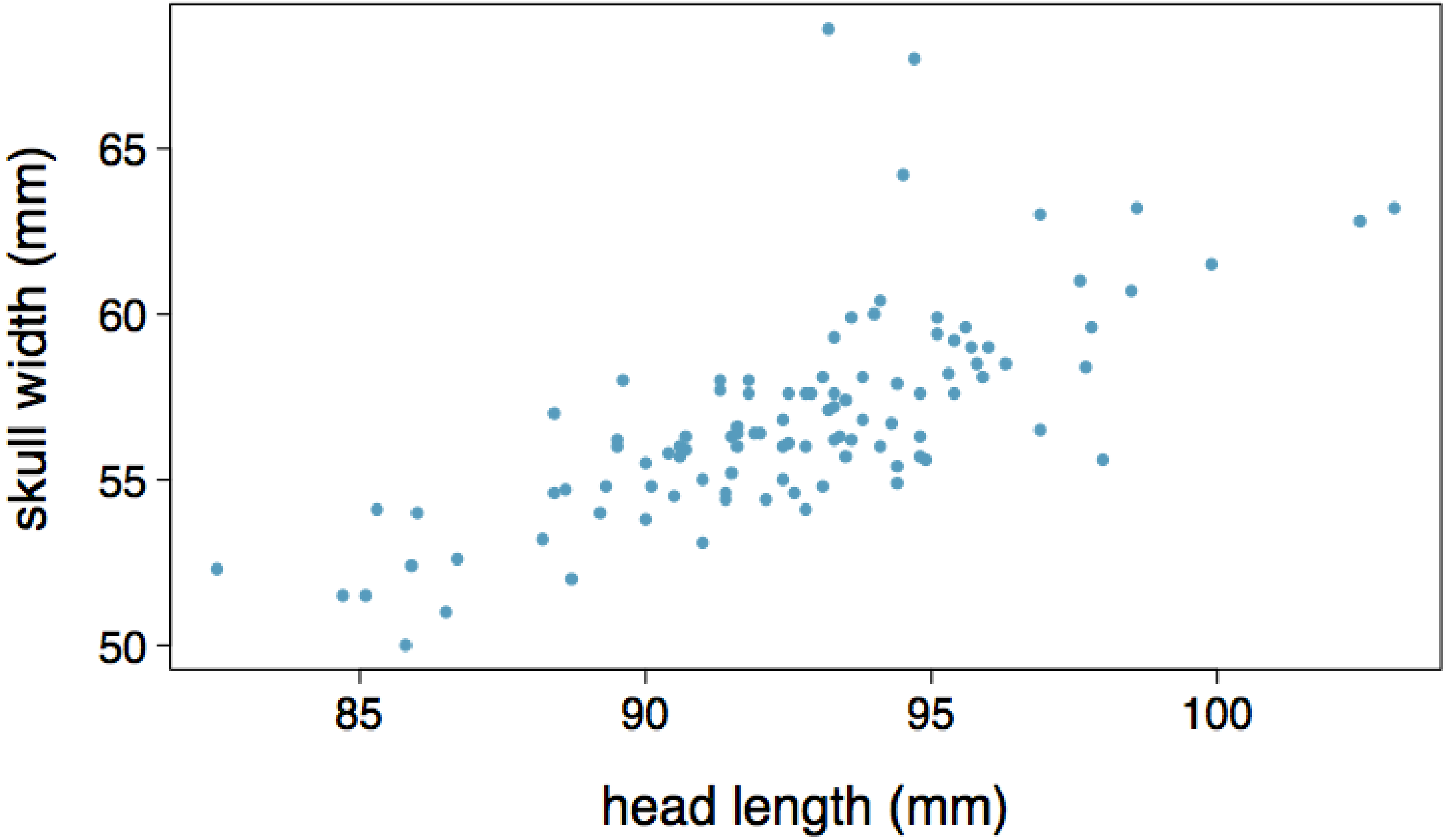
(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(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?

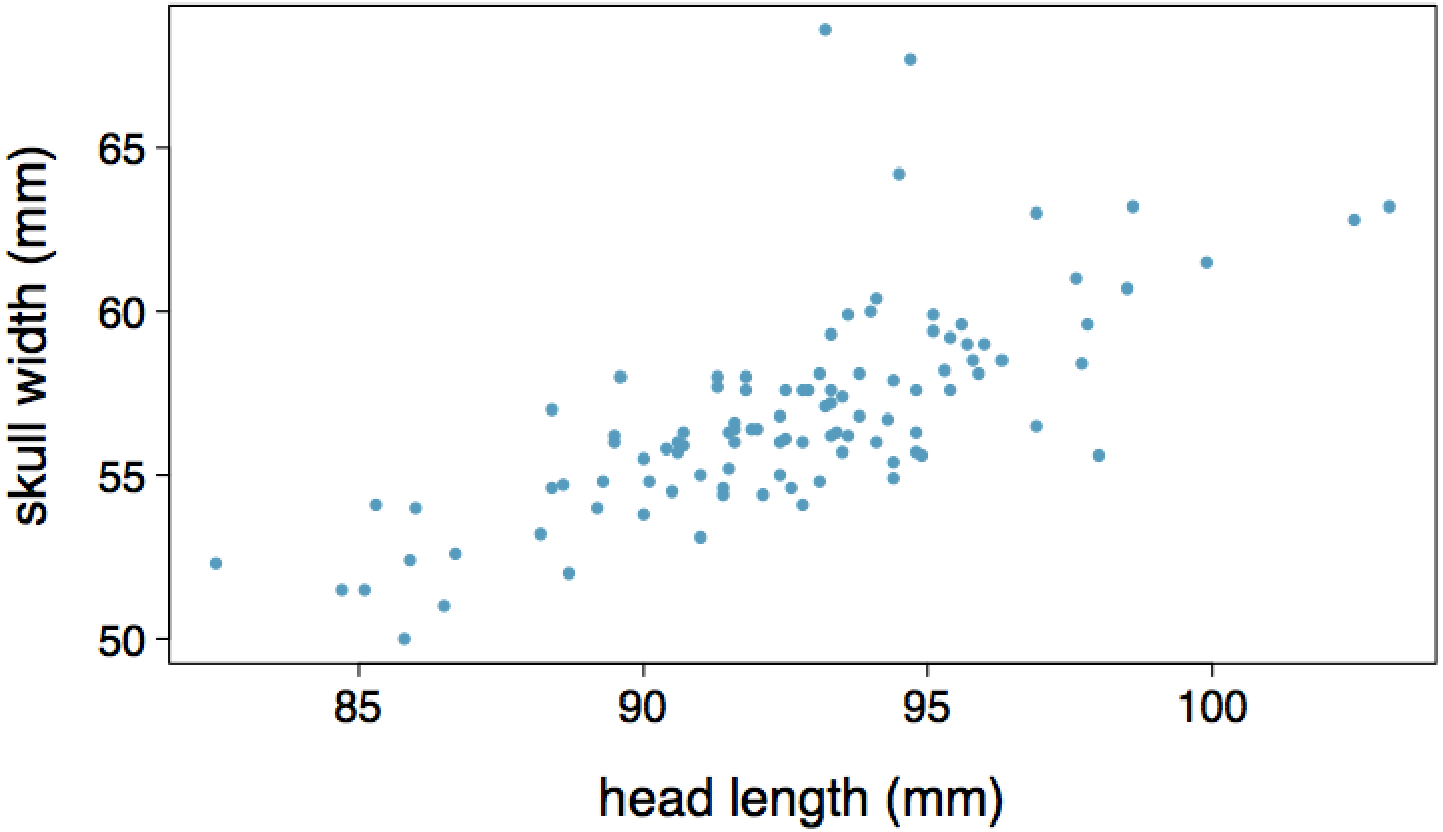
(a) There is no relationship between head length and skull width, i.e. the variables are independent.

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

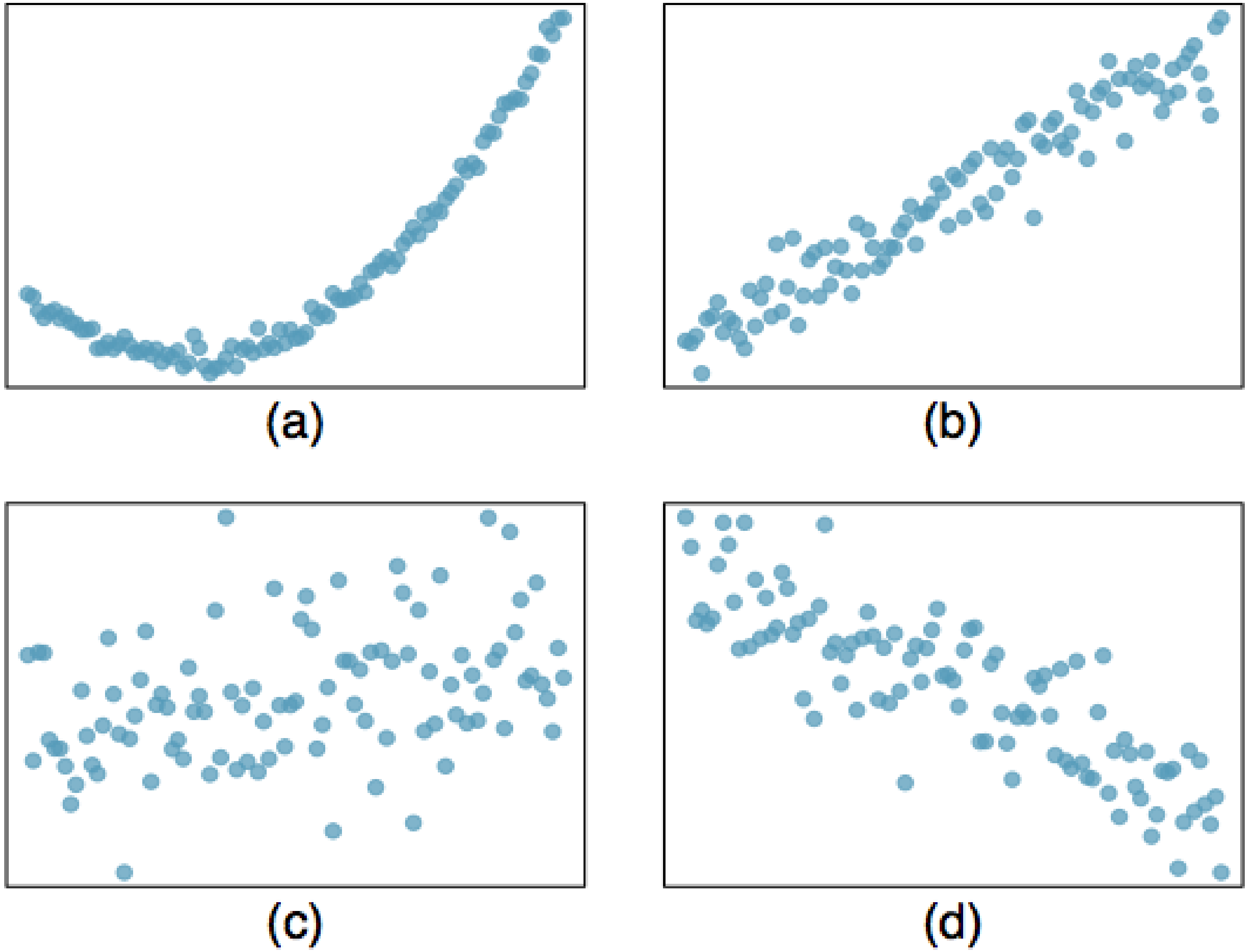
(c) Skull width and head length are negatively associated.

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

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



ASSESSING THE CORRELATION



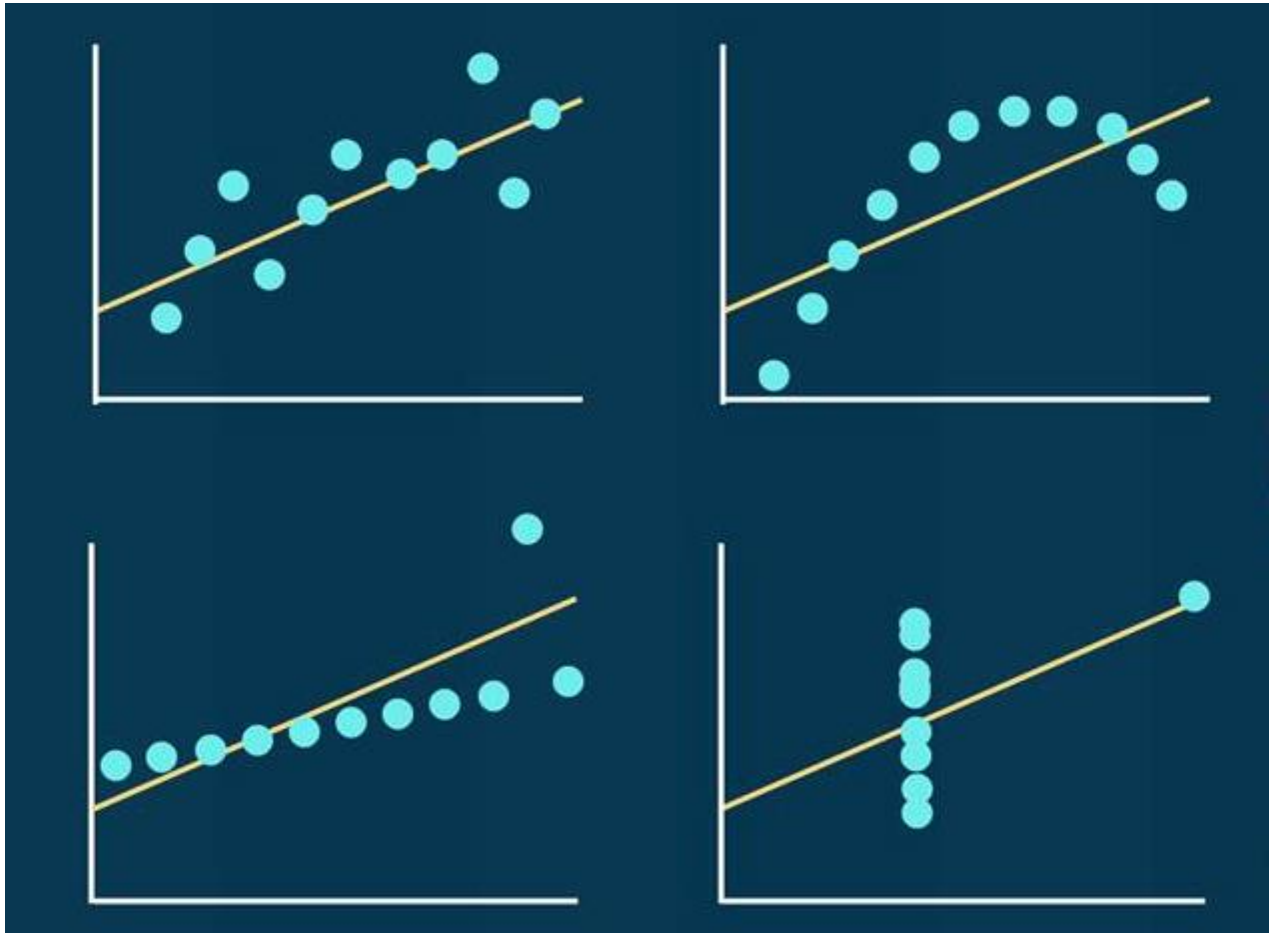
(a)

(b)

(c)

(d)

BE CAREFUL WHEN MAKING CONCLUSIONS BASED ON STATISTICS



ref: The Anscombe's Quartet, 1973

Fitting a line by least squares regression

FITTING

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

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

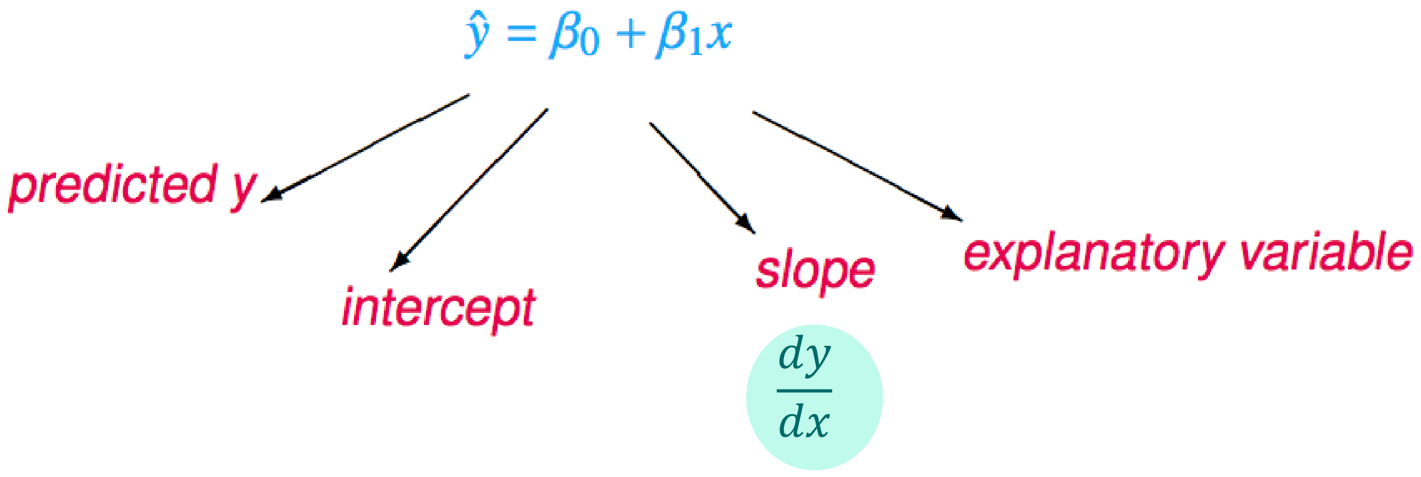
  $$|e_1| + |e_2| + \dots + |e_n|$$

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

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

- 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

EQUATION



$$\hat{y} = \beta_0 + \beta_1 x$$

predicted y

intercept

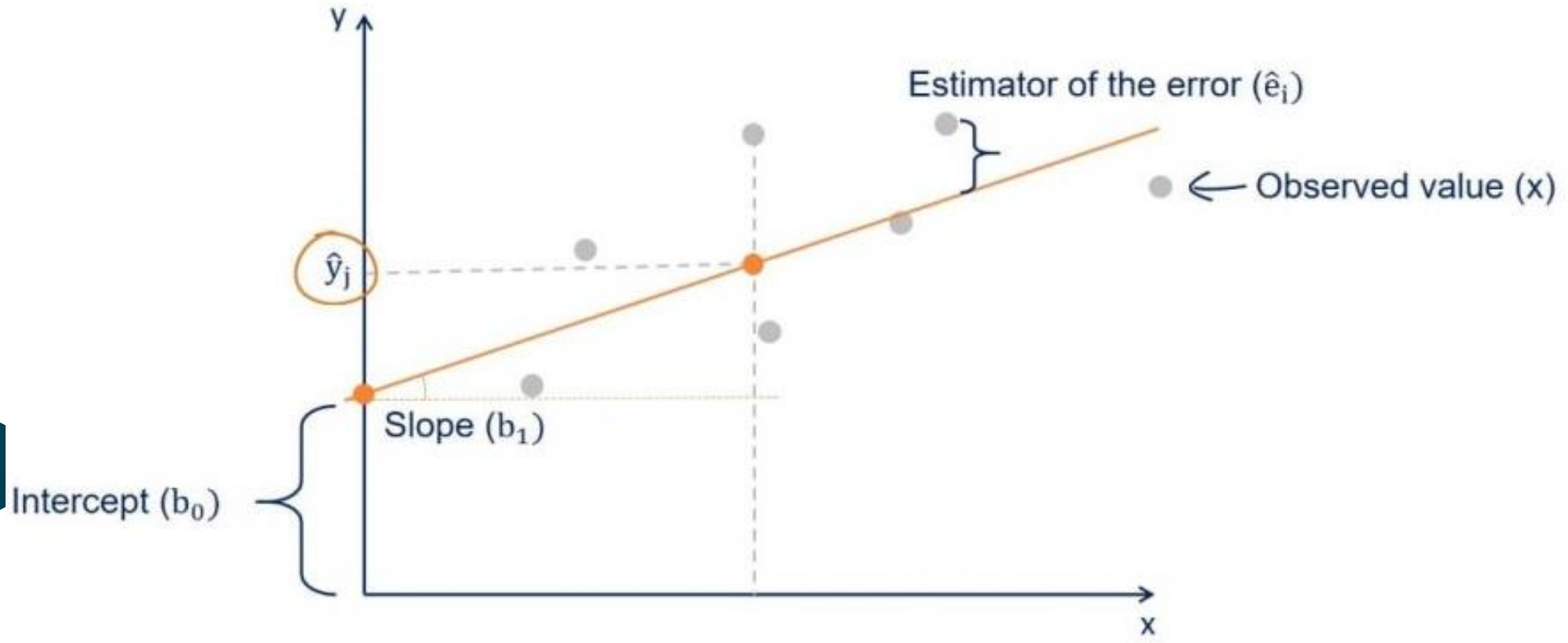
slope

$$\frac{dy}{dx}$$

explanatory variable

REGRESSION COEFFICIENTS

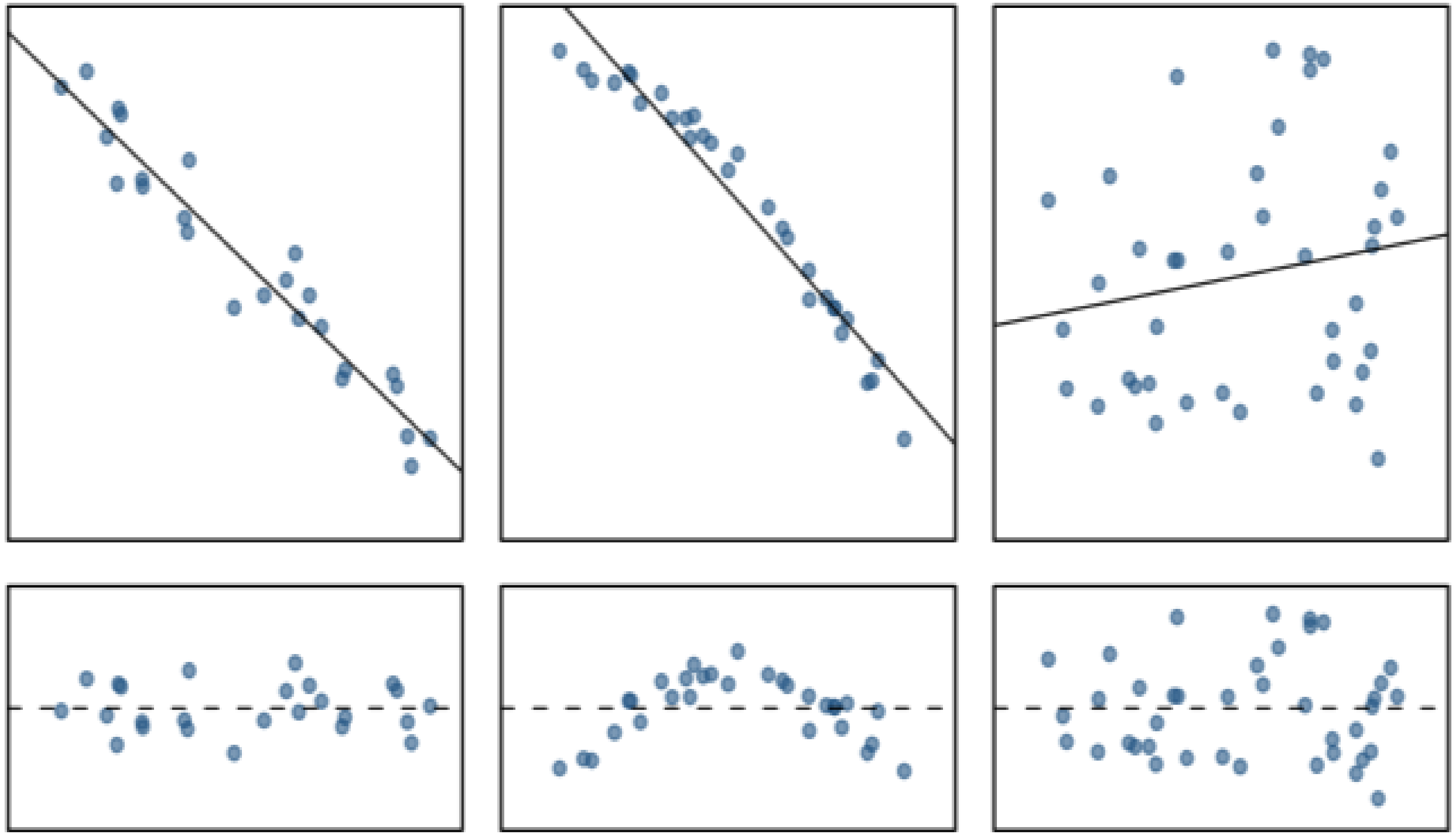
$$Y_i = \beta_0 + \beta_1 X_i + \hat{e}_i$$



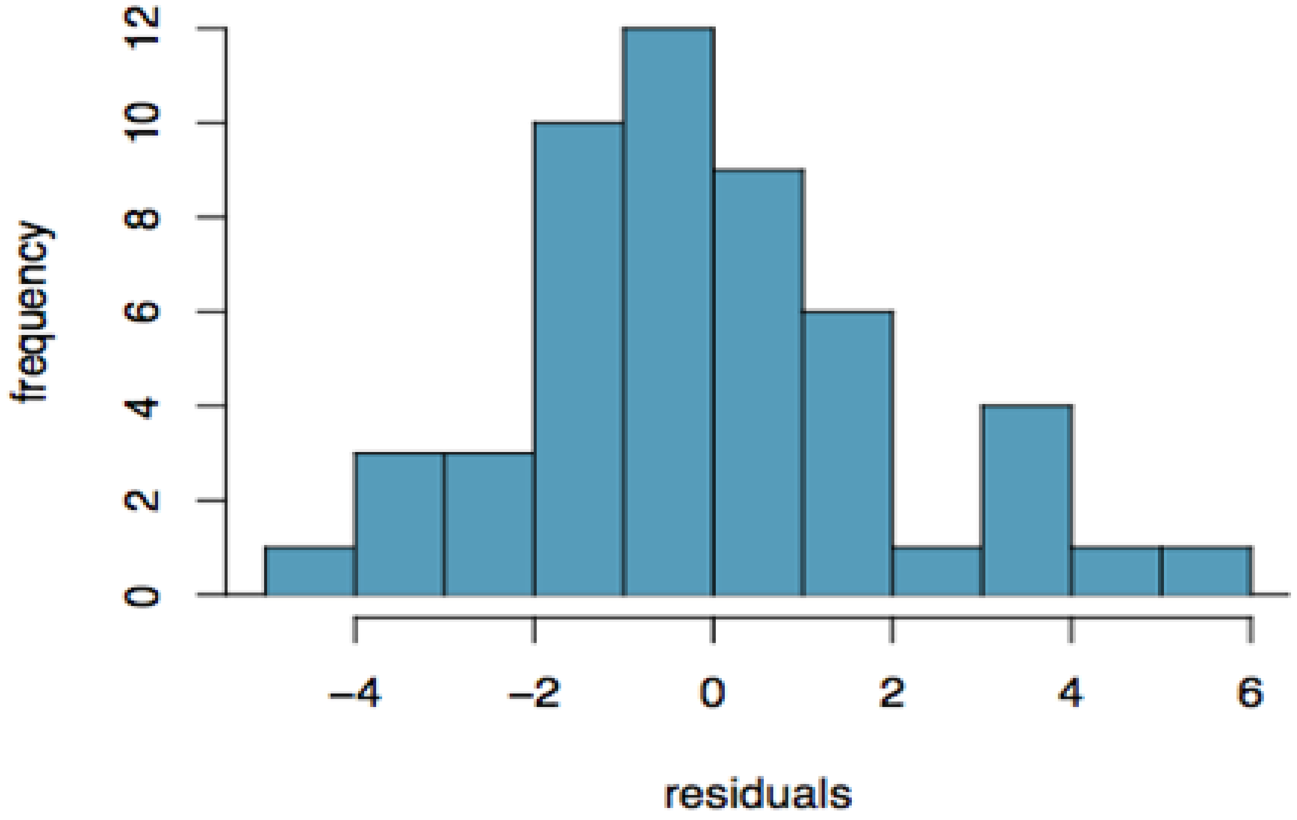
LEAST SQUARES LINE: CONDITIONS

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

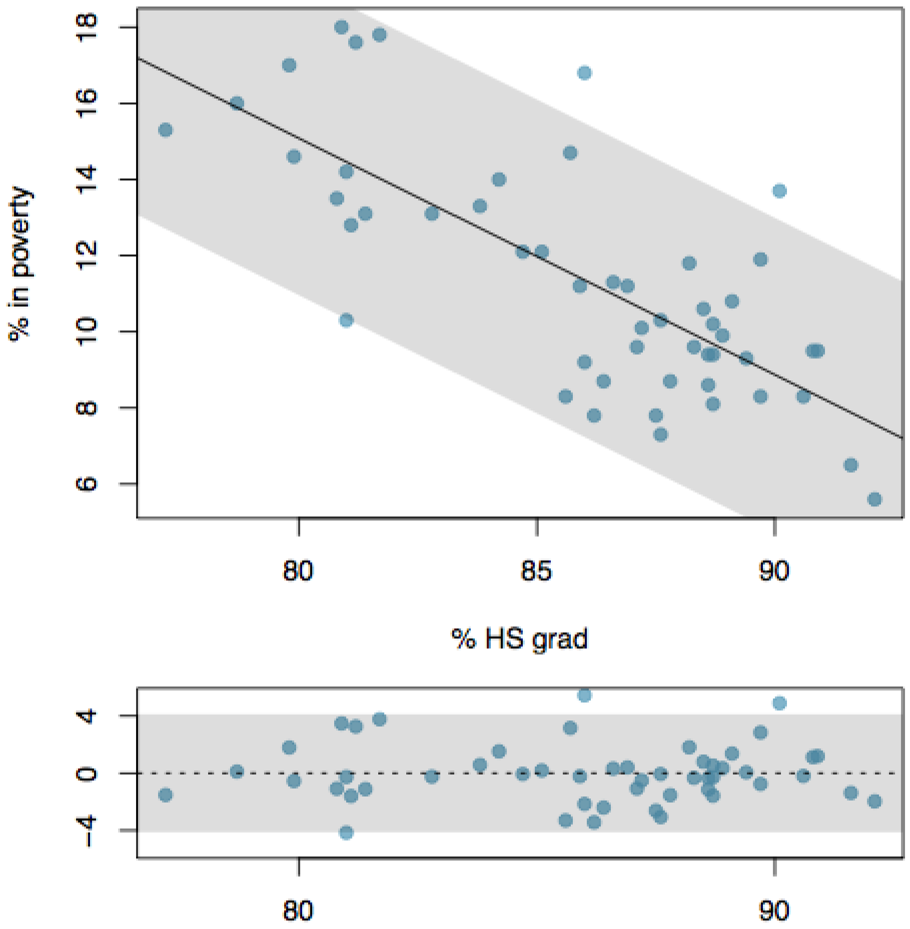
LS LINE: CONDITION: LINEARITY



LS LINE: CONDITION: NORMAL ε_i



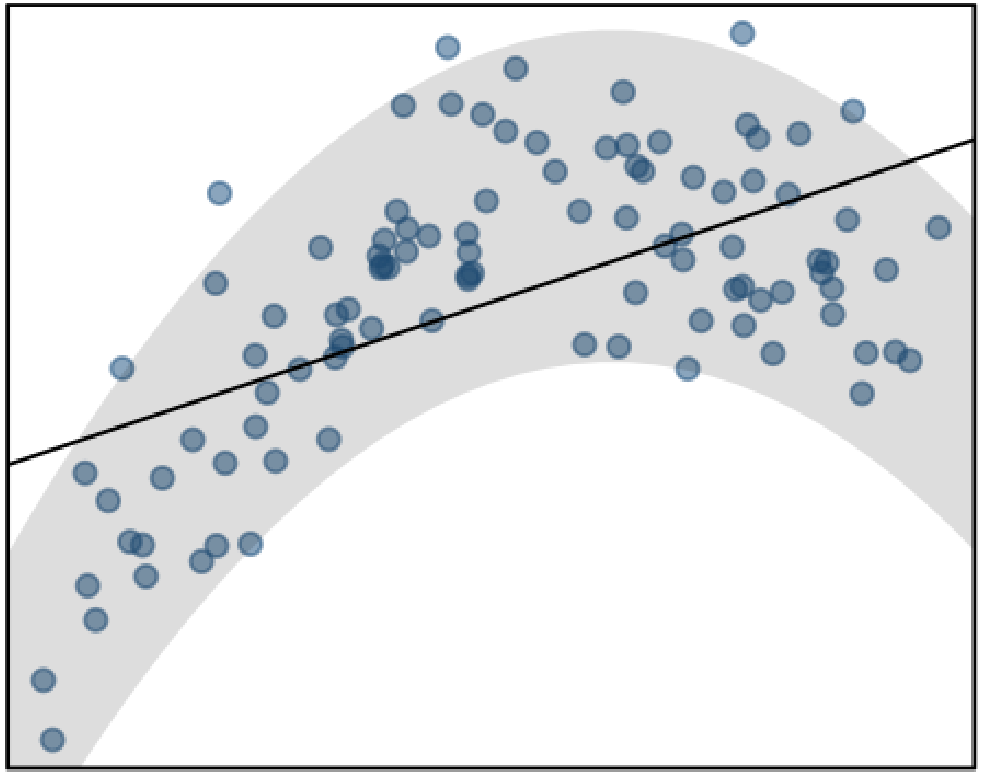
LS LINE: CONDITION: CONSTANT σ^2

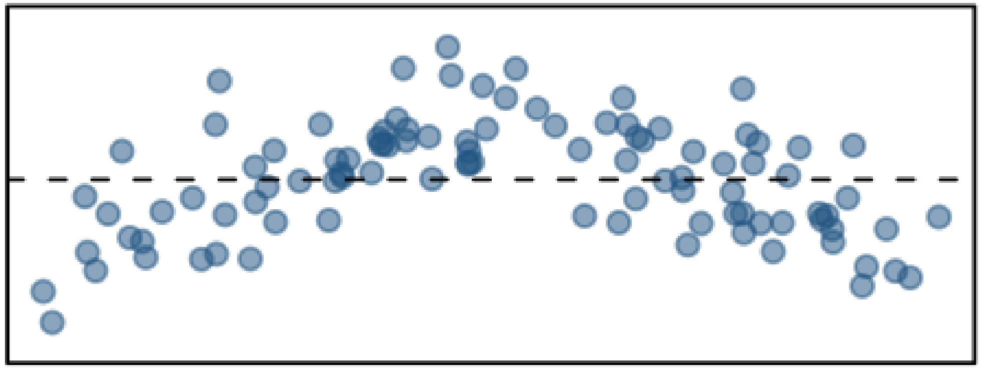


LS LINE: CONDITION: CHECK

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

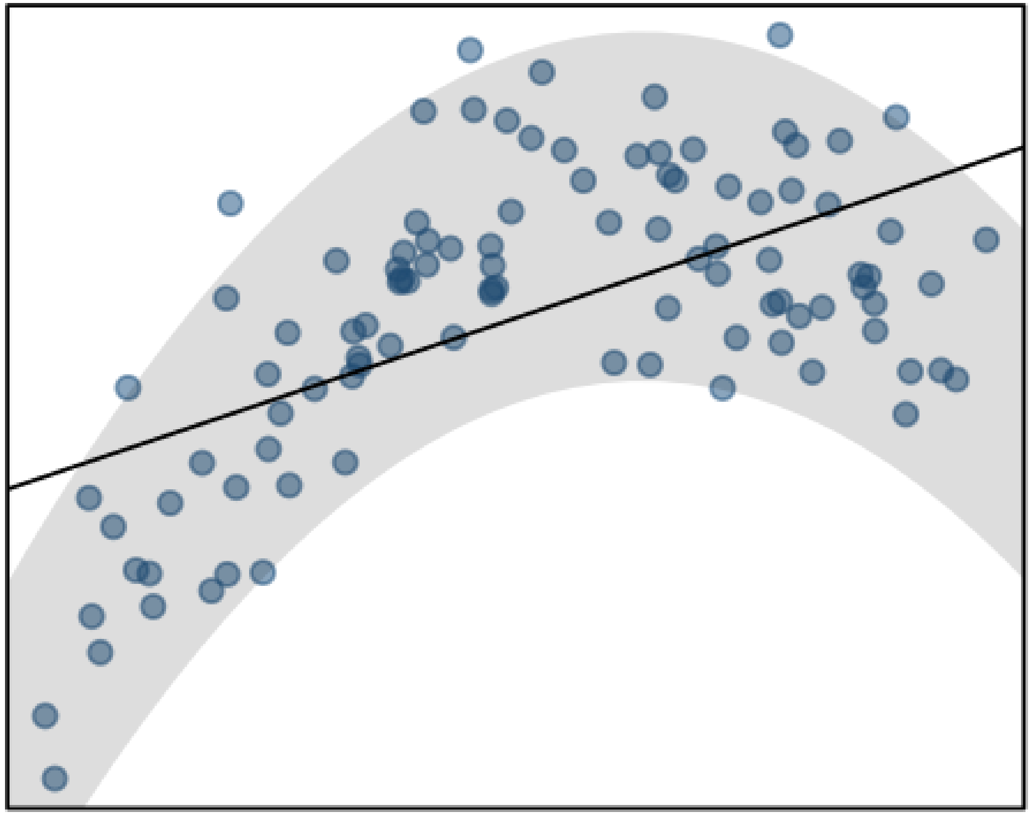
What condition is this linear model obviously violating?

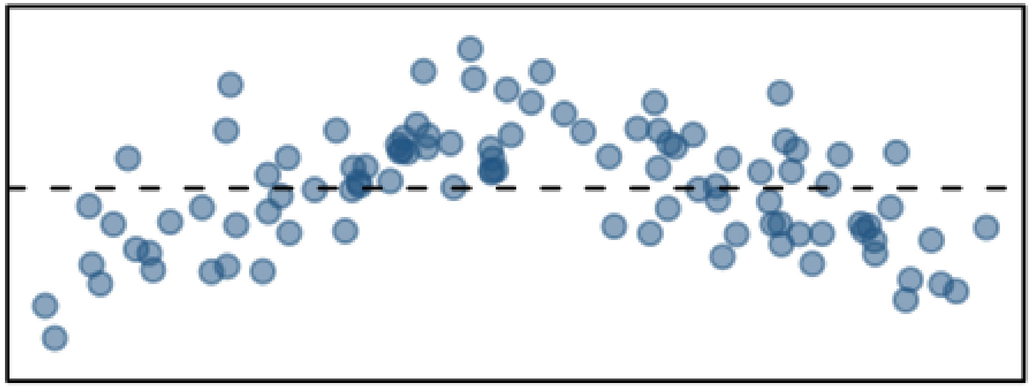
(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers

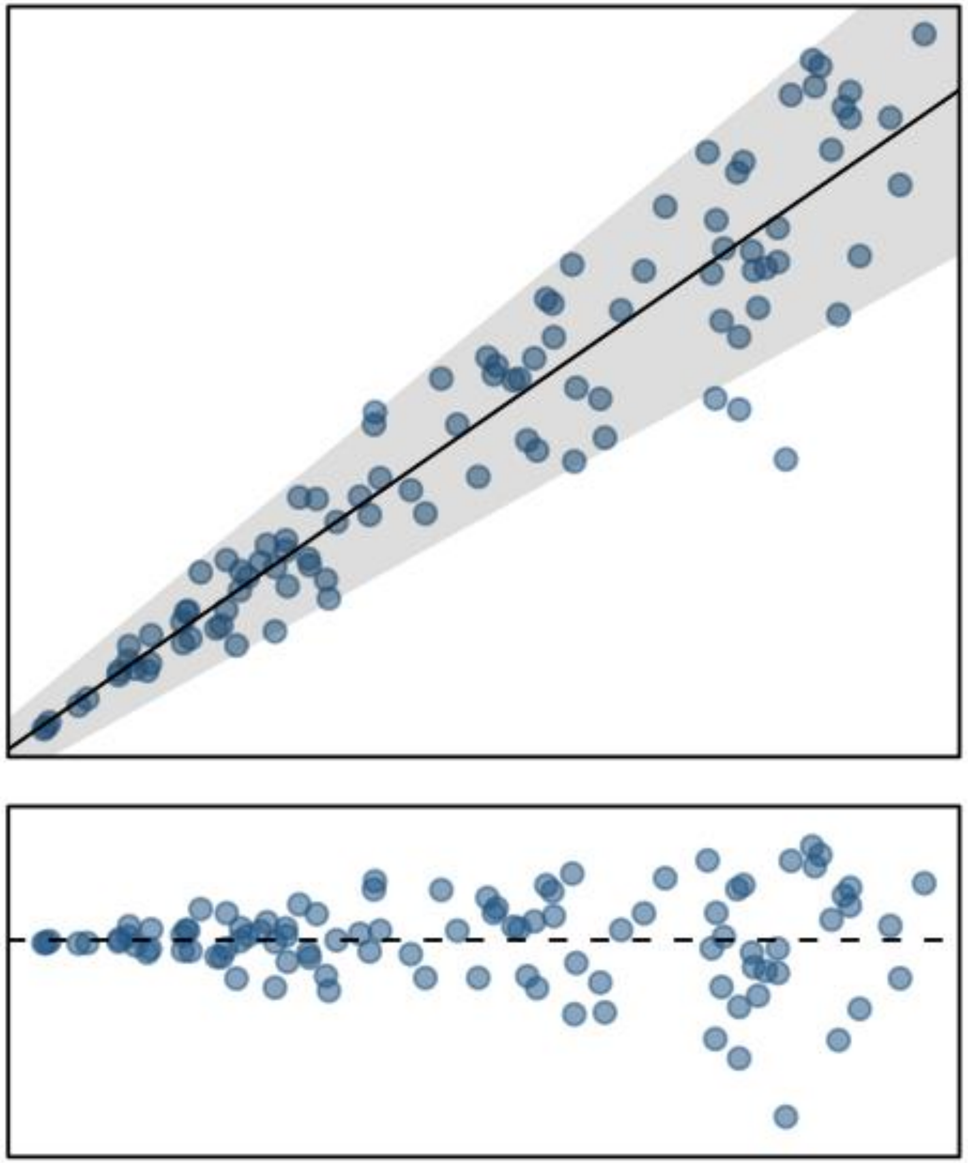




LS LINE: CONDITION: CHECK

What condition is this linear model obviously violating?

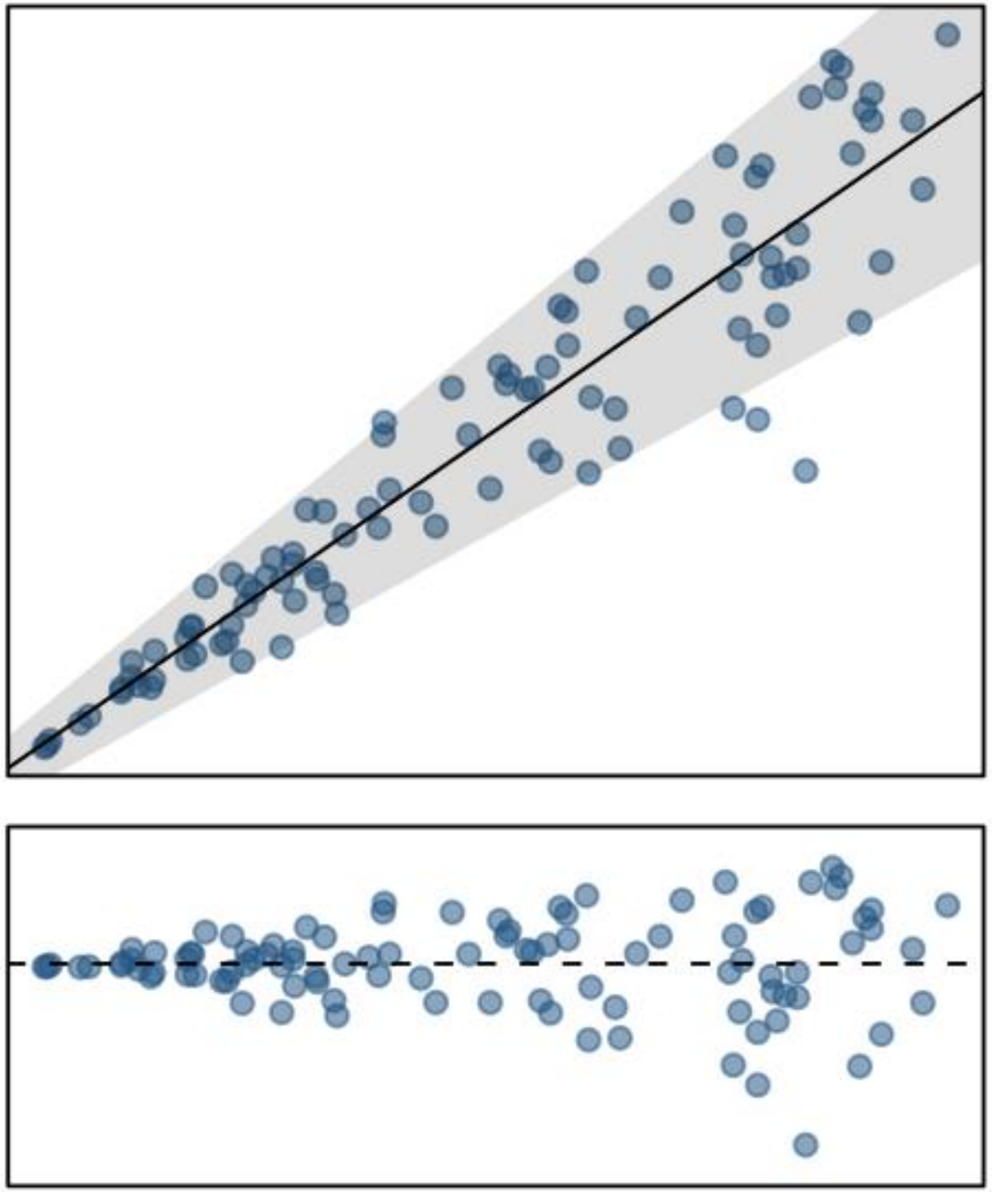
(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



LS LINE: CONDITION: CHECK

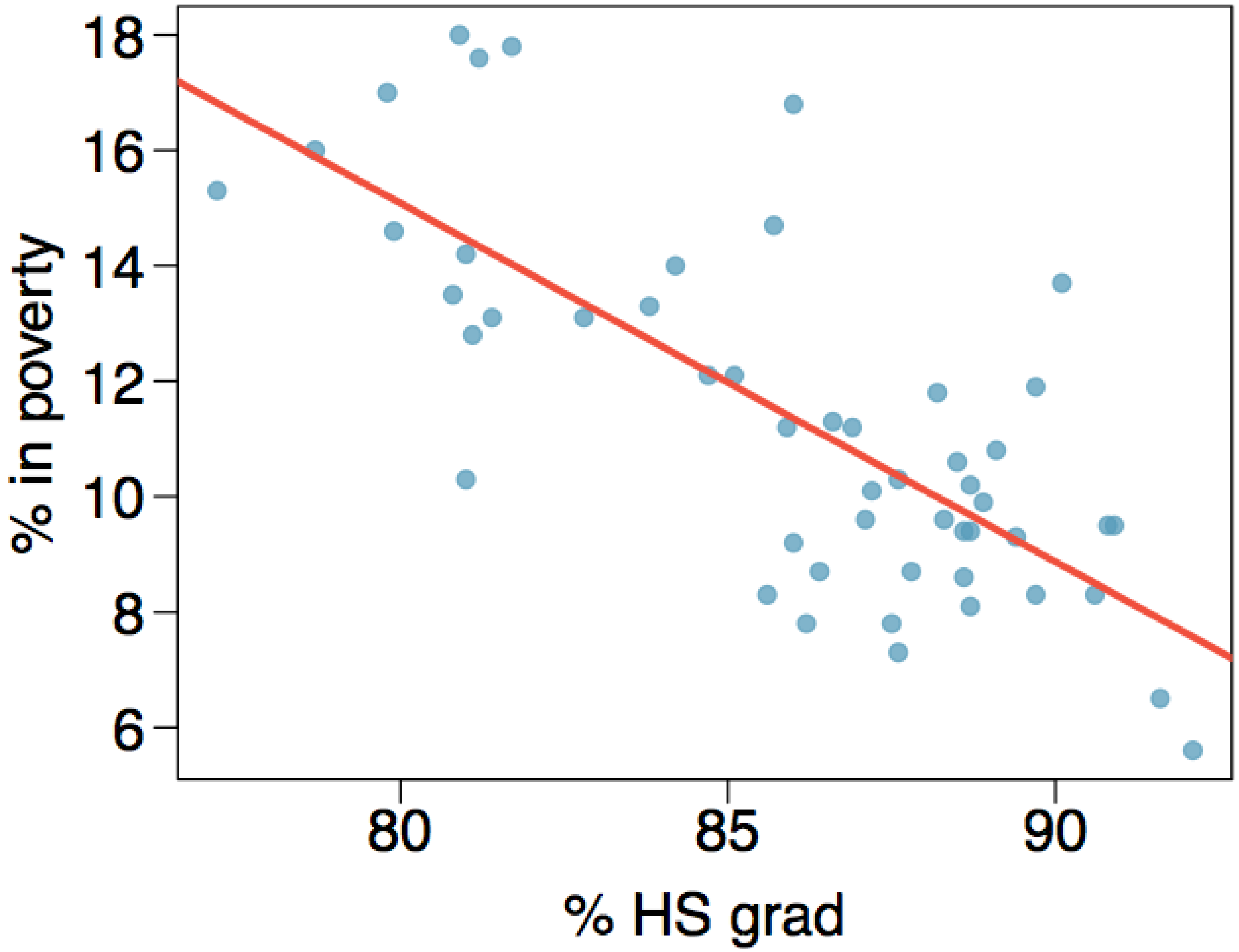
What condition is this linear model obviously violating?

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

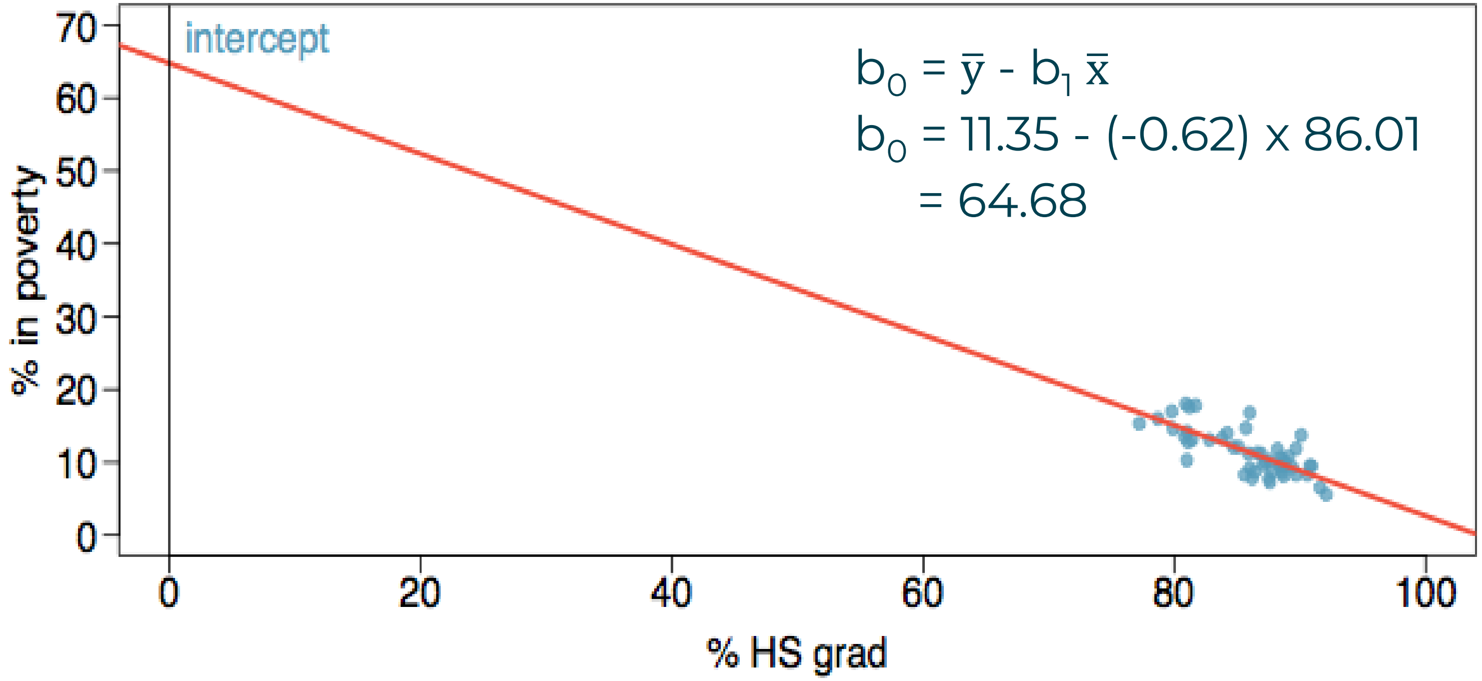


REGRESSION LINE

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



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%.

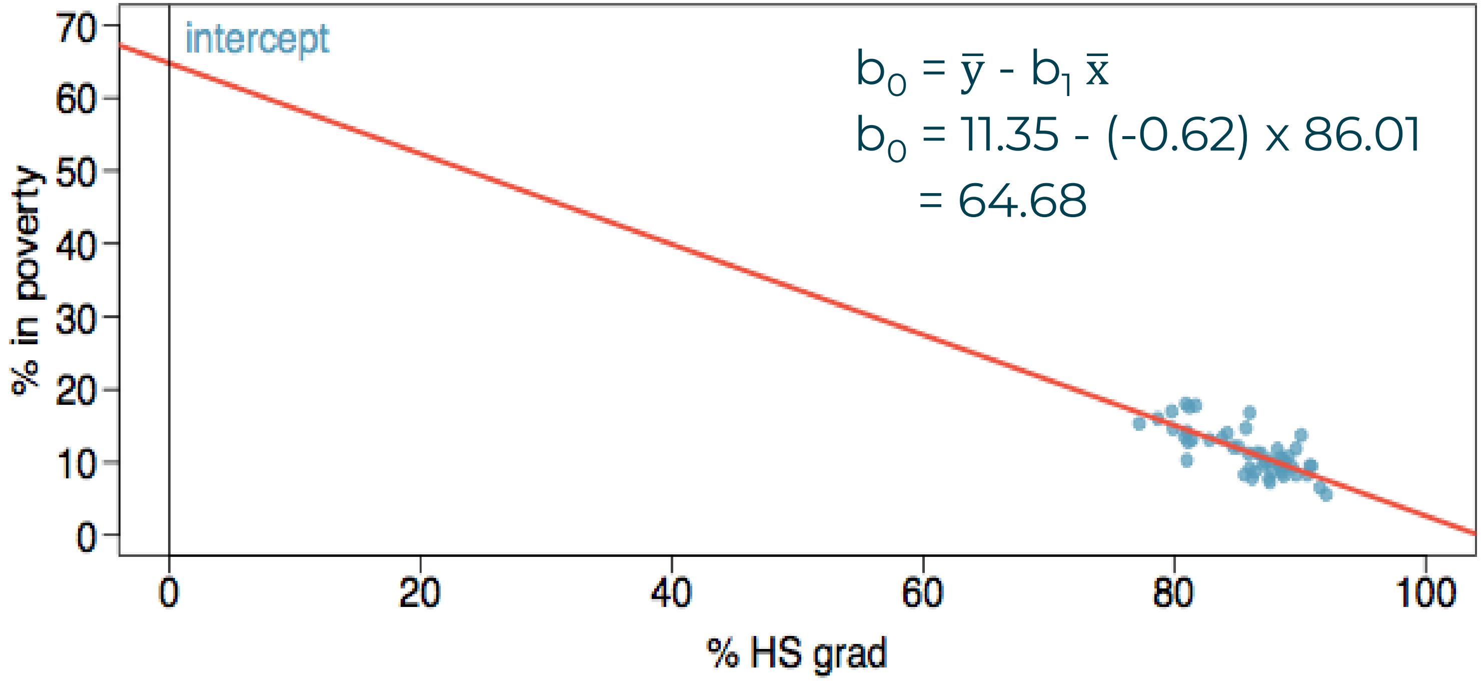
(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(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



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

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

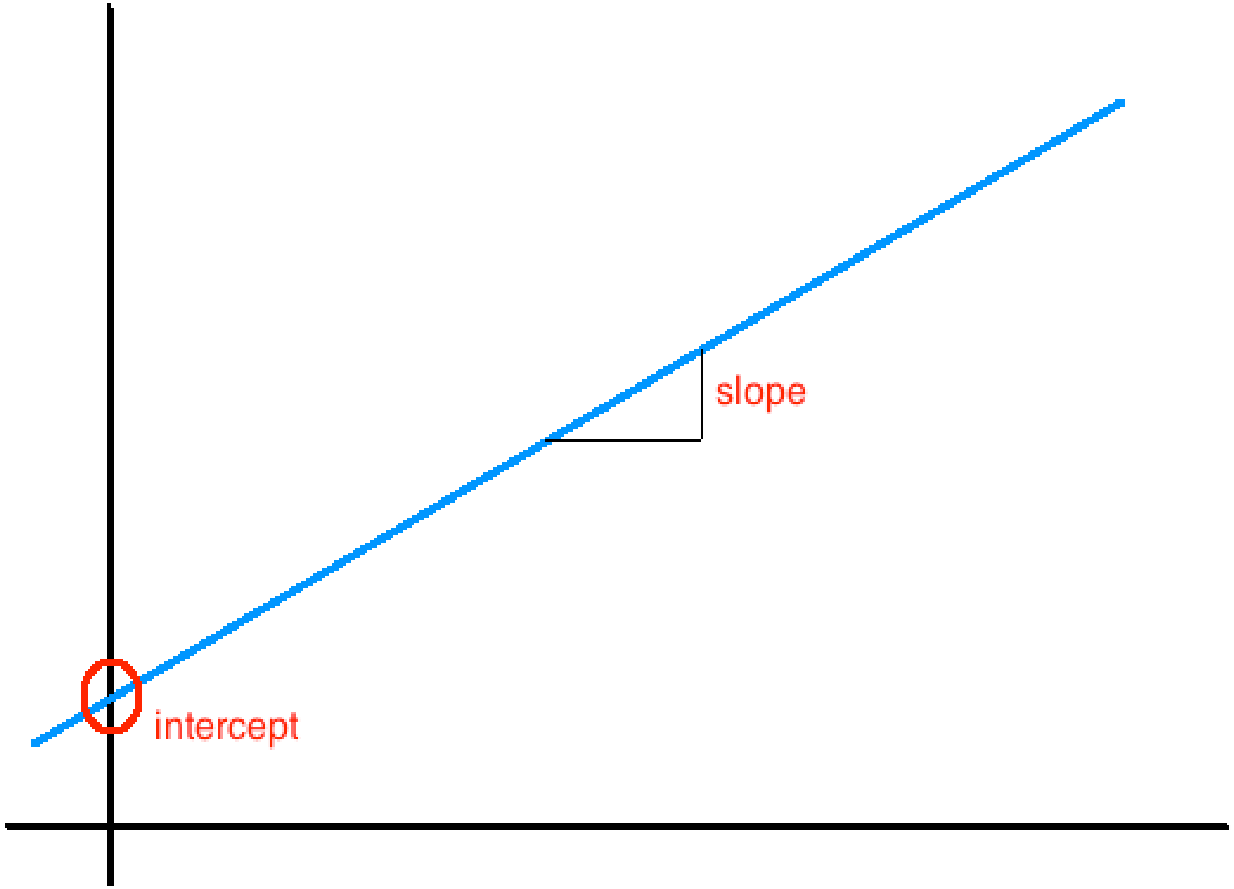
$$= 64.68$$

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

(d)

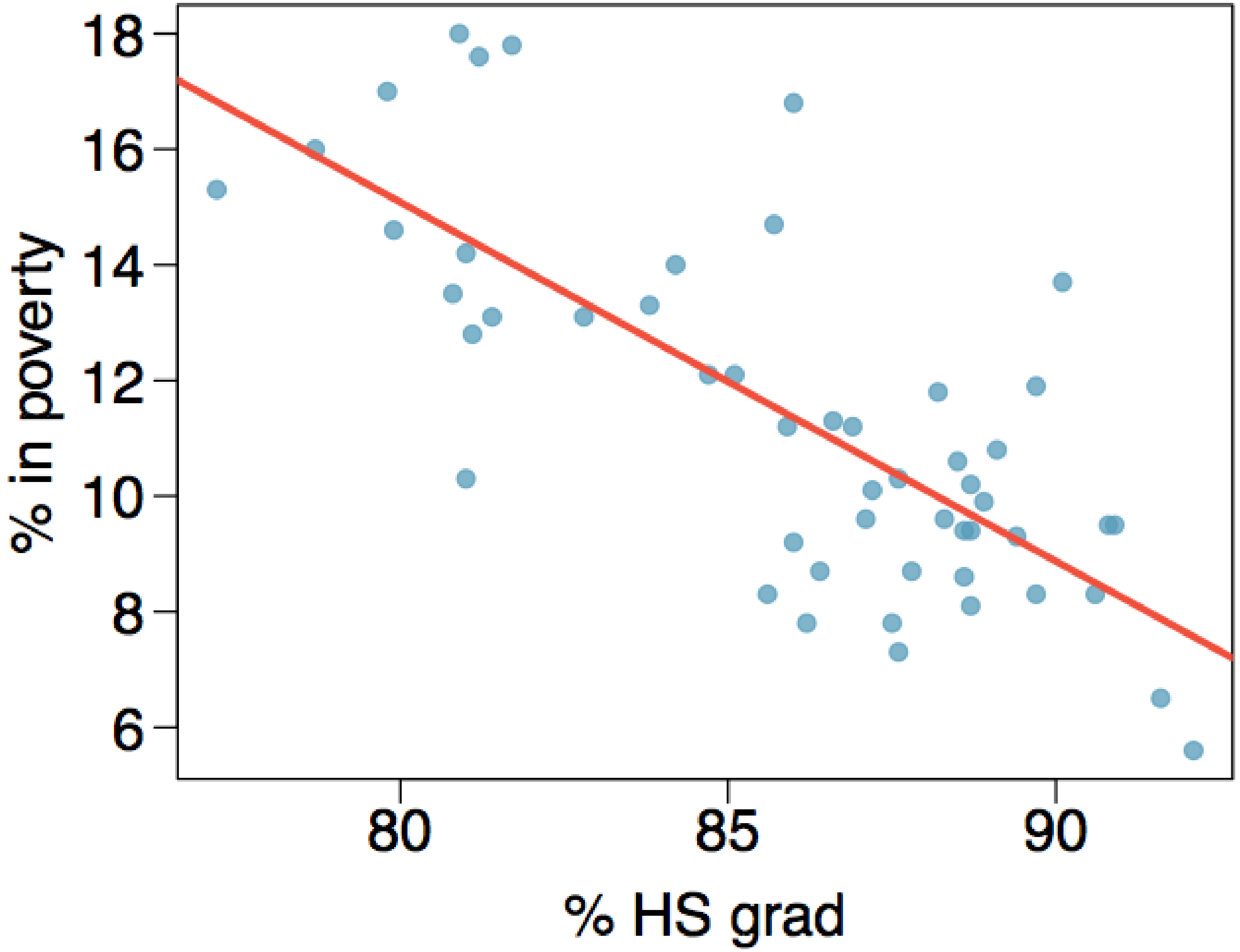
States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line

INTERPRETATION

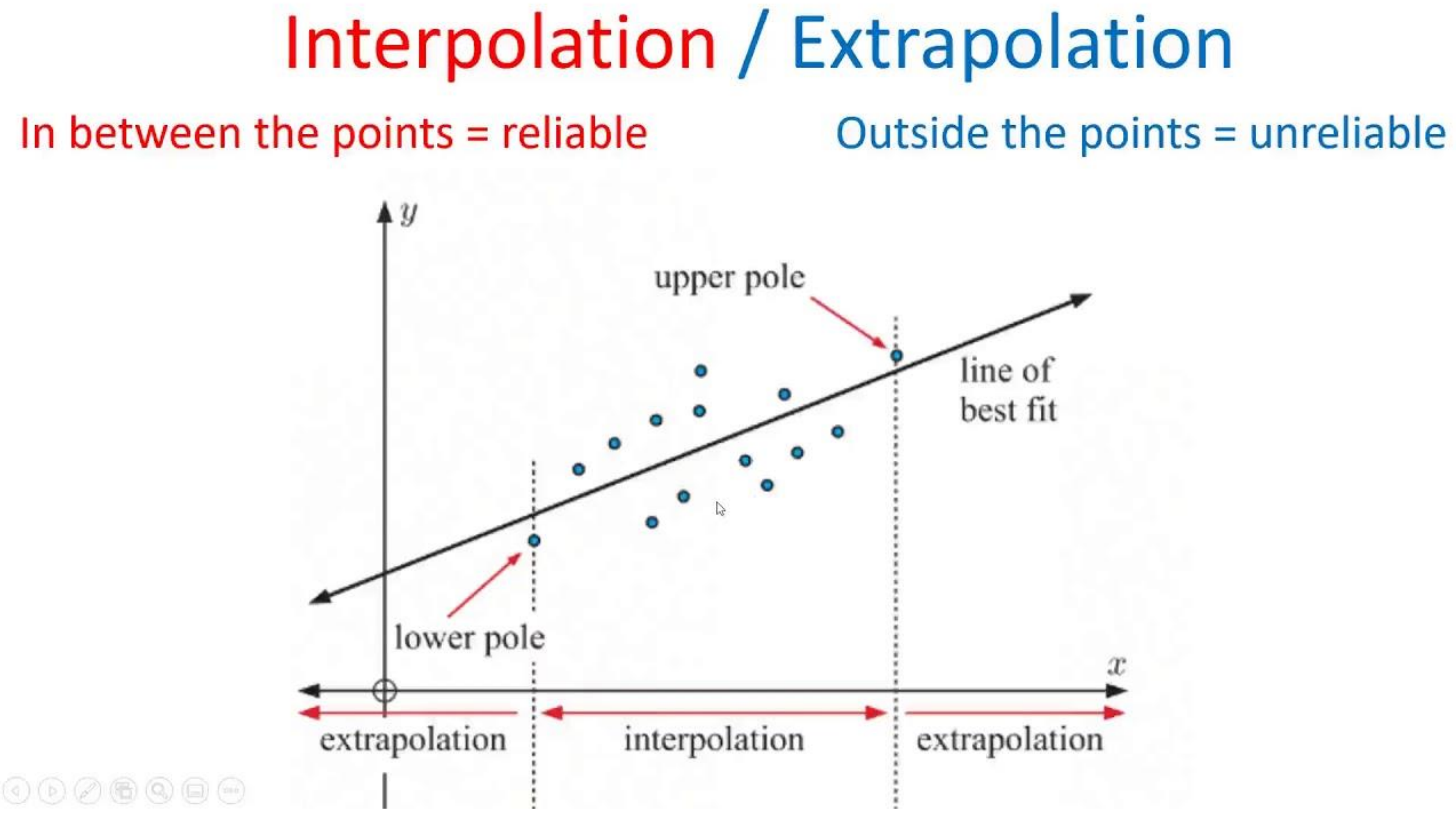


PREDICTION

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



EXTRAPOLATION



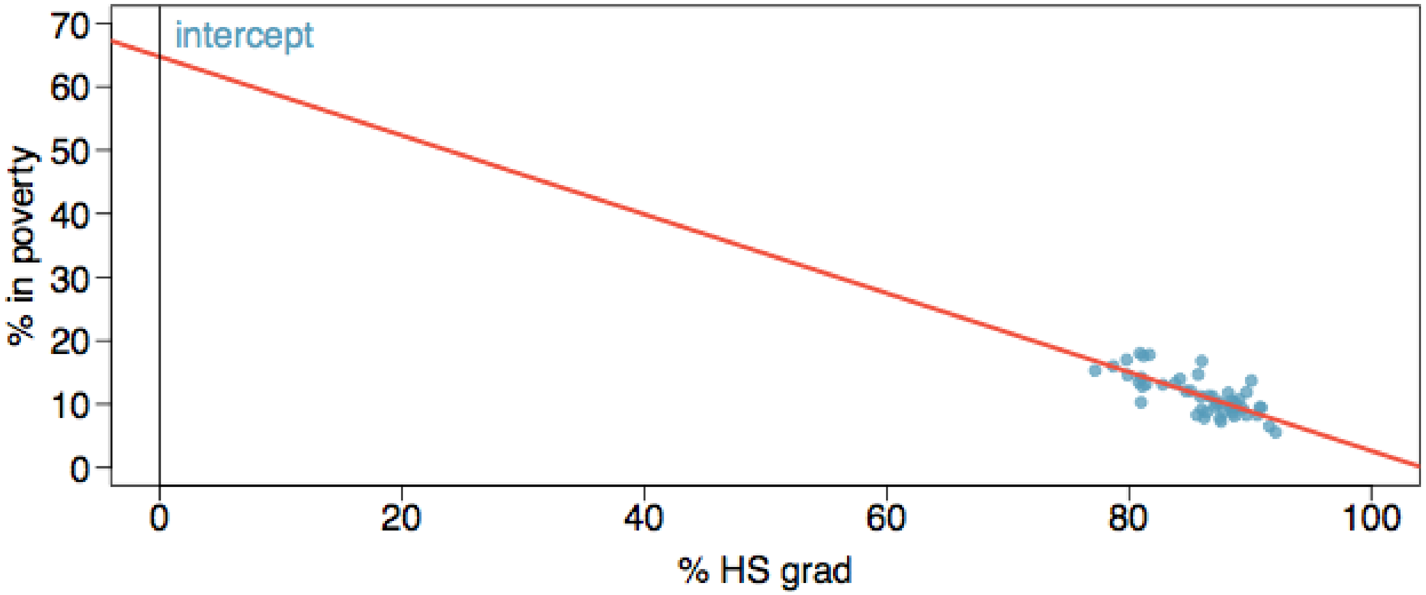
Interpolation / Extrapolation

In between the points = reliable

Outside the points = unreliable

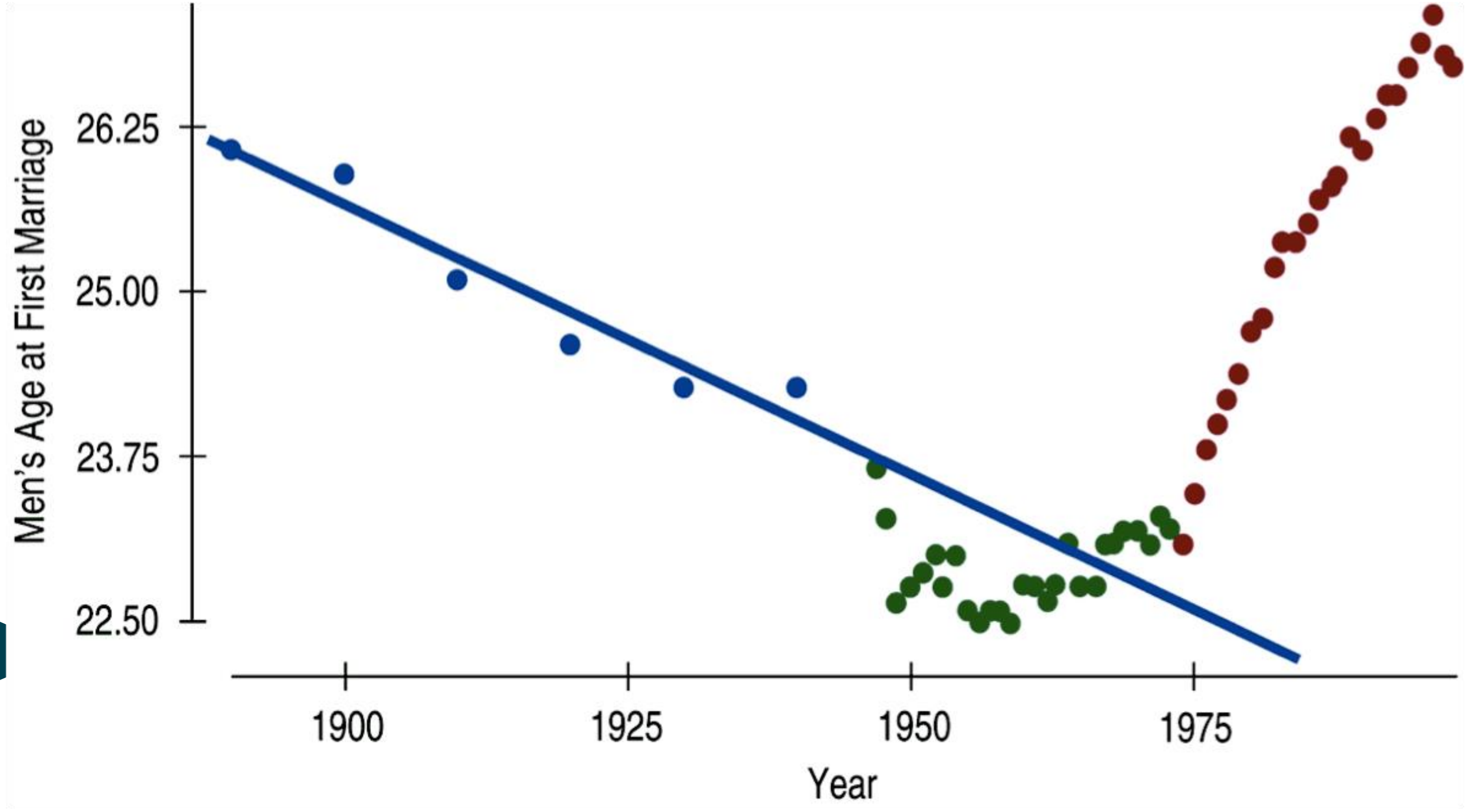
EXTRAPOLATION

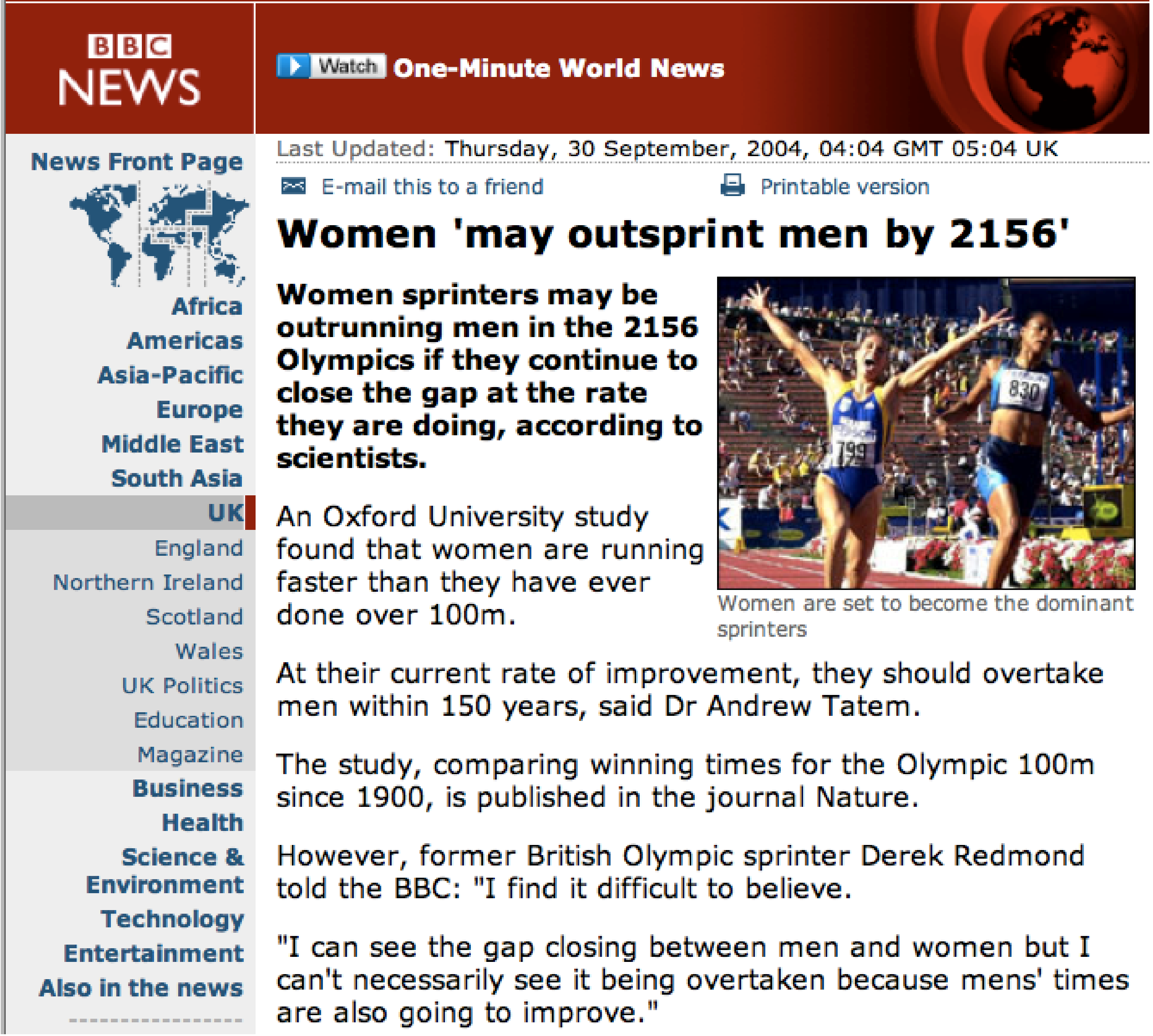
Our Example of finding the intercept



EXTRAPOLATION

Applying a model estimate to values outside of the realm of the original data – if a linear model was built in 1950





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

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.

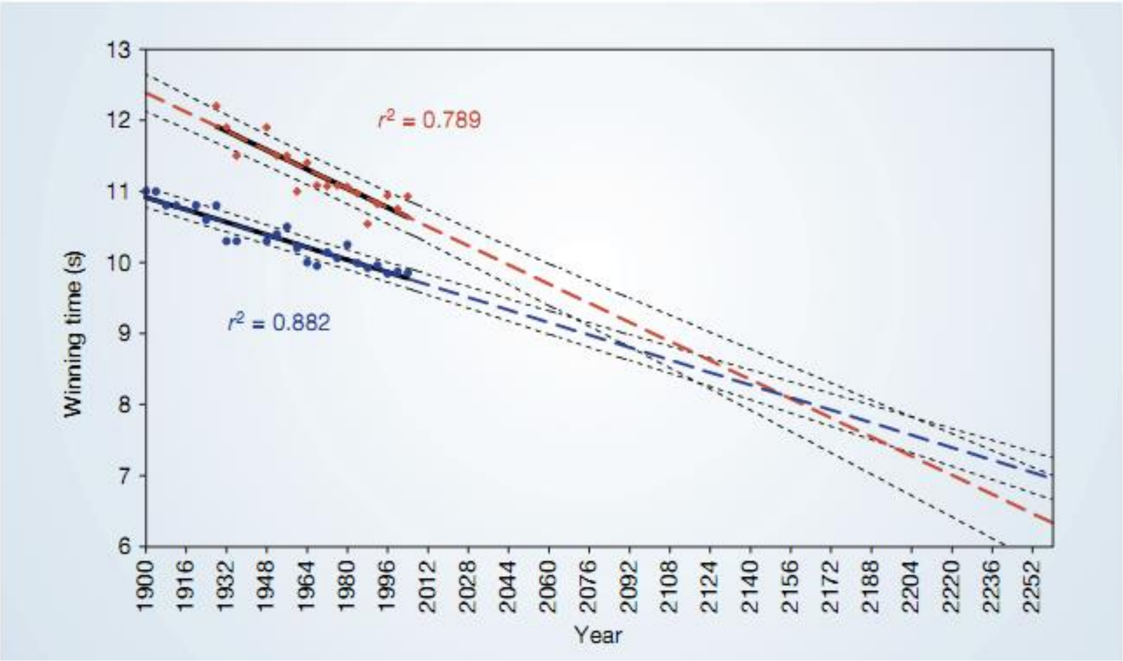
However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"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."

EXTRAPOLATION

Momentous sprint at the 2156 Olympics?

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



WHAT IS R² ?

What is **R²** ?

R² evaluates the strength of the fit of a linear model

R² = square of the correlation coefficient

R² = “coefficient of determination”

What does **R²** tell us?

The 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?

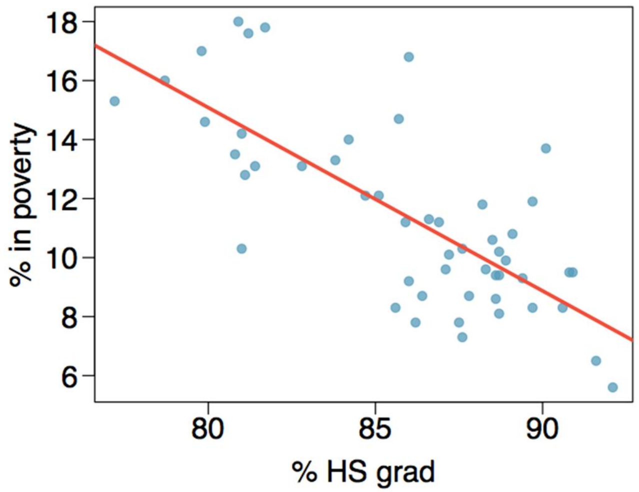
a) 38% of the variability in the % of HS graduates among the 51 states is explained by the model.

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

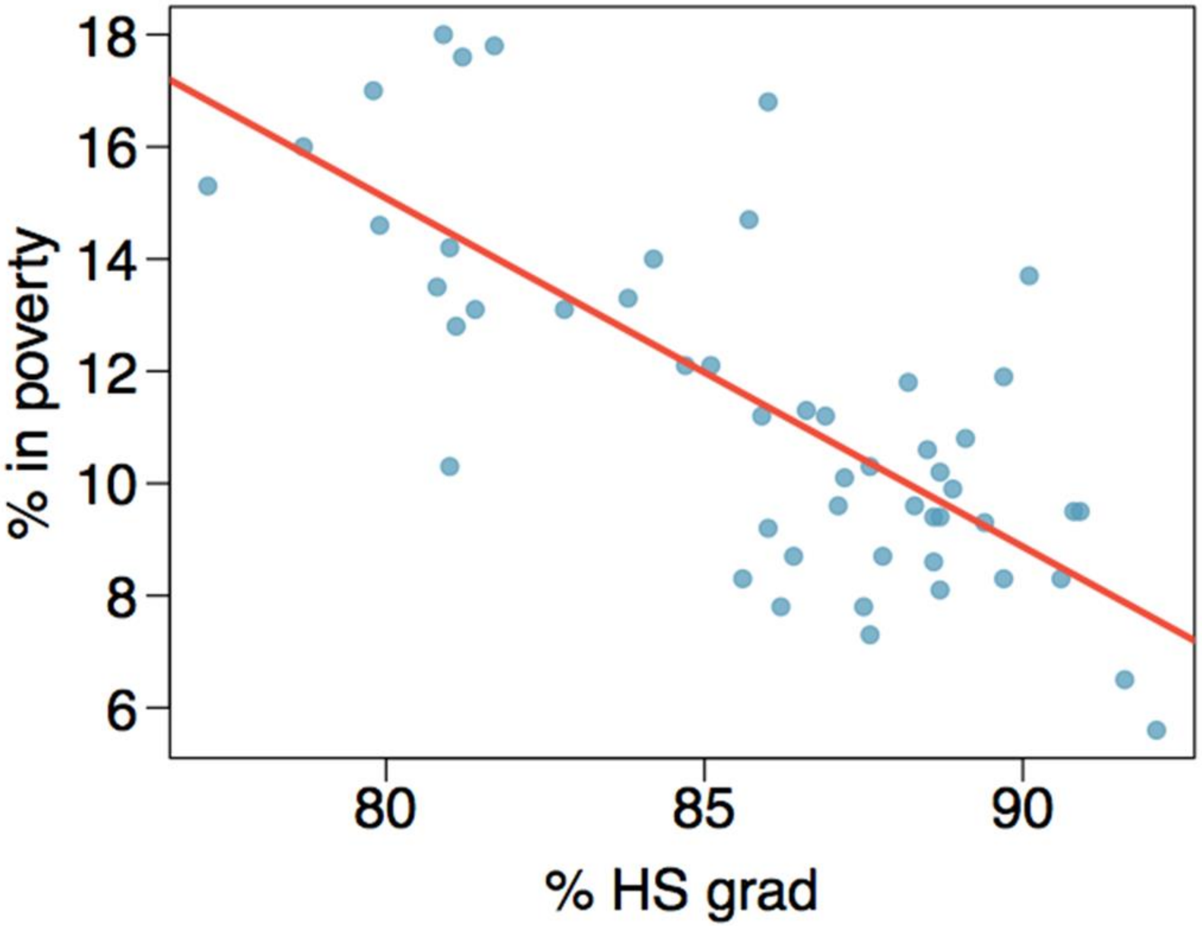
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



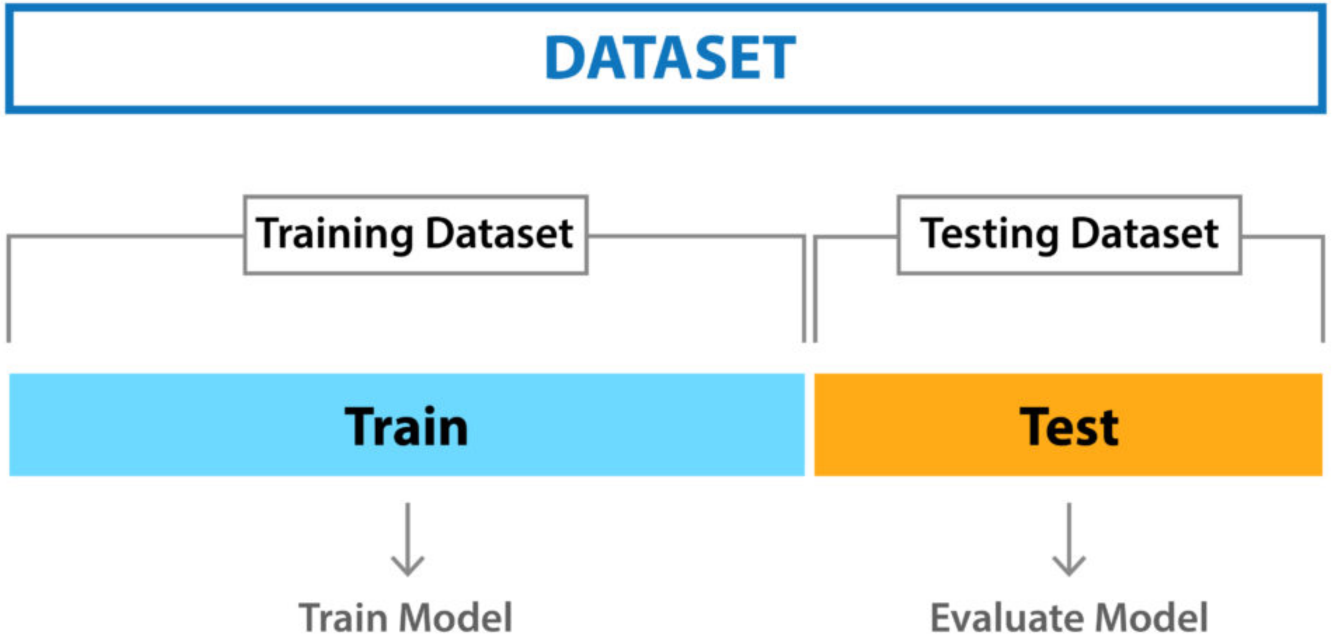
R^2 = 0.38 INTERPRETATION?

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

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



IN MACHINE LEARNING



DATASET

Training Dataset

Testing Dataset

Train

Test

Train Model

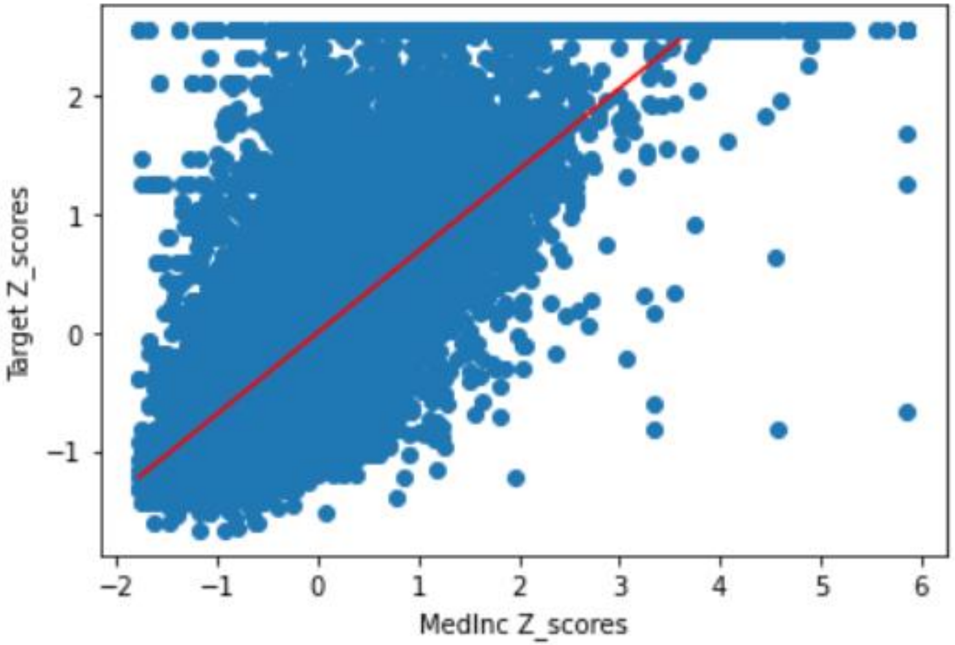
Evaluate Model

COMMON METRICS FOR REGRESSION

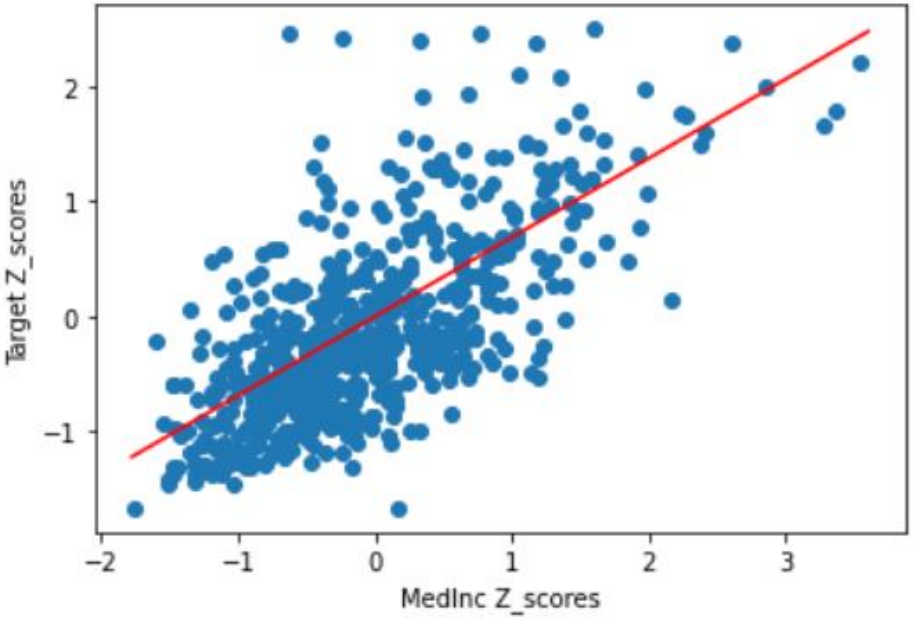
We have 3 main metrics for regression models:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)

EXAMPLE



Training Data

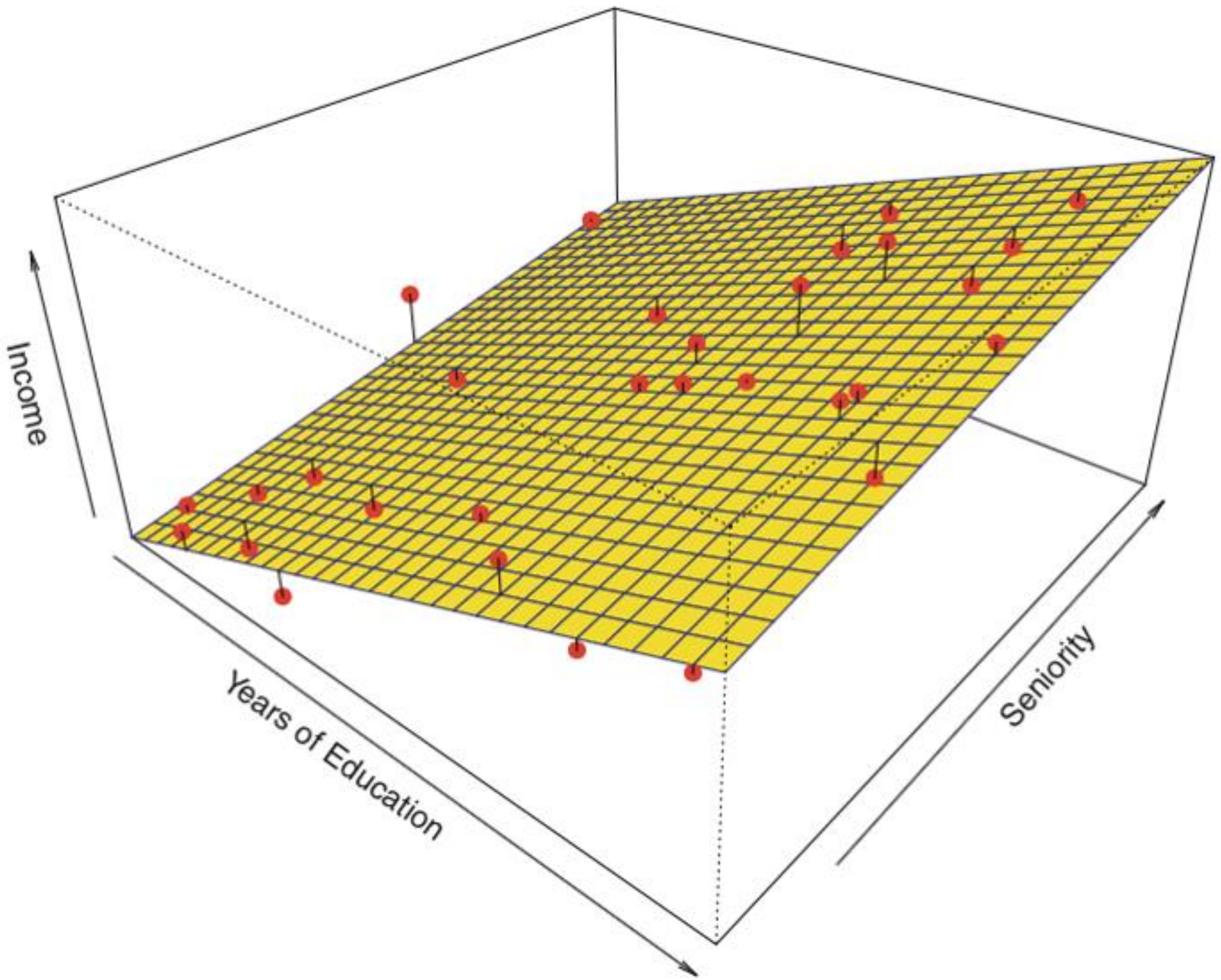


Test Data

MULTI DIMENSIONAL

Multi-Linear Regression is simply ‘tagging’ on more input variables.

As we increase beyond 3 dimensions the model becomes impossible to visualise.



THE MATHS

It is impossible to visualize as we increase dimension, but it is very easy and intuitive to display mathematically

In 2 dimensions, 1 target and 1 feature our regression model was:

$$Y = \beta_0 + \beta_1 x$$

In 3 dimensional space, 1 target and 2 features:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

In 4 dimensional space, 1 target and 3 features:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

And so forth

CURSE OF DIMENSIONALITY

More Dimensions

=

More problems

DEMO AND EXERCISES

Open the linear regression notebook

We will go through a demo, building a linear regression model, first with one feature then with more.