Advanced Statistics
DS2003 (BDS-4A)
Lecture 17

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

19 April, 2022

Previous Lecture

- Difference in more than two means (ANOVA), continued
- Which group is different?
  - The Bonferroni correction (more stringent significance level) & pairwise comparisons
- More 1-way ANOVA tests
  - Example: Vertical Jump height measurements of three groups of males
  - Using Excel for ANOVA tests: https://www.youtube.com/watch?v=ZvfO7-J5u34

Plan for Today

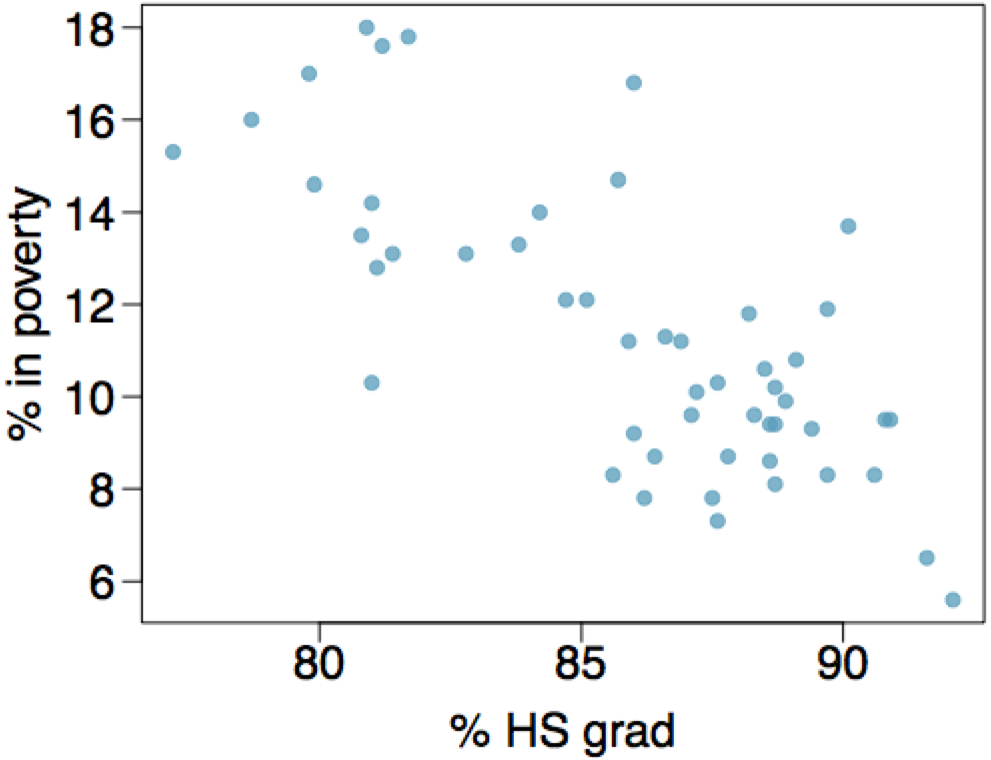
- Line Fitting, Residuals, and Correlation

Modeling numerical variables

- In this unit we will learn to quantify the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.

Poverty vs. High School (HS) graduate rate

- 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)*.

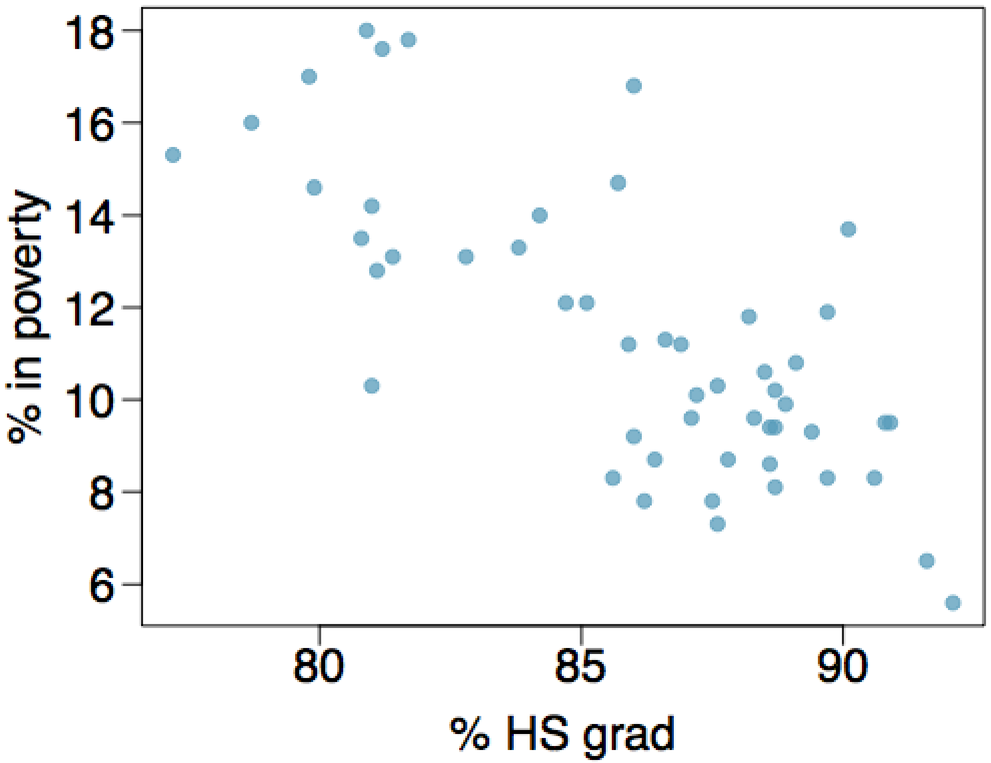


Response variable?

% in poverty

Poverty vs. High School (HS) graduate rate

- 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?

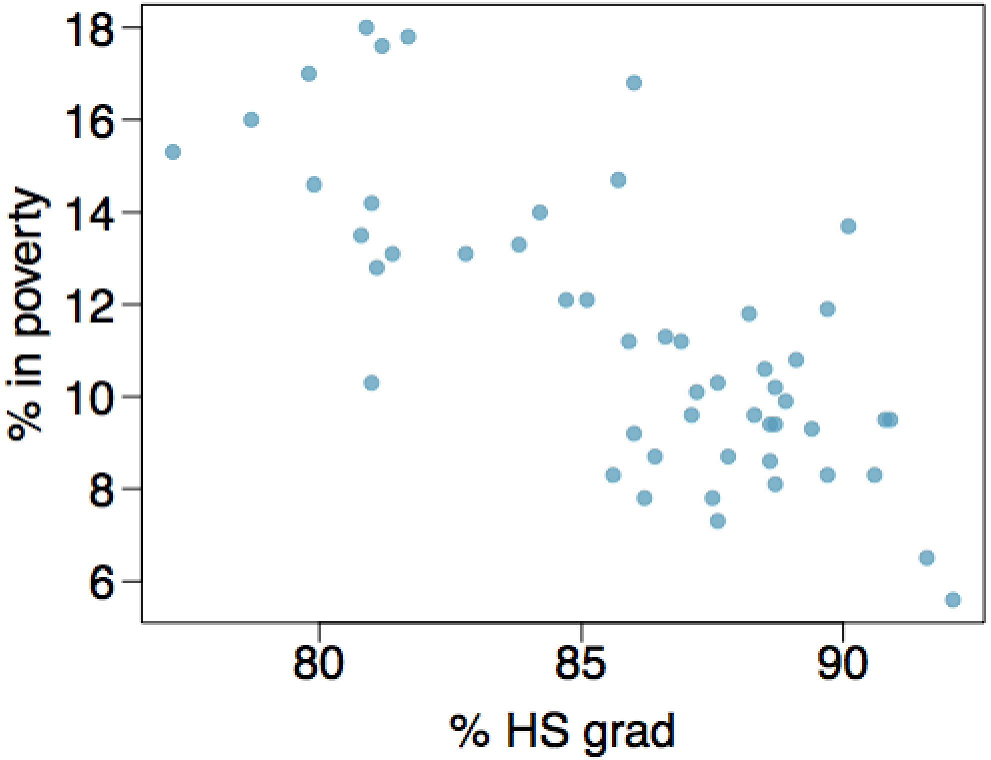
% in poverty

Explanatory variable?

% HS grad

Poverty vs. High School (HS) graduate rate

- 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

Poverty vs. HS graduate rate

The linear model for predicting poverty from high school graduation rate in the US is:

$$\hat{poverty} = 64.78 - 0.62 * HS_{grad}$$

The hat ^ is used to signify that this is an estimate.

Poverty vs. HS graduate rate

The linear model for predicting poverty from high school graduation rate in the US is:

$$\widehat{poverty} = 64.78 - 0.62 * HS_{grad}$$

The hat ^ is used to signify that this is an estimate.

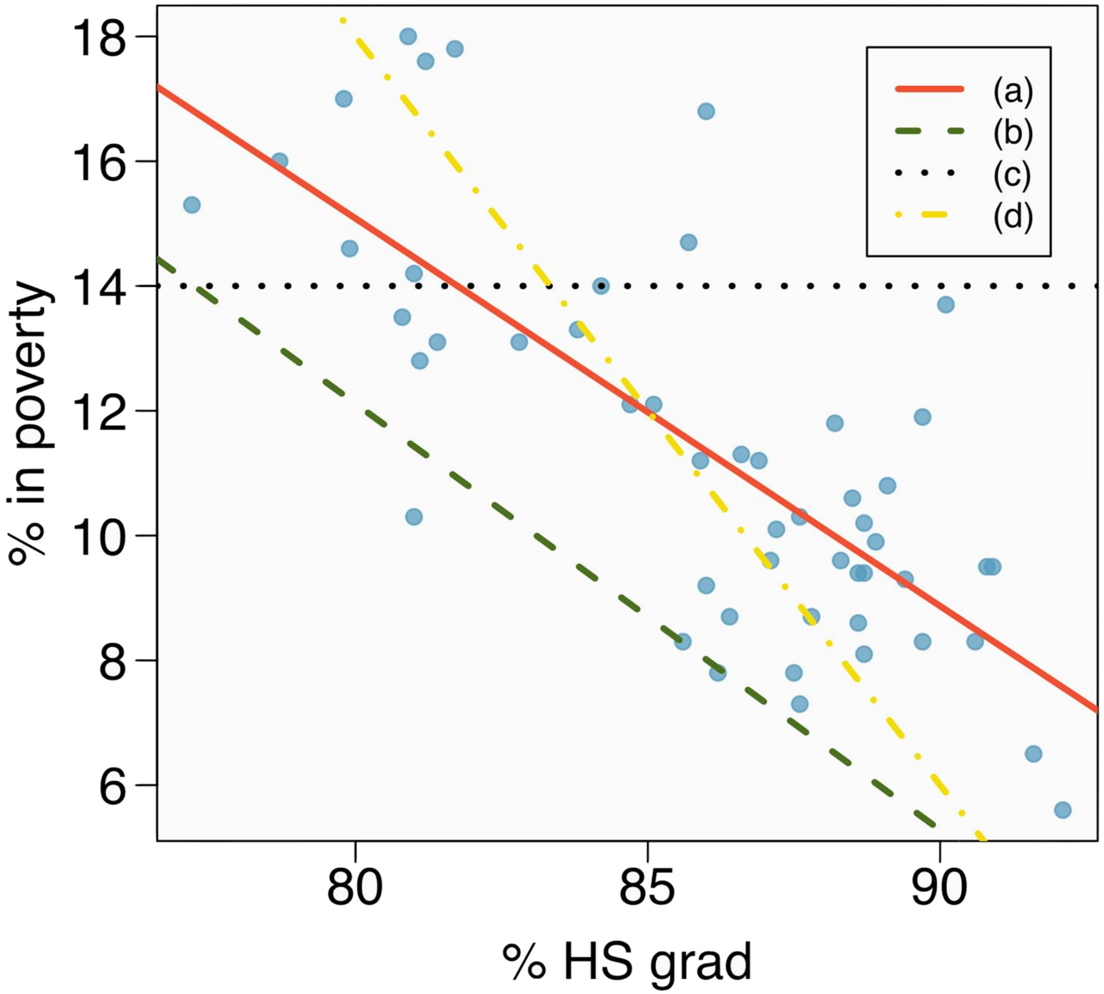
The high school graduate rate in Georgia is 85.1%. What poverty level does the model predict for this state?

64.78 – 0.62 x 85.1 = 12.018

Eyeballing the line

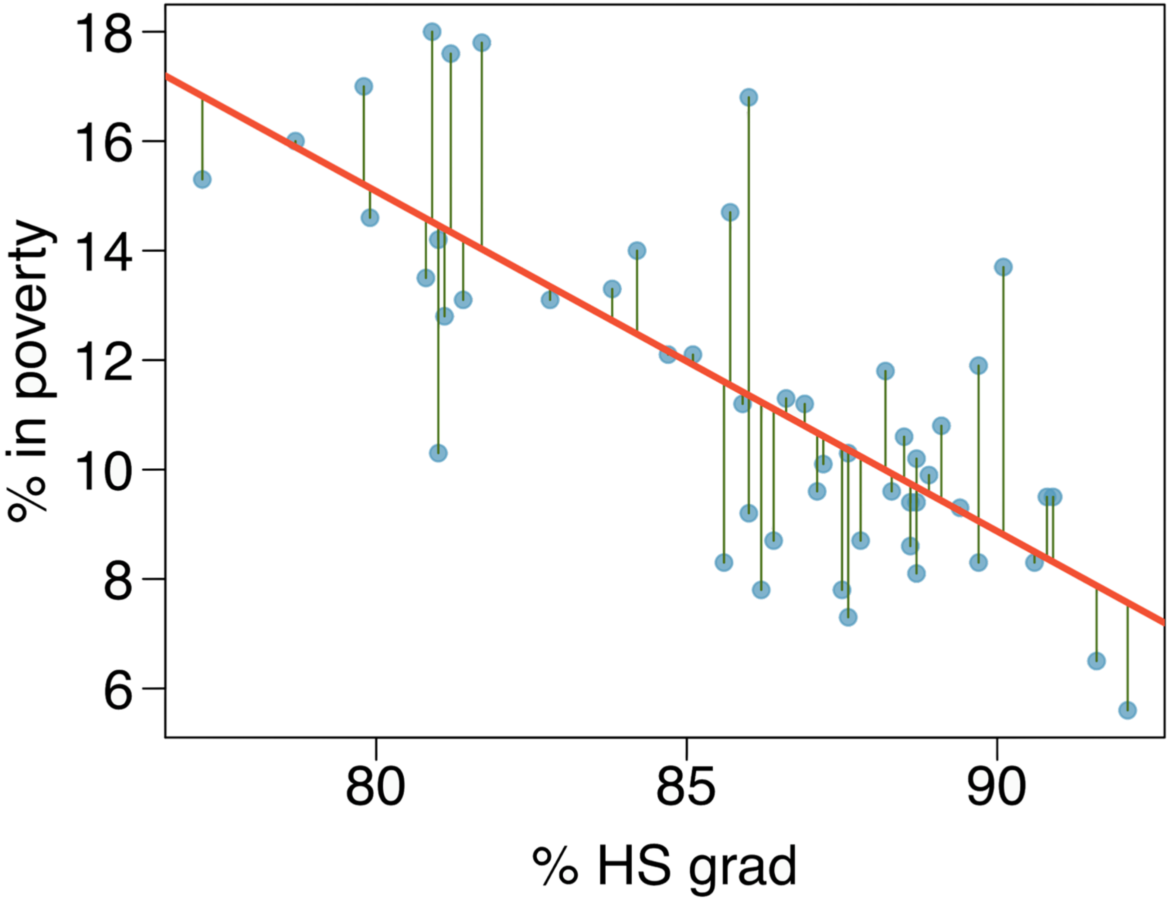
- Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



Residuals

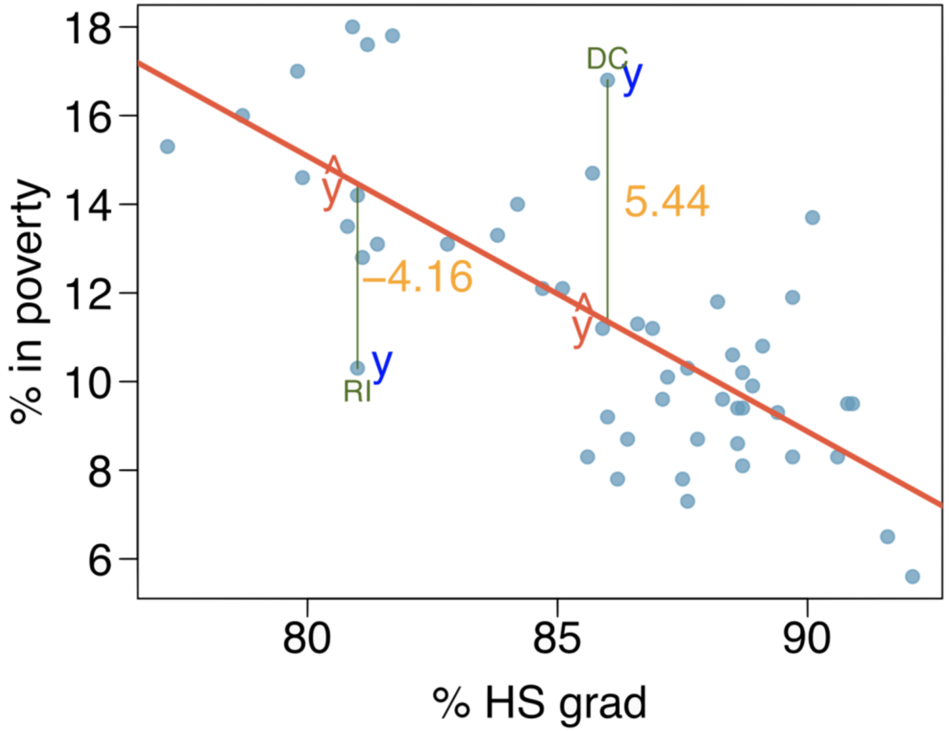
Residuals are the leftovers from the model fit: *Data = Fit + Residual*



Residuals (cont.)

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

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

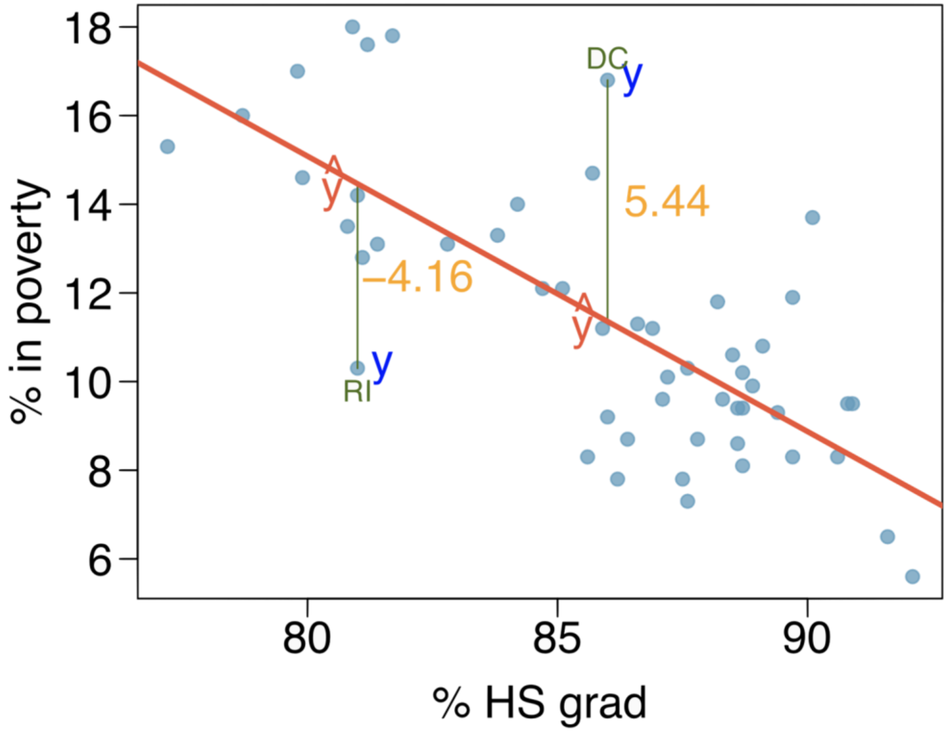


% living in poverty in DC is 5.44% more than predicted.

Residuals (cont.)

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

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



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

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

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.

Guessing the correlation

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

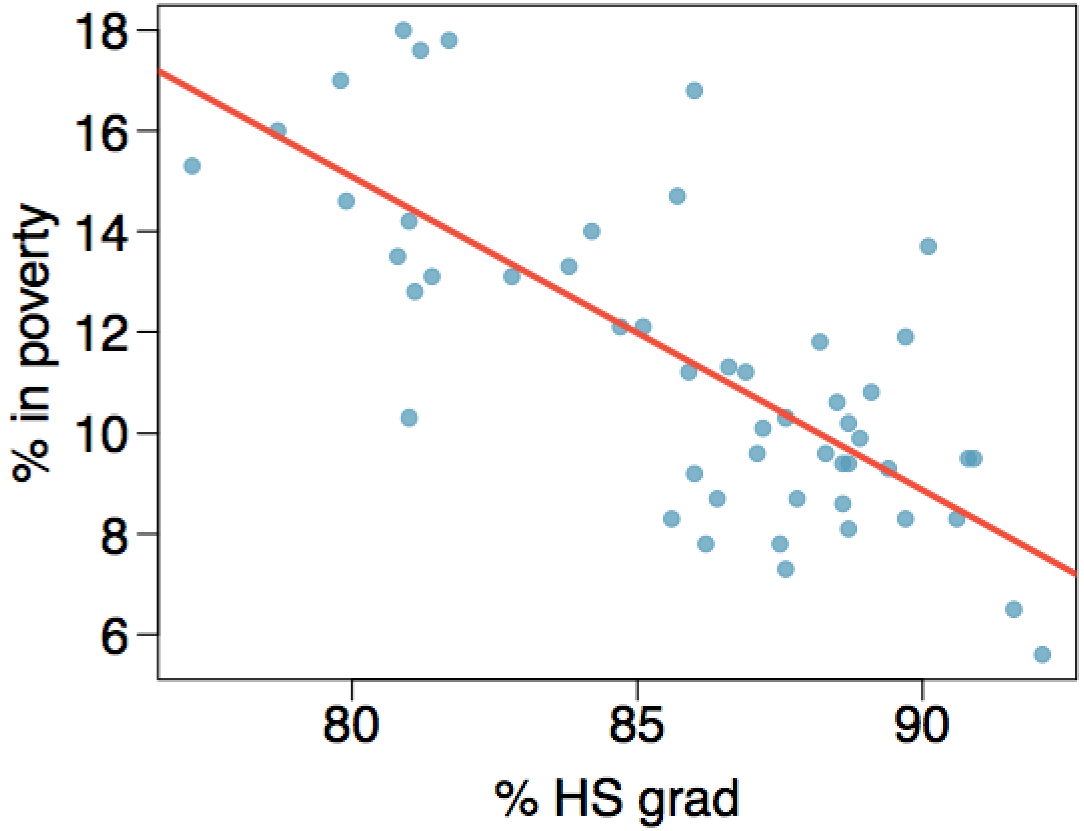
(a) 0.6

(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?

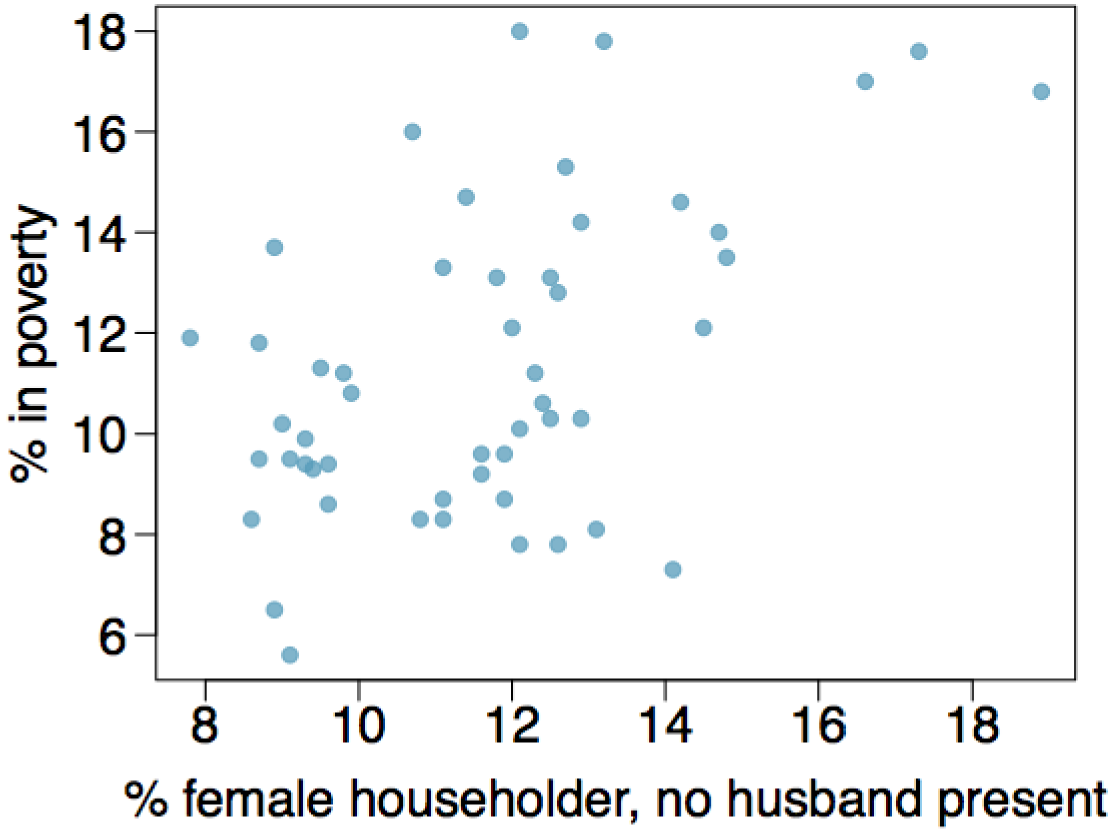
(a) 0.1

(b) -0.6

(c) -0.4

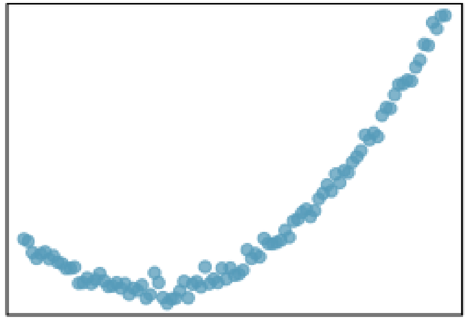
(d) 0.9

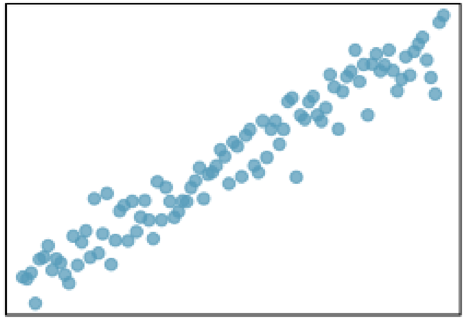
(e) 0.5 ⬅

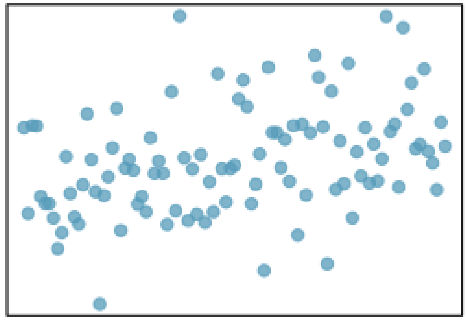


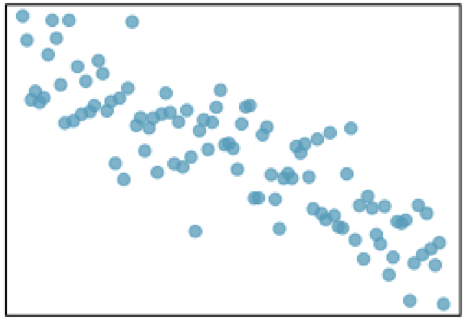
Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?


(a)


(b)


(c)


(d)

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

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

  2. Option 2: Minimize the sum of squared residuals -- *least squares*

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

A measure for the best line

- We want a line that has small residuals
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

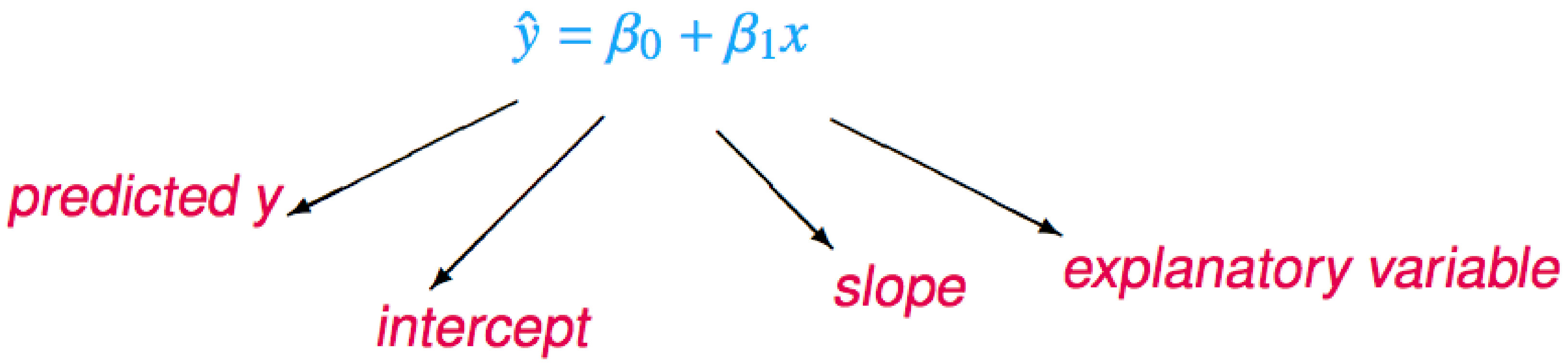
     $$|e_1| + |e_2| + \ldots + |e_n|$$

  2. 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

The least squares line



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

predicted y

intercept

slope

explanatory variable

Notation:

- Intercept:
    - Parameter: β_0
    - Point estimate: b_0

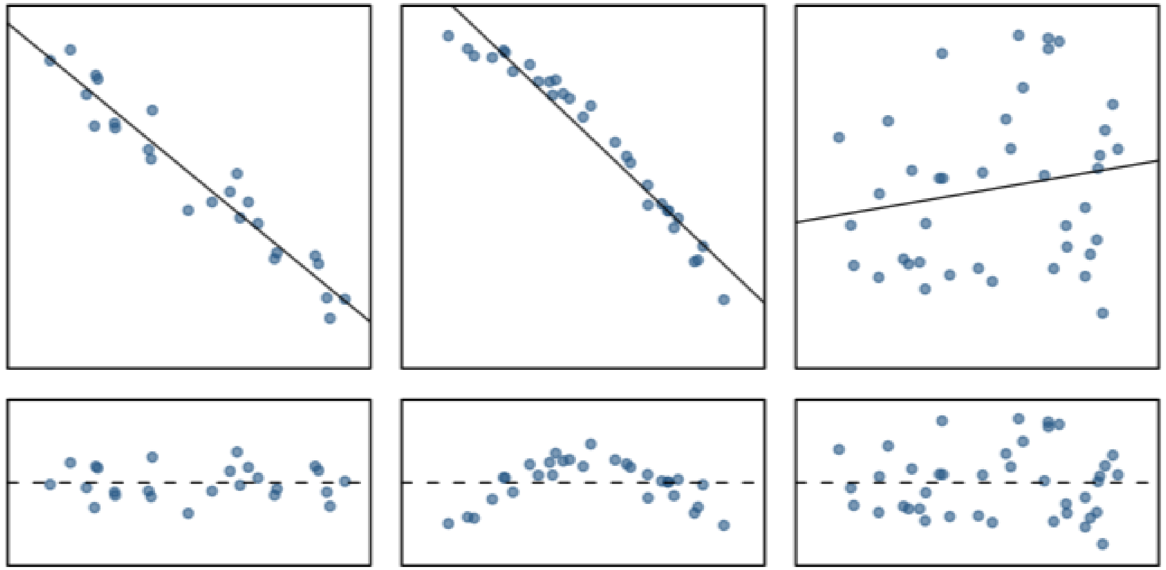
- Slope:
    - Parameter: β_1
    - Point estimate: b_1

Conditions for the least squares line

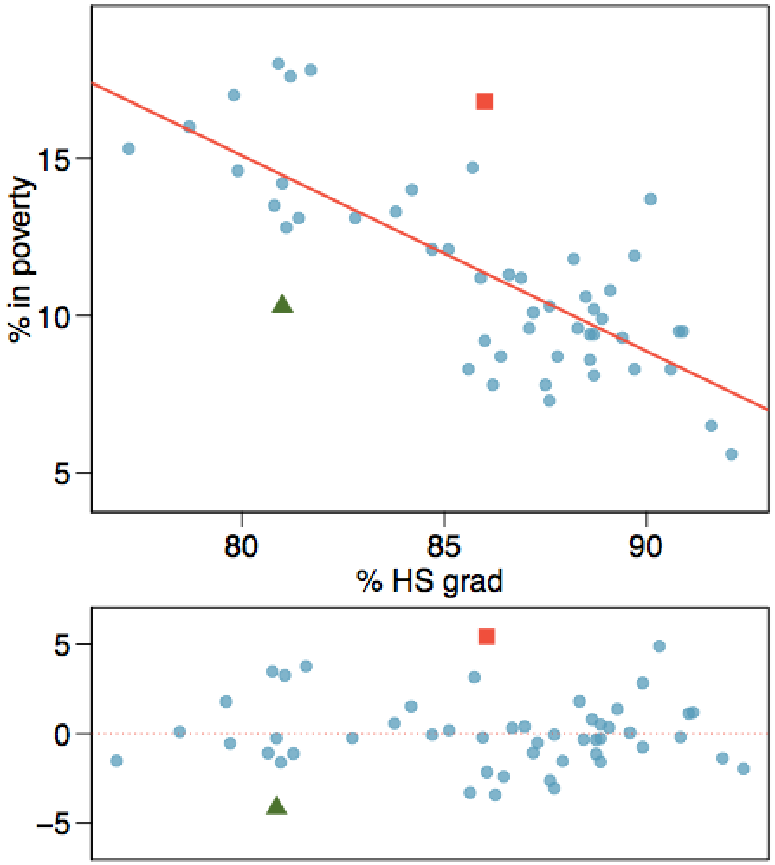
1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class.
  - If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a *residuals plot*.



Anatomy of a residuals plot



▲ RI:

$$\% \ HS\ grad = 81 \qquad \% \ in\ poverty = 10.3$$

$$\% \ \widehat{in\ poverty} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \ in\ poverty - \% \ \widehat{in\ poverty}$$

$$= 10.3 - 14.46 = -4.16$$

■ DC:

$$\% \ HS\ grad = 86 \qquad \% \ in\ poverty = 16.8$$

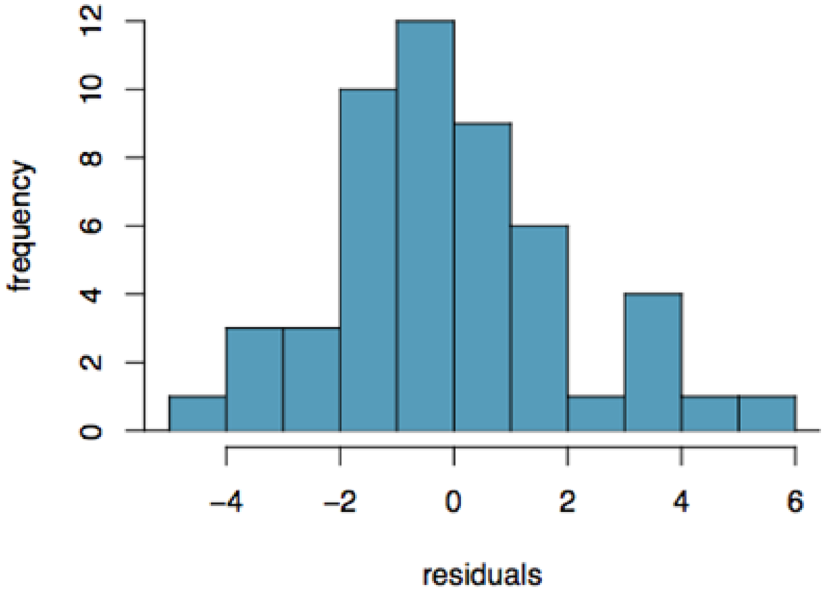
$$\% \ \widehat{in\ poverty} = 64.68 - 0.62 * 86 = 11.36$$

$$e = \% \ in\ poverty - \% \ \widehat{in\ poverty}$$

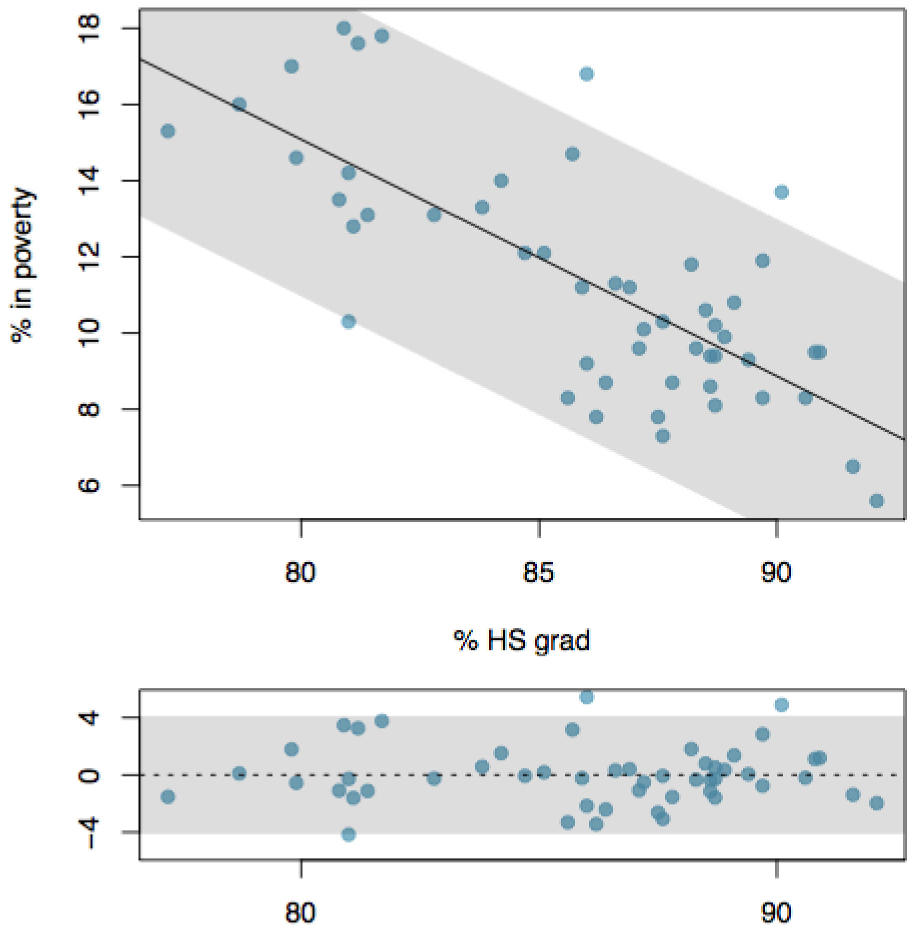
$$= 16.8 - 11.36 = 5.44$$

Conditions: (2) Nearly normal residuals

- 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.
- Check using a histogram or normal probability plot of residuals.



Conditions: (3) Constant variability



- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a histogram or normal probability plot of residuals.

Checking conditions

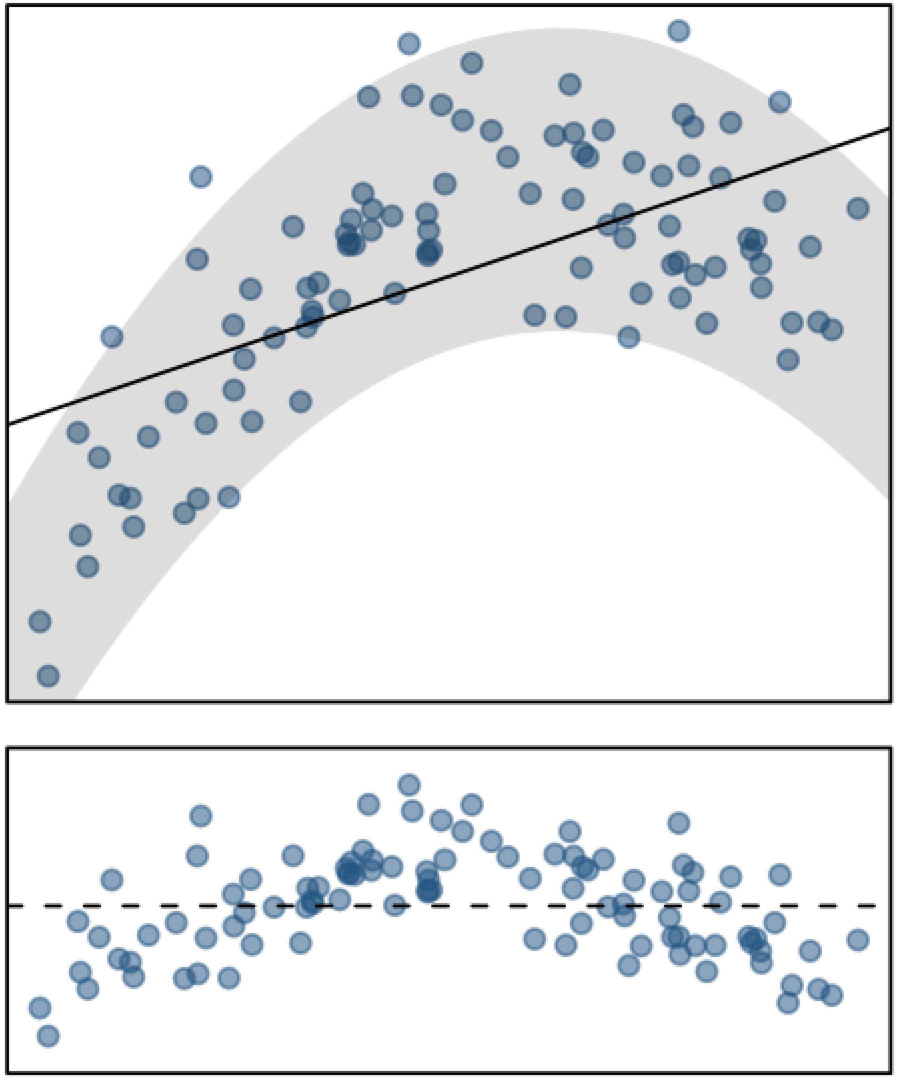
What condition is this linear model obviously violating?

(a) Constant variability

[✓ (b) Linear relationship]

(c) Normal residuals

(d) No extreme outliers



Checking conditions

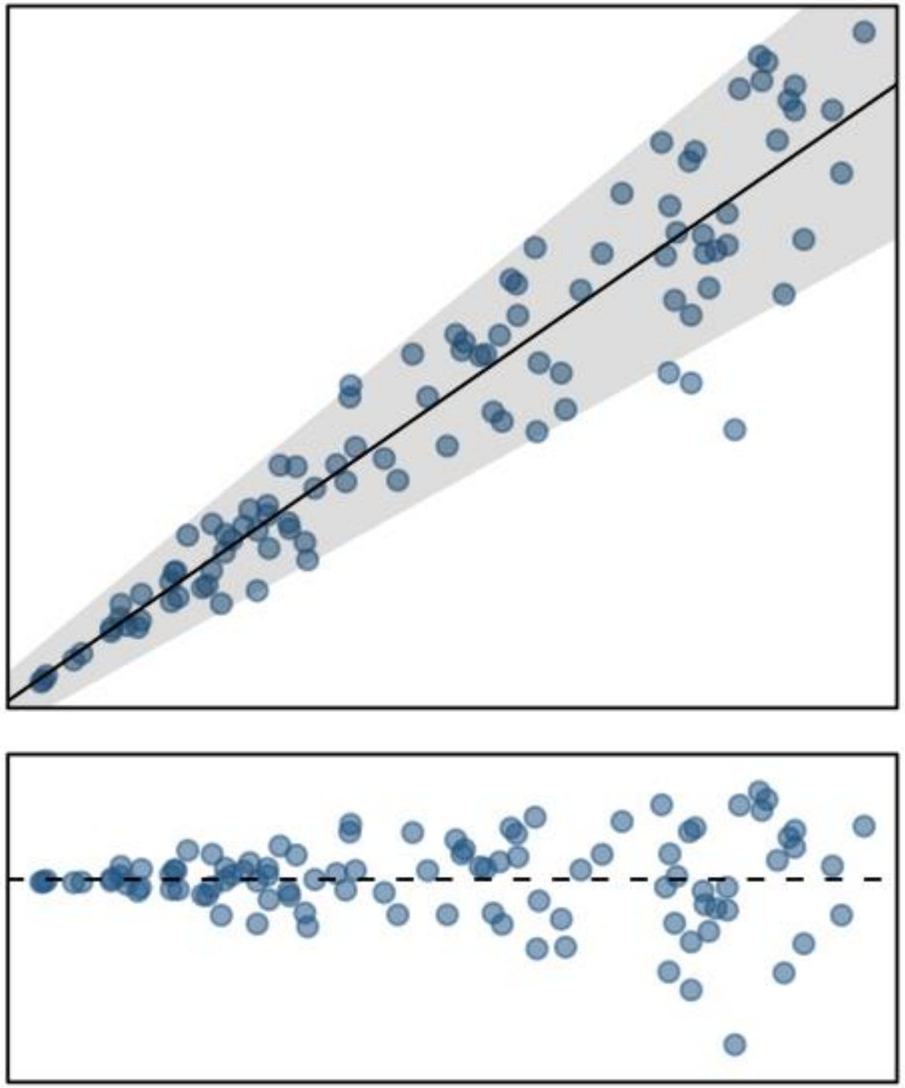
What condition is this linear model obviously violating?

(a) Constant variability ←

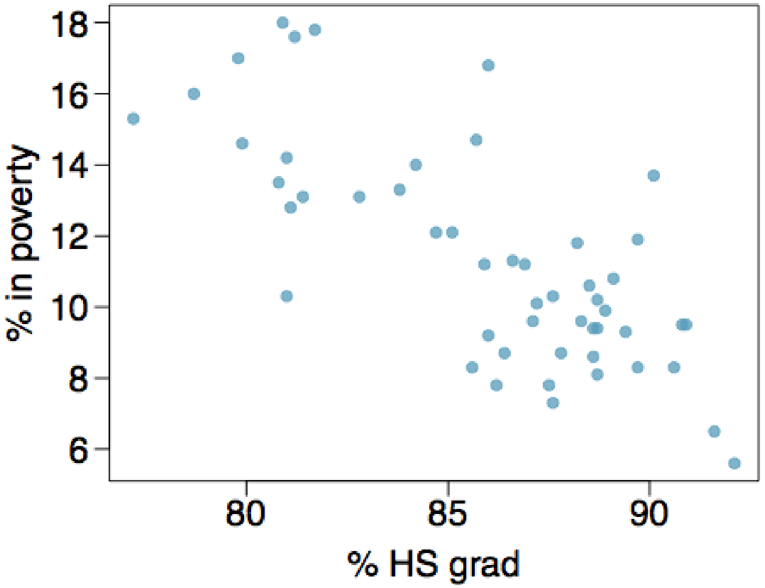
(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Given…



	% 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

The slope of the regression can be calculated as:

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

In context...

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

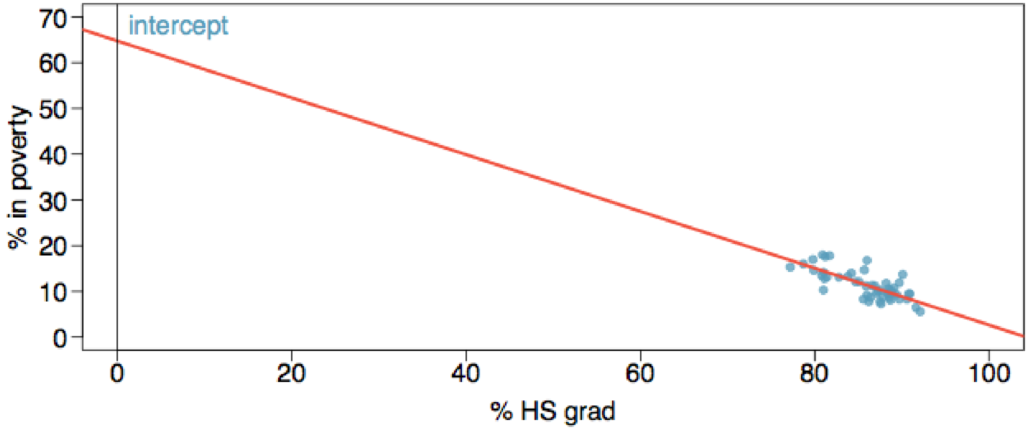
Interpretation

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

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact that a regression line always passes through $(\bar{x}, \bar{y})$.

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



$$\begin{aligned} b_0 &= 11.35 - (-0.62) \times 86.01 \\ &= 64.68 \end{aligned}$$

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%.

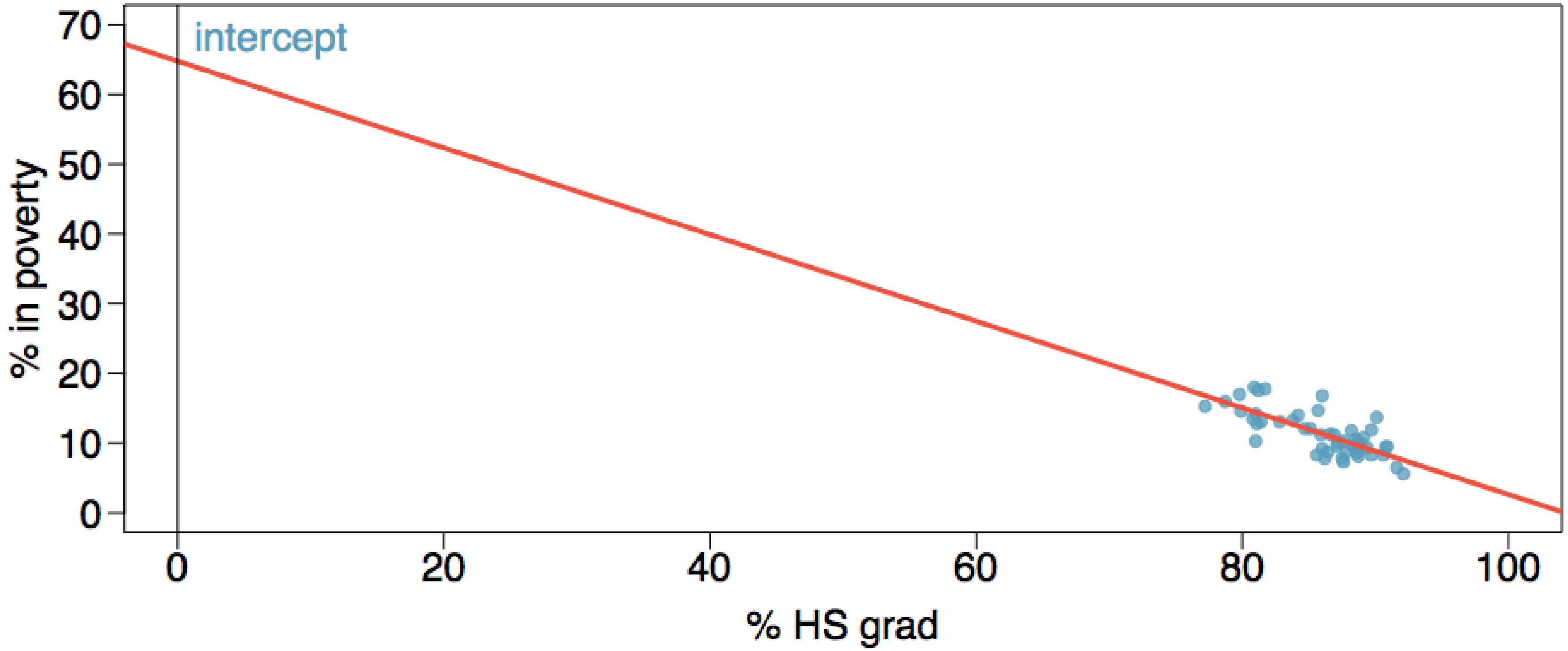
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

[✓ (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%.

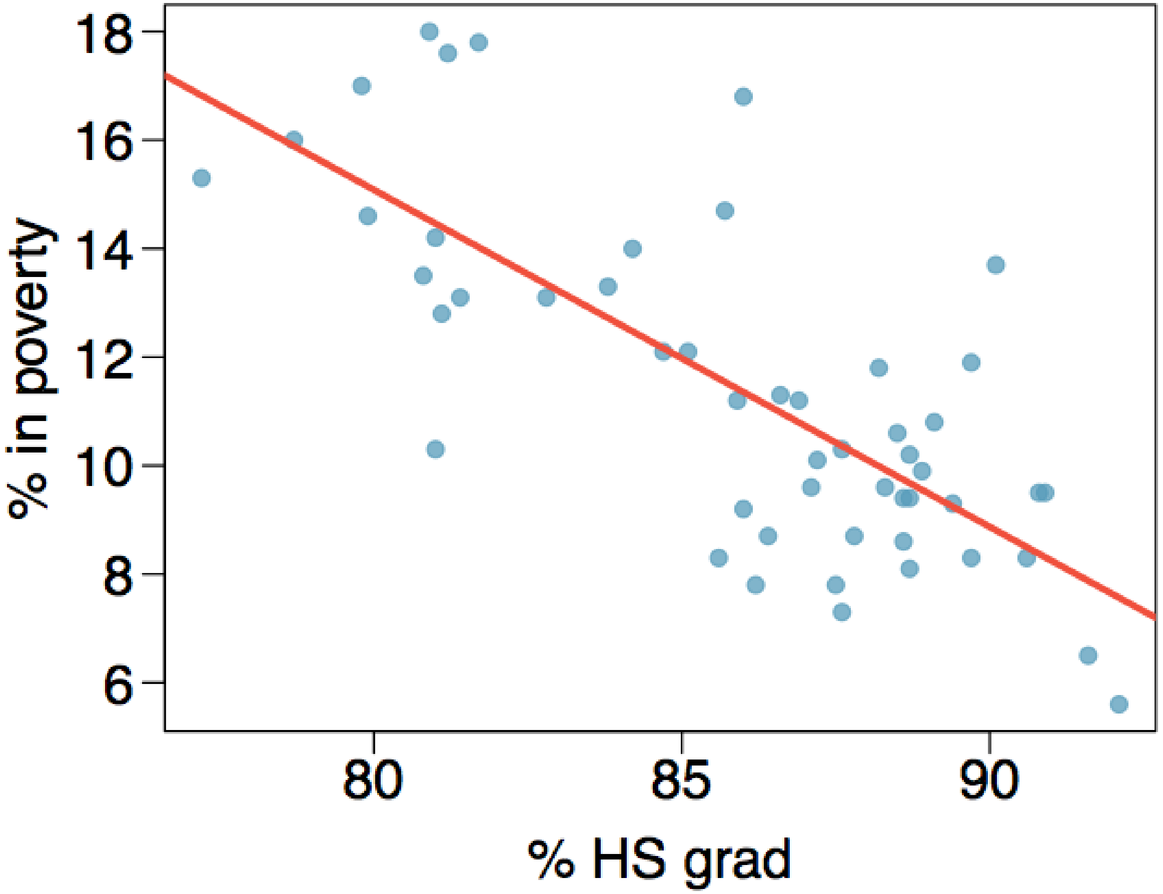
More on the intercept

- Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.



Regression line

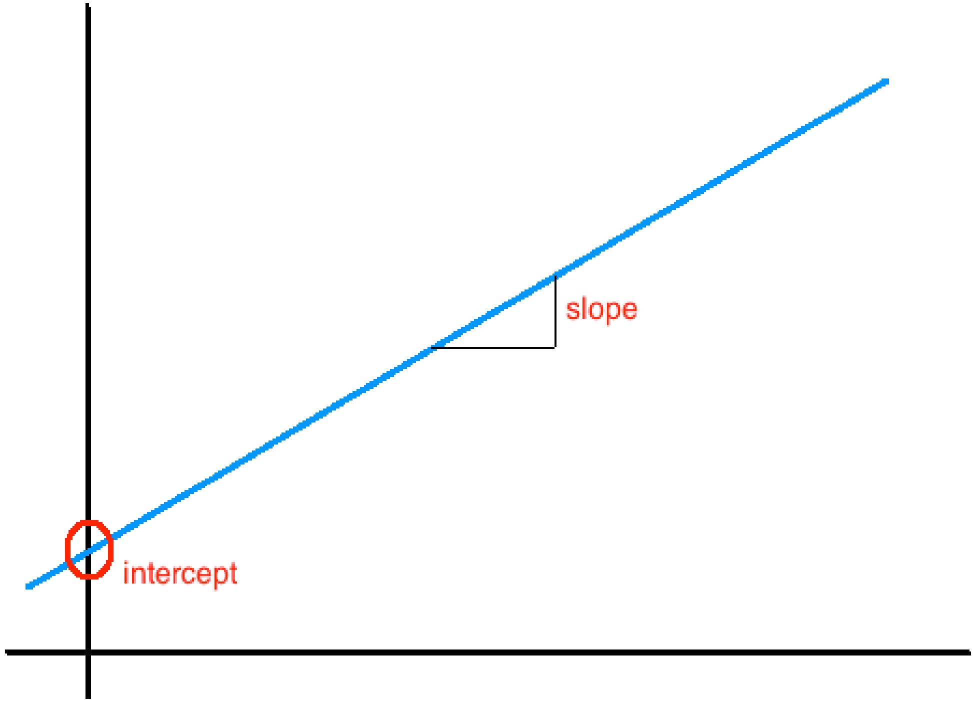
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretation of slope and intercept

- *Intercept*: When $x = 0$, y is expected to equal the intercept.

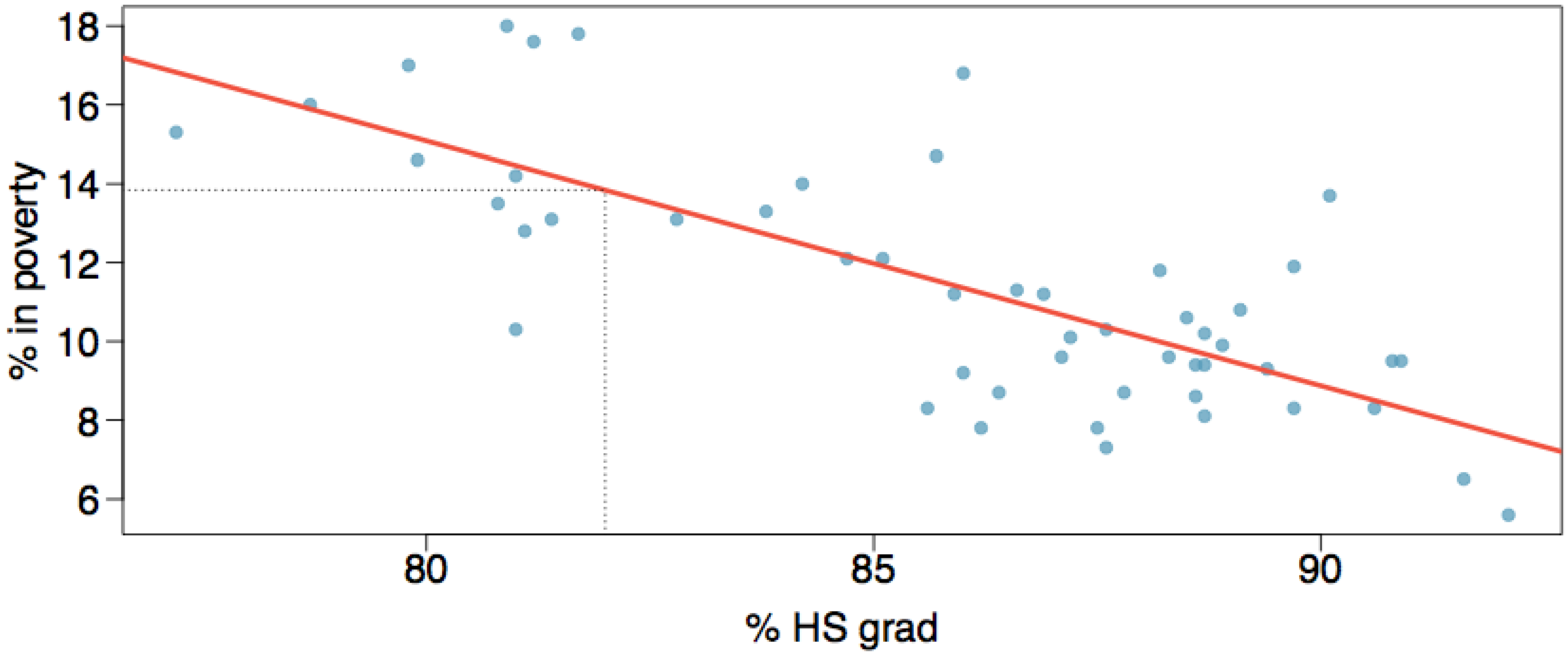
- *Slope*: For each unit in x, y is expected to increase / decrease on average by the slope.



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

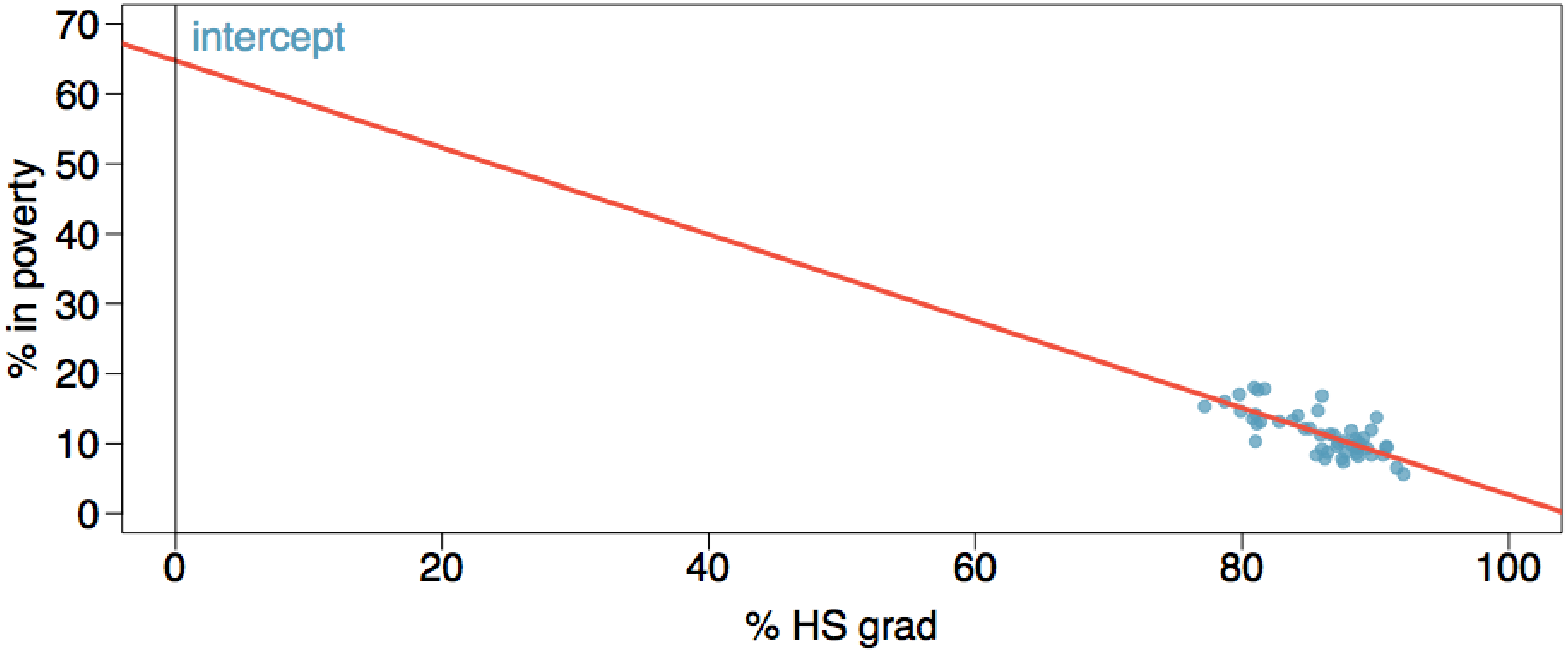
Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the *predicted value*.

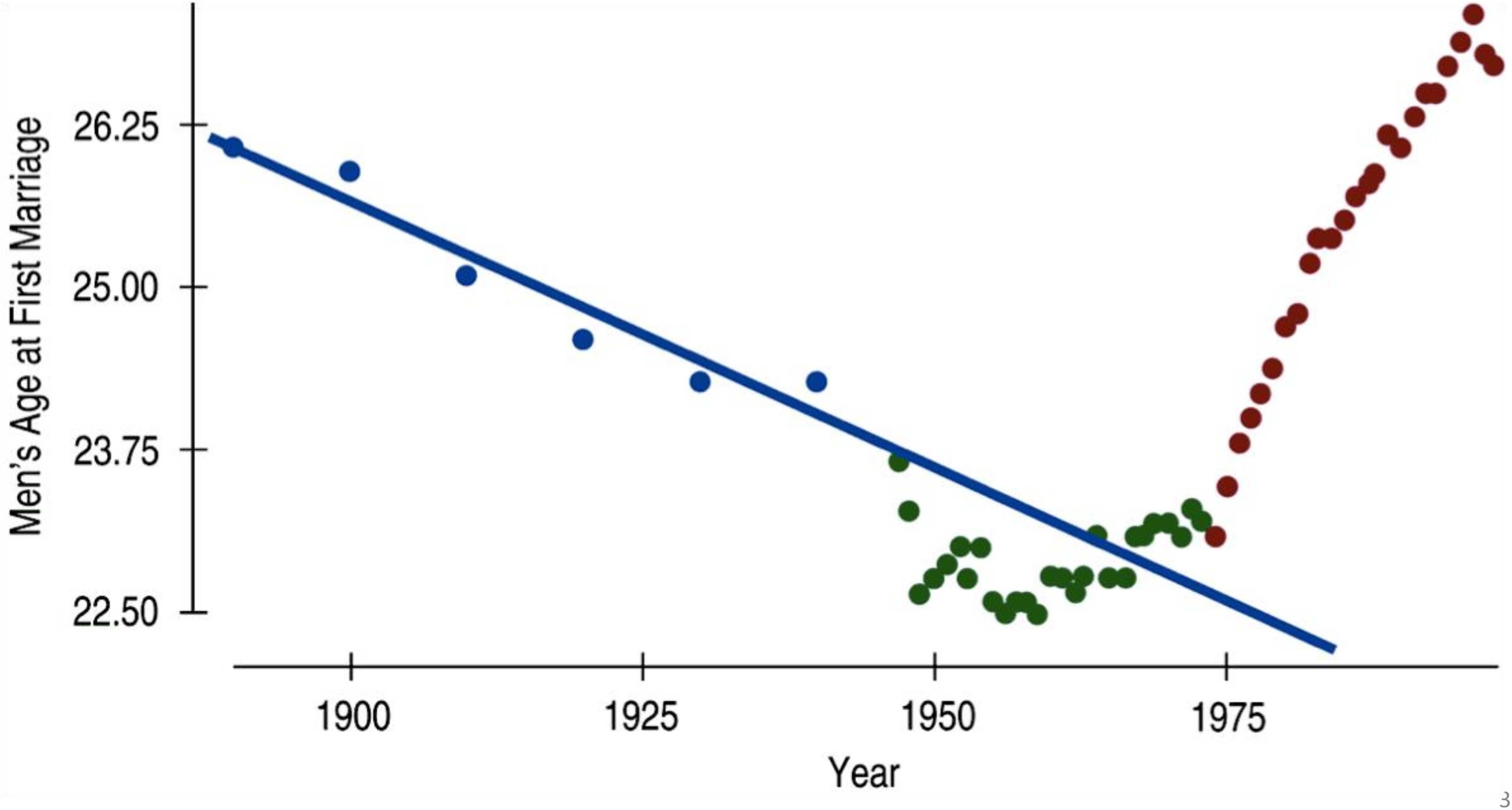


Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = -0.62^2 = 0.38$.

Interpretation of R^2

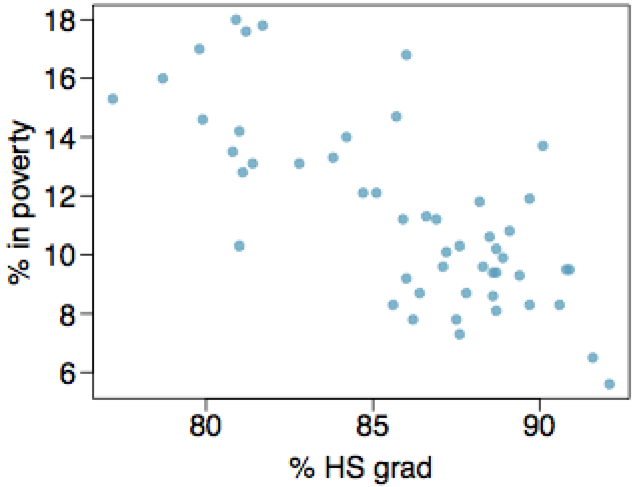
- Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

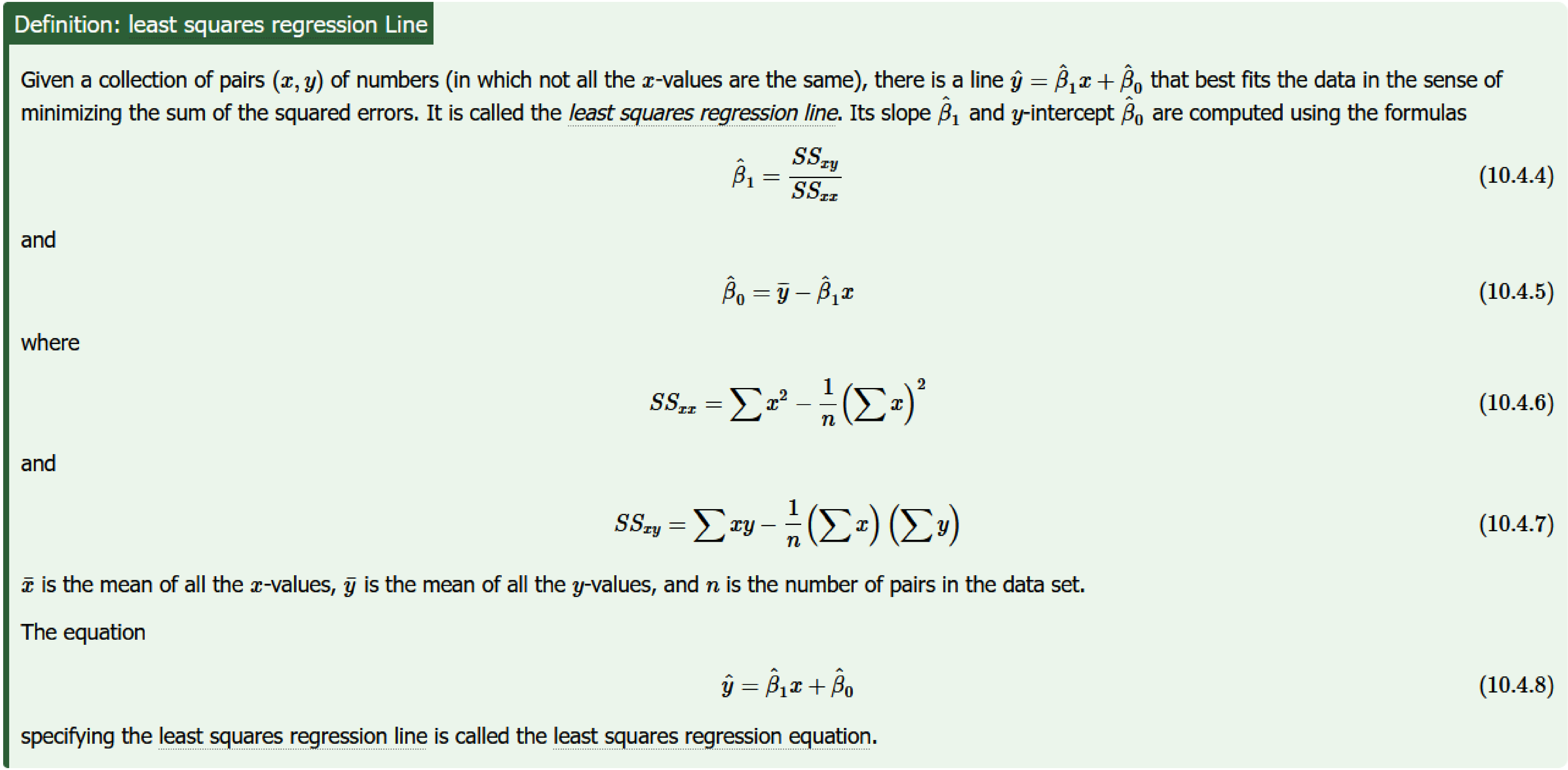
(a) 38% of the variability in the % of HG graduates among the 51 states is explained by the model.

(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.





Definition: least squares regression Line

Given a collection of pairs (x, y) of numbers (in which not all the x-values are the same), there is a line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ that best fits the data in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Its slope $\hat{\beta}_1$ and y-intercept $\hat{\beta}_0$ are computed using the formulas

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{10.4.4}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x \tag{10.4.5}$$

where

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 \tag{10.4.6}$$

and

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) \tag{10.4.7}$$

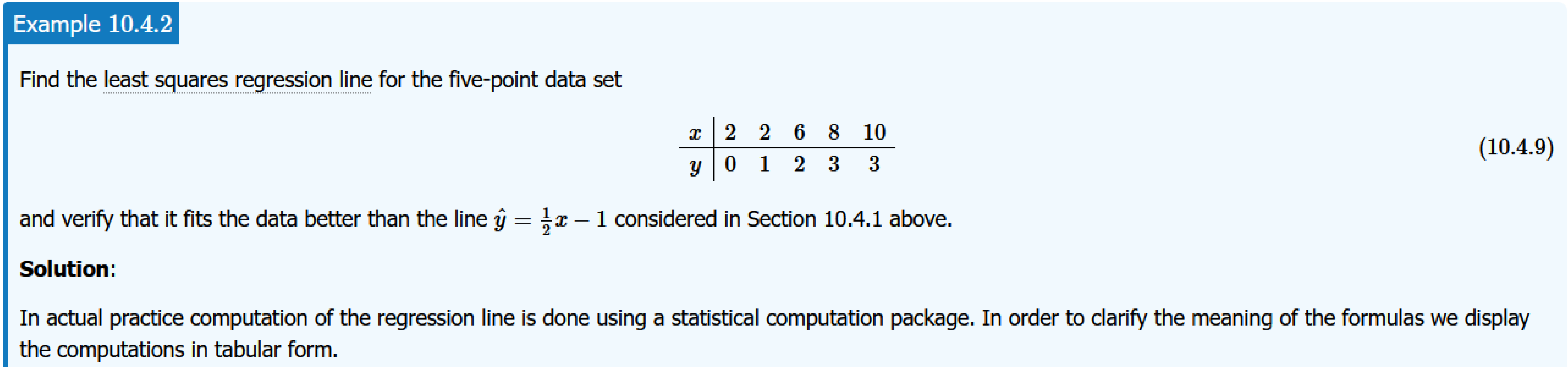
$\bar{x}$ is the mean of all the x-values, $\bar{y}$ is the mean of all the y-values, and n is the number of pairs in the data set.

The equation

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \tag{10.4.8}$$

specifying the least squares regression line is called the least squares regression equation.

Example



Example 10.4.2

Find the least squares regression line for the five-point data set

x	2	2	6	8	10
y	0	1	2	3	3

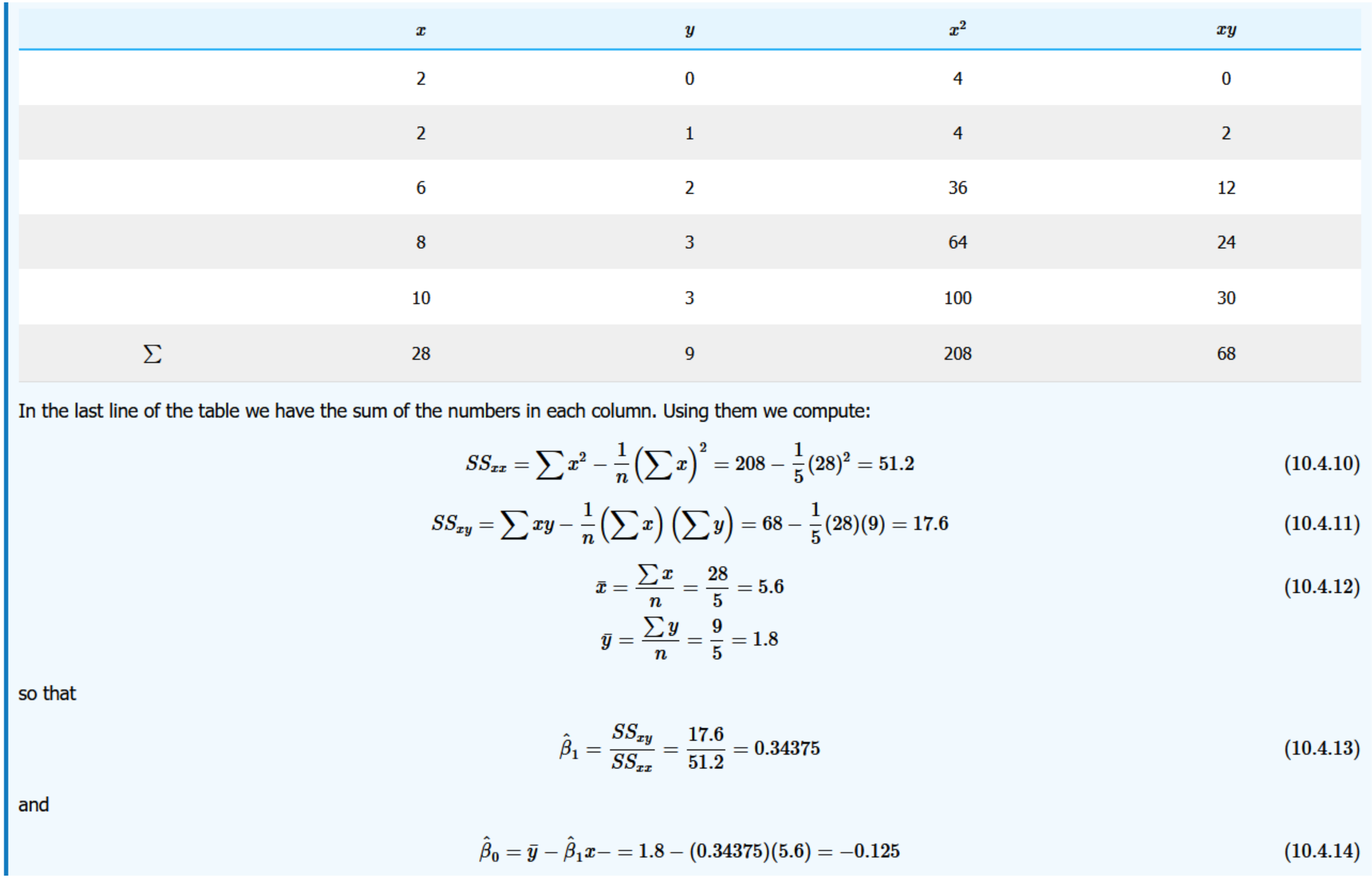
(10.4.9)

and verify that it fits the data better than the line $\hat{y} = \frac{1}{2}x - 1$ considered in Section 10.4.1 above.

Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

Solution



	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
$\sum$	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 = 208 - \frac{1}{5}(28)^2 = 51.2 \tag{10.4.10}$$

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right) = 68 - \frac{1}{5}(28)(9) = 17.6 \tag{10.4.11}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{5} = 5.6 \tag{10.4.12}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \tag{10.4.13}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x - = 1.8 - (0.34375)(5.6) = -0.125 \tag{10.4.14}$$

Finally

- The least squares regression line for these data is:

$$\hat{y} = 0.34375x \ - 0.12$$

Practice

- Try example 10.4.3
  - Source: https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/10%3A_Correlation_and_Regression/10.04%3A_The_Least_Squares_Regression_Line

Other Examples

- https://www.technologynetworks.com/informatics/articles/calculating-a-least-squares-regression-line-equation-example-explanation-310265

Sources

- openintro.org/os (Chapter 8, Section 8.1, 8.2)

Helpful Links:

- https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/10%3A_Correlation_and_Regression/10.04%3A_The_Least_Squares_Regression_Line