Chapter 8

Introduction to linear regression[1]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Fitting a line by least squares regression

An objective measure for finding the best line

- We want a line that has small residuals:

An objective measure for finding the best line

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

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

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
  $$|e_1| + |e_2| + \ \ + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – **least squares**
  $$e_1^2 + e_2^2 + \ \ + e_n^2$$

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
  $$|e_1| + |e_2| + \ \ + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – **least squares**
  $$e_1^2 + e_2^2 + \ \ + e_n^2$$
- Why least squares?

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
  $$|e_1| + |e_2| + \ \ + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – **least squares**
  $$e_1^2 + e_2^2 + \ \ + e_n^2$$
- Why least squares?
  1. Most commonly used

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
  $$|e_1| + |e_2| + \ \ + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – **least squares**
  $$e_1^2 + e_2^2 + \ \ + e_n^2$$
- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software

An objective measure for finding the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
  $$|e_1| + |e_2| + \ \ + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – **least squares**
  $$e_1^2 + e_2^2 + \ \ + e_n^2$$
- 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

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

- $\hat{y}$: Predicted value of the response variable, y.
- β_0 : Intercept, parameter.
  - b_0 : Intercept, point estimate.
- β_1 : Slope, parameter
  - b_1 : Slope, point estimate
- x : Explanatory variable

Conditions for the least squares line

1. Linearity.

Conditions for the least squares line

1. Linearity.
2. Nearly normal residuals.

Conditions for the least squares line

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

Conditions: Linearity

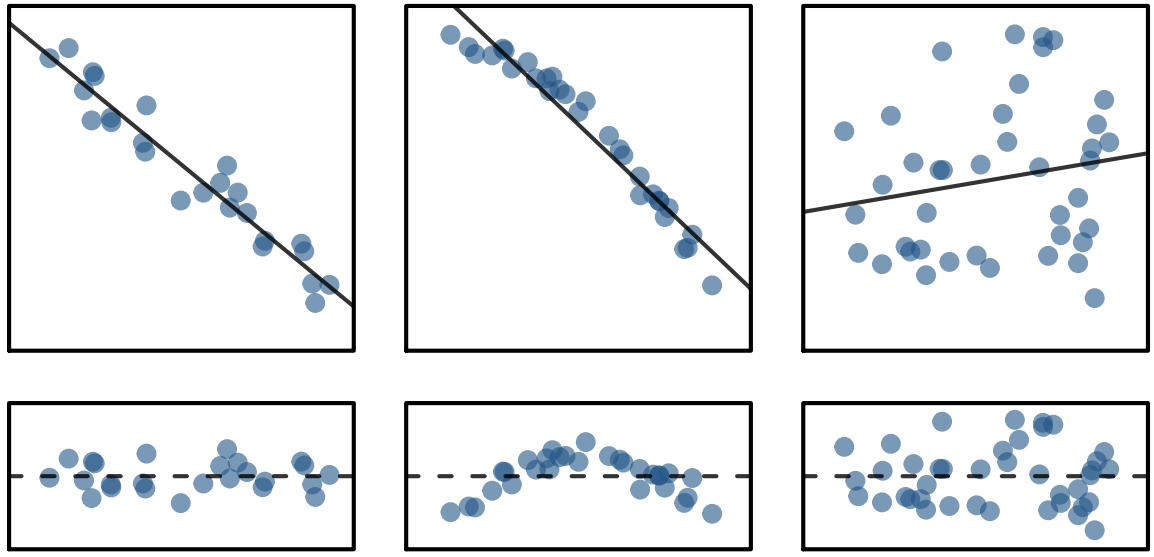
- The relationship between the explanatory and the response variable should be linear.

Conditions: 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.

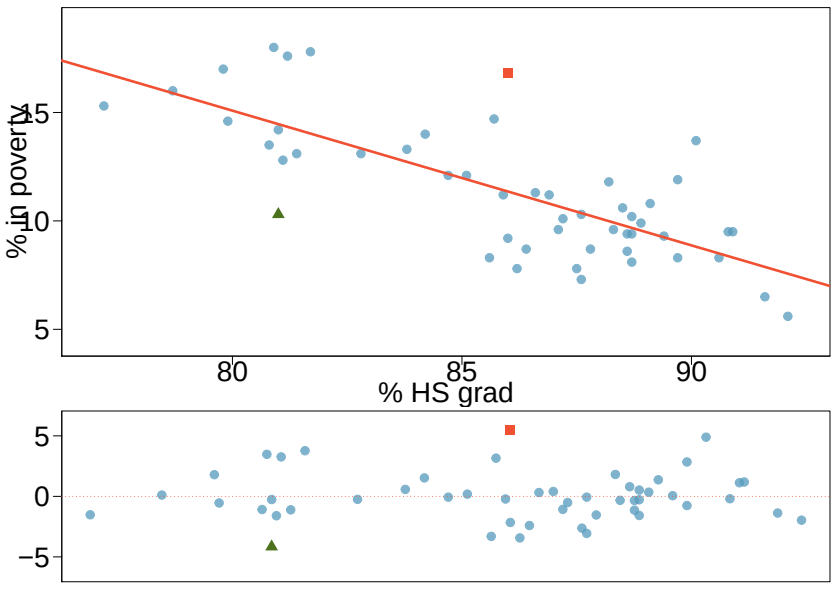
Conditions: Linearity

- 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:**



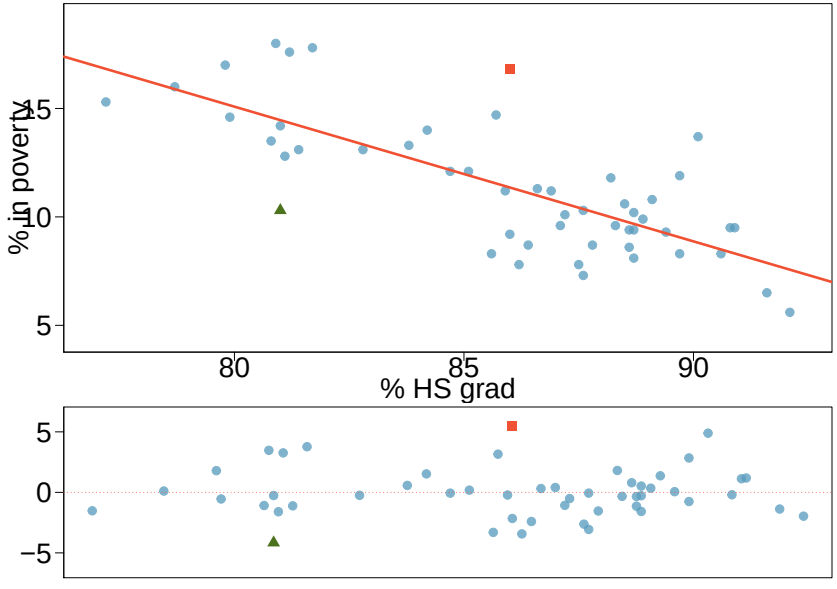
$$\% \; HS \; grad = 81 \qquad \% \; in \; poverty = 10.3$$

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

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

$$= 10.3 - 14.46 = -4.16$$

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: Nearly normal residuals

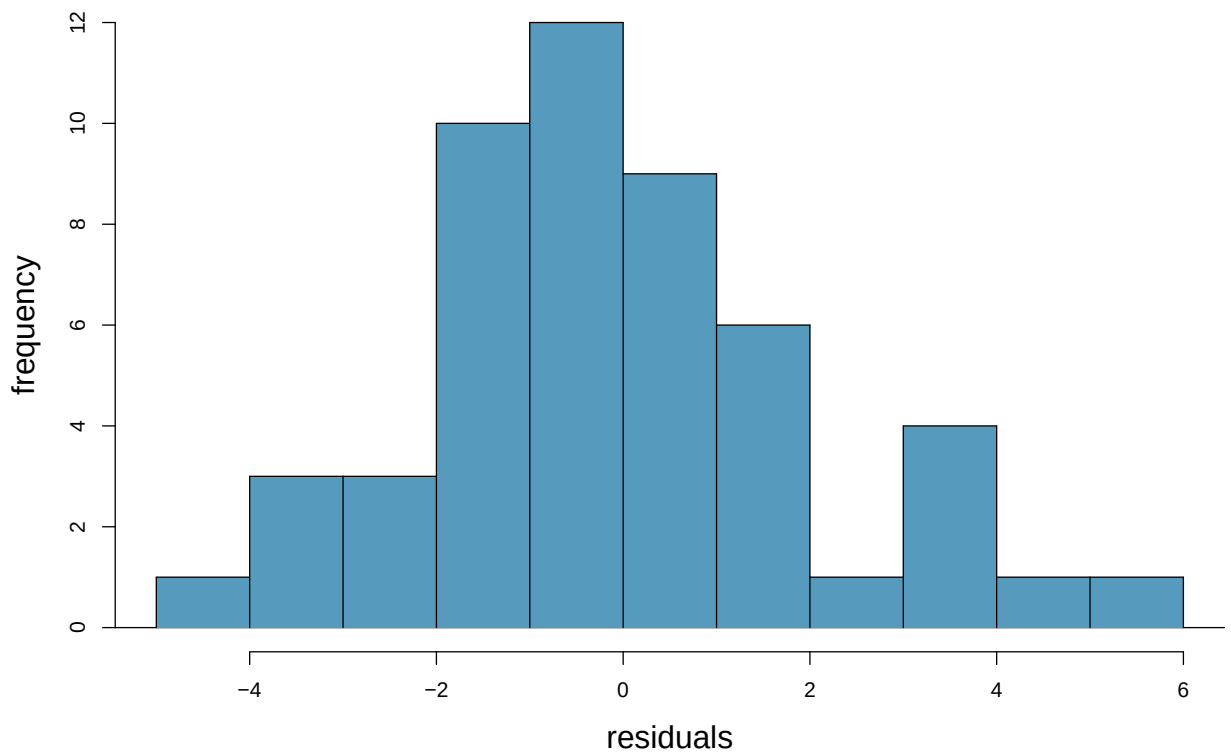
- The residuals should be nearly normal.

Conditions: Nearly normal residuals

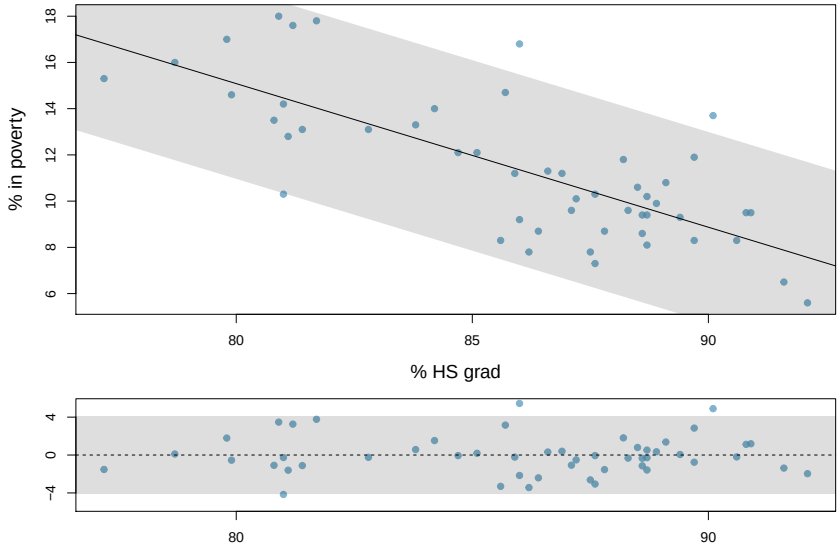
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

Conditions: 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.

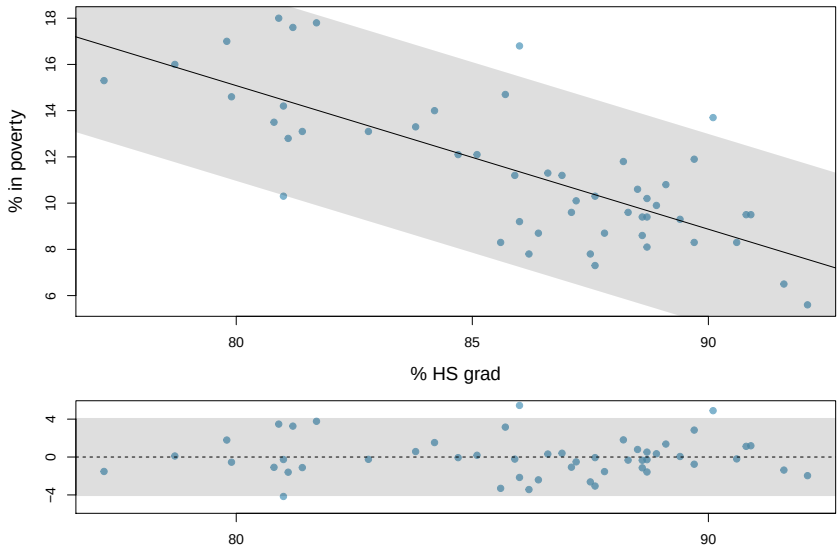


Conditions: Constant variability



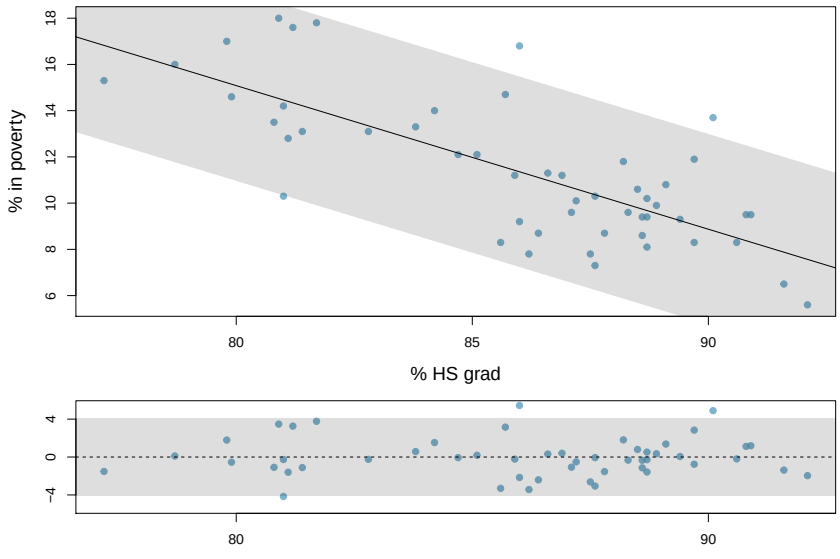
- The variability of points around the least squares line should be roughly constant.

Conditions: Constant variability



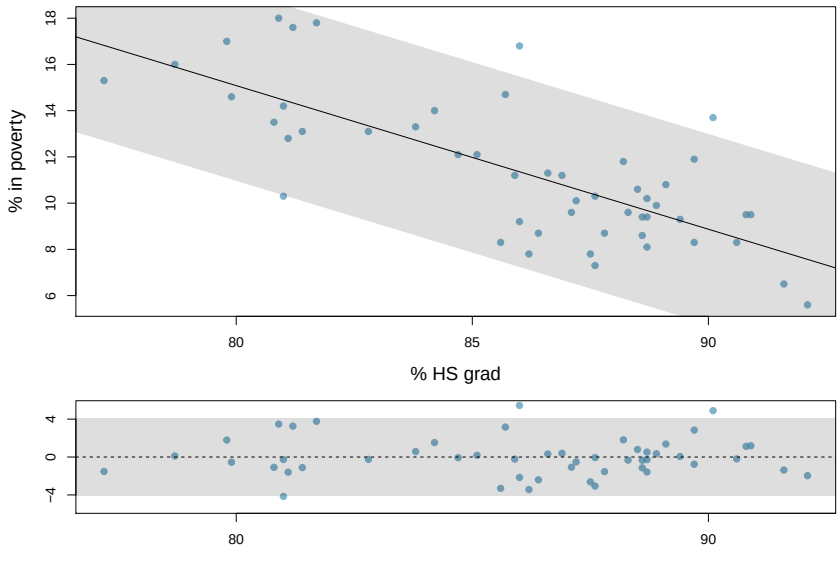
- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.

Conditions: 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**.

Conditions: Constant variability

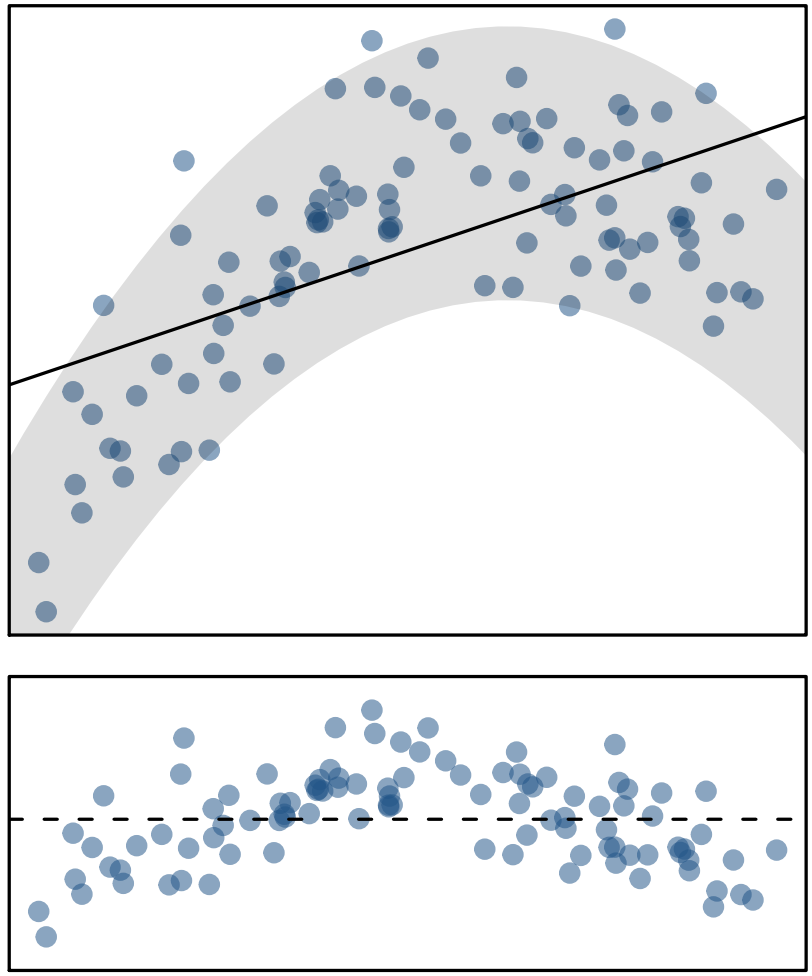


- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called **homoscedasticity**.
- Check using a residuals plot.

Checking conditions

What condition is this linear model obviously violating?

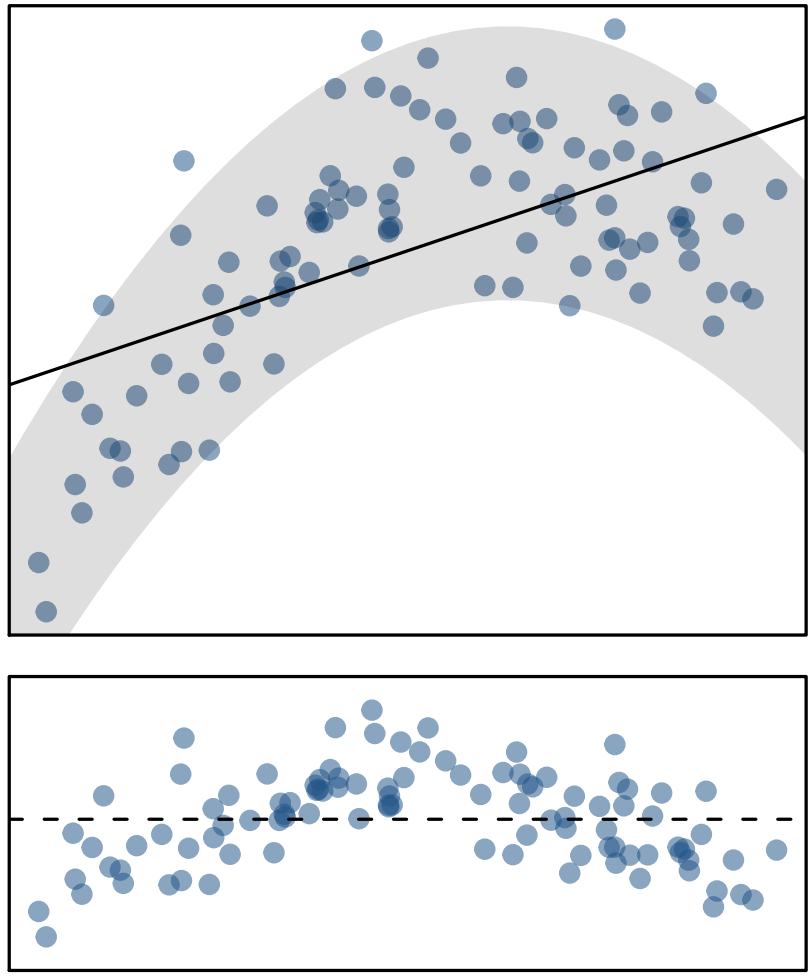
A) Constant variability
B) Linear relationship
C) Normal residuals
D) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

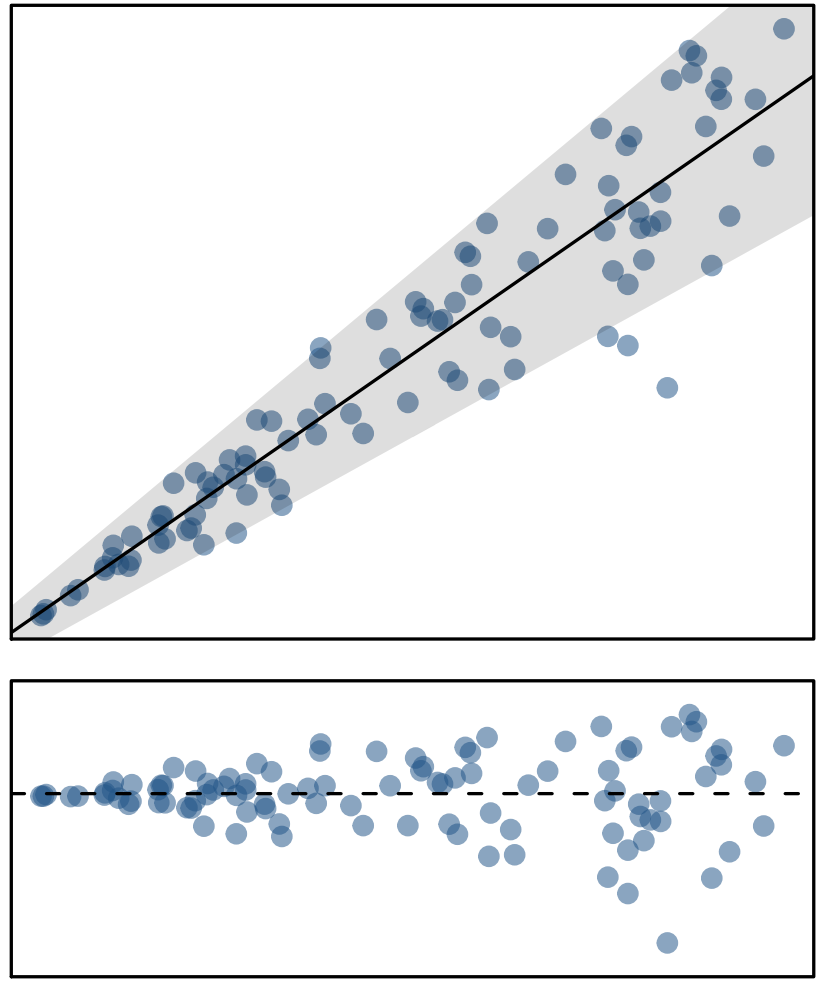
A) Constant variability
B) Linear relationship
C) Normal residuals
D) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

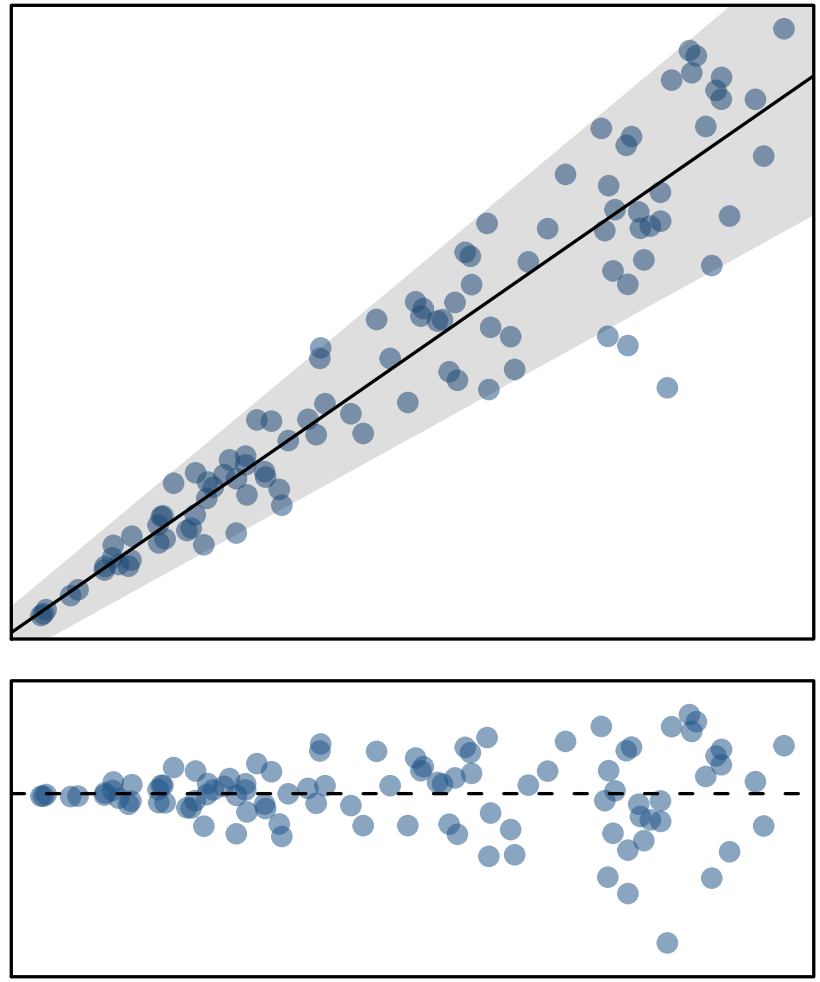
A) Constant variability
B) Linear relationship
C) Normal residuals
D) No extreme outliers



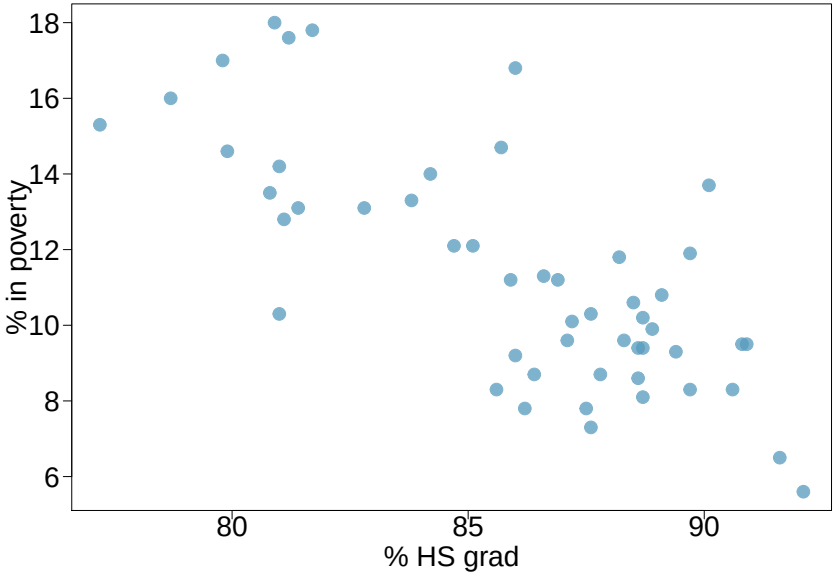
Checking conditions

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

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

Slope

The slope of the regression can be calculated as

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

In context...

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

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

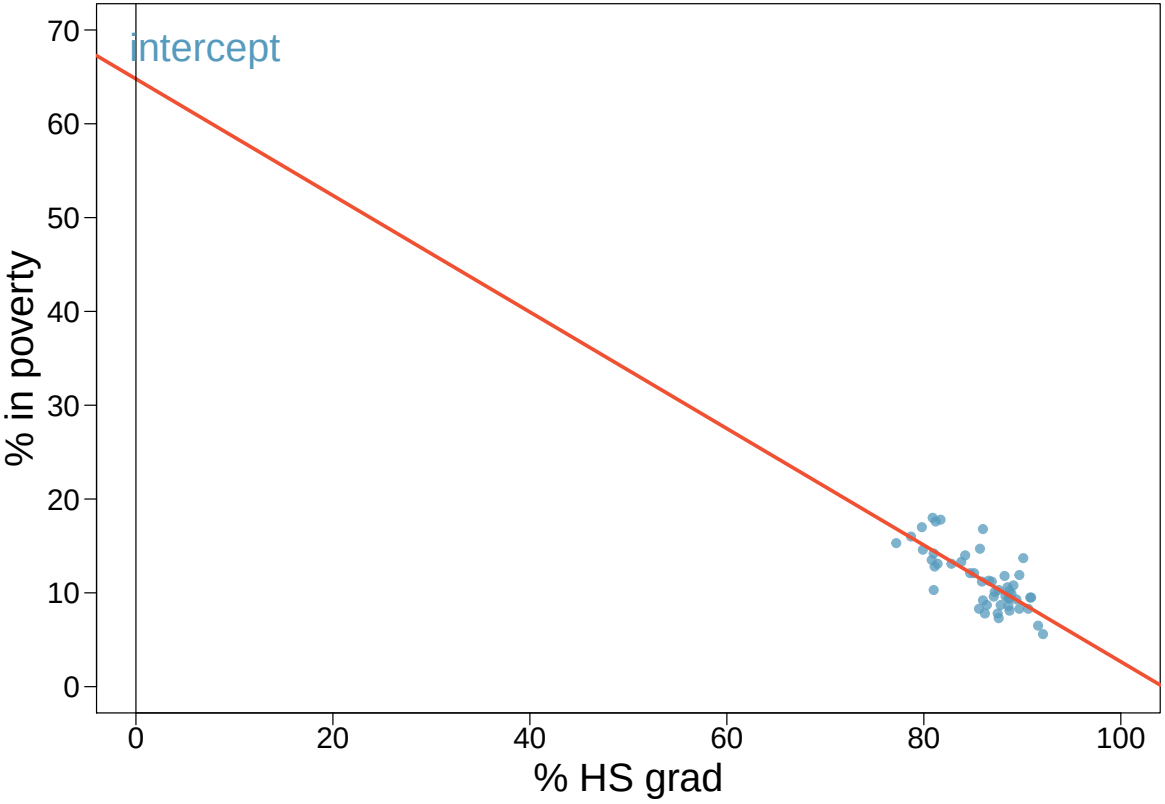
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

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

Intercept

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

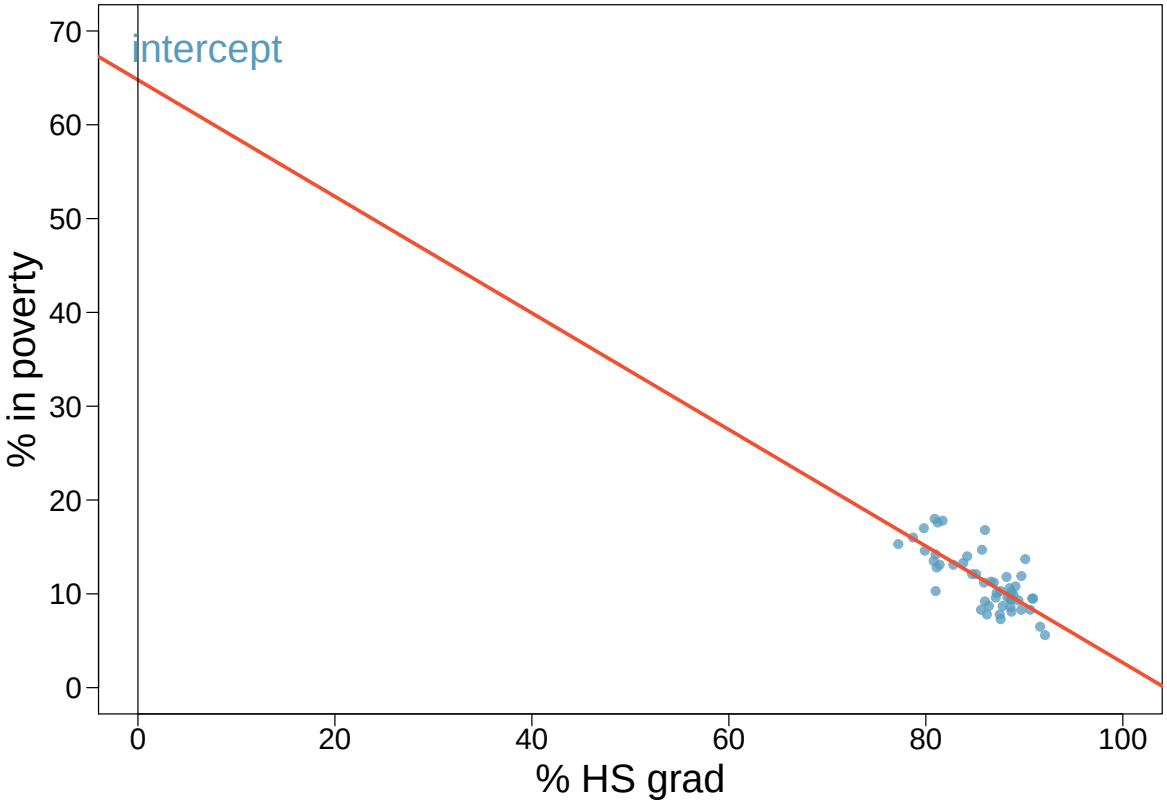
$$b_0 = \bar{y} - b_1 \bar{x}$$



Intercept

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

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



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

Practice

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%.

Practice

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%.

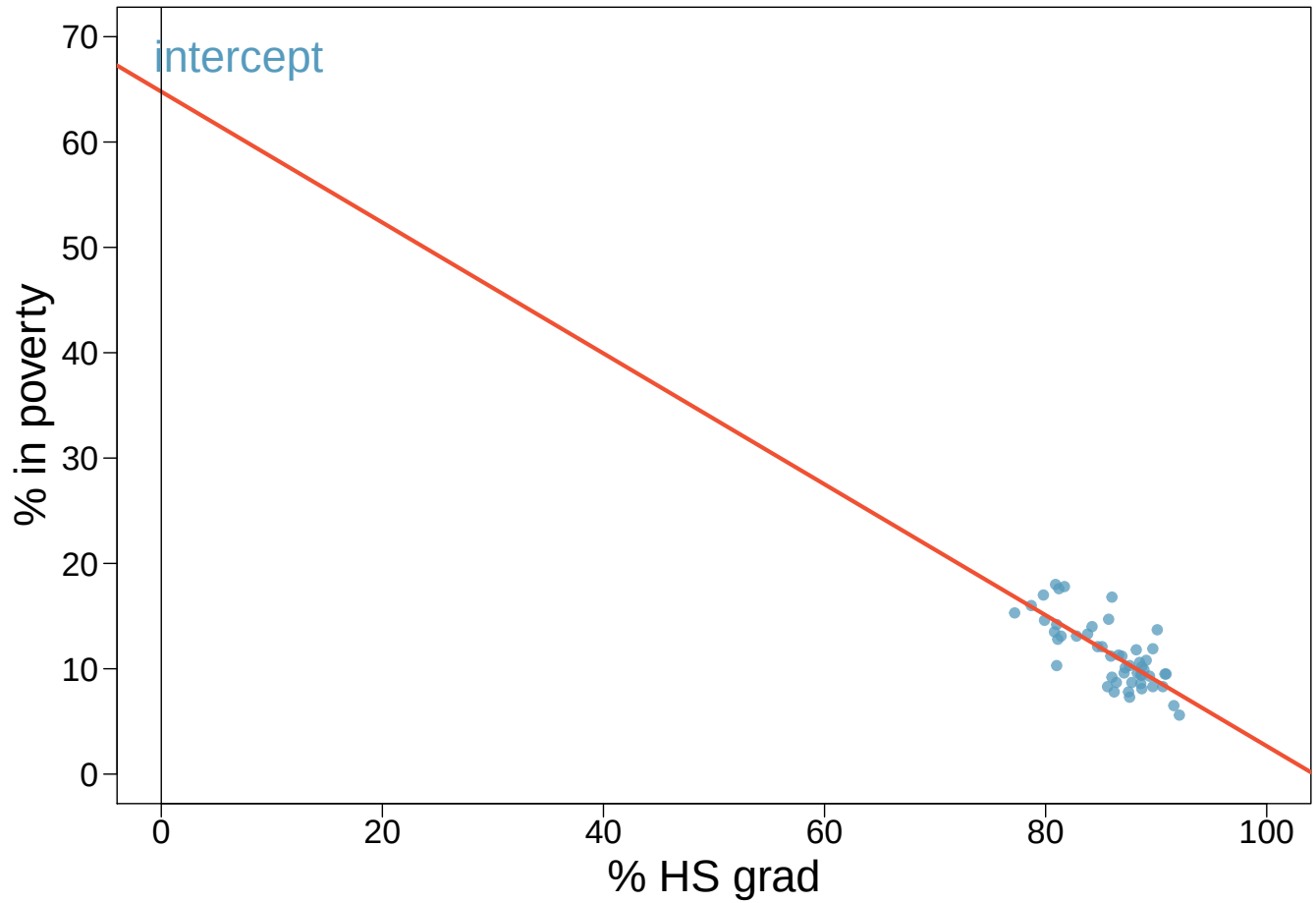
C) Having no HS graduates leads to 64.68% of residents living below the poverty line.

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%.

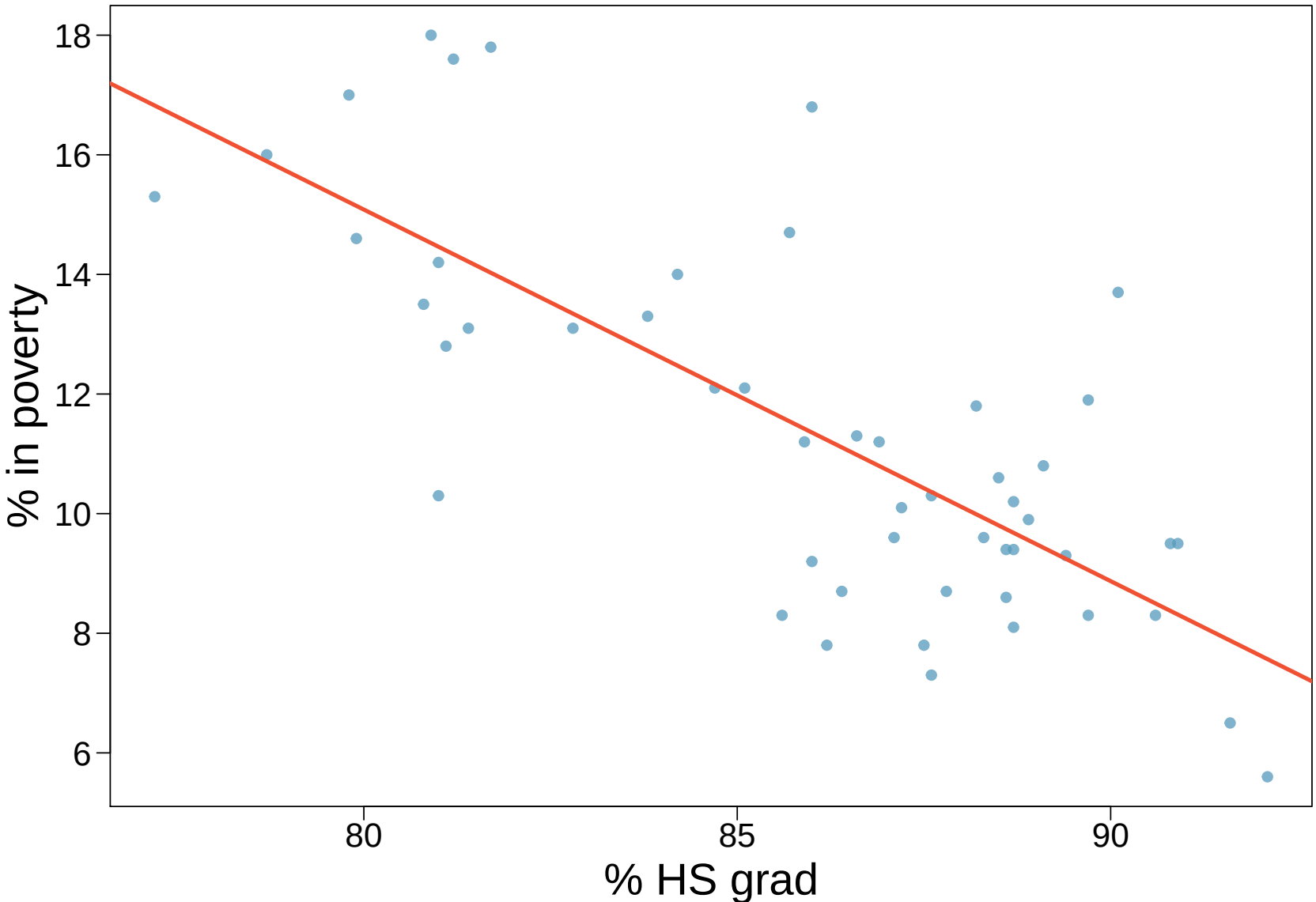
More on the intercept

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.



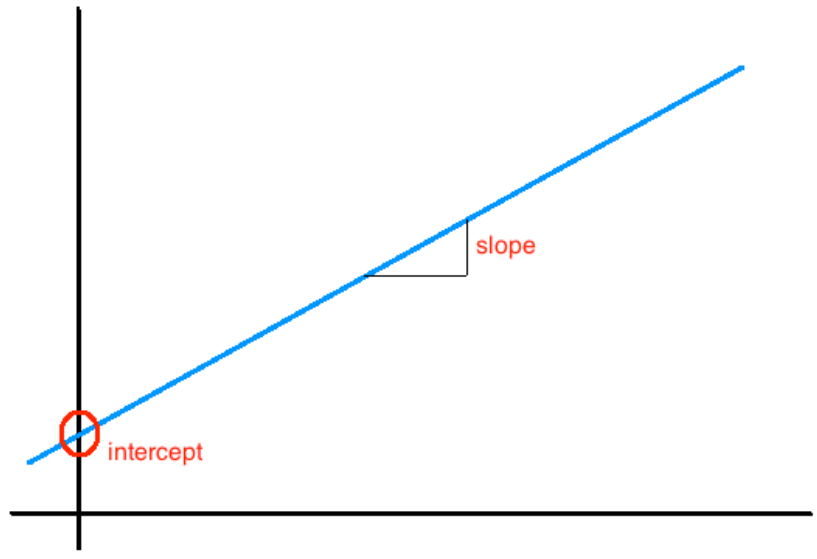
Regression line

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



Interpretation of slope and intercept

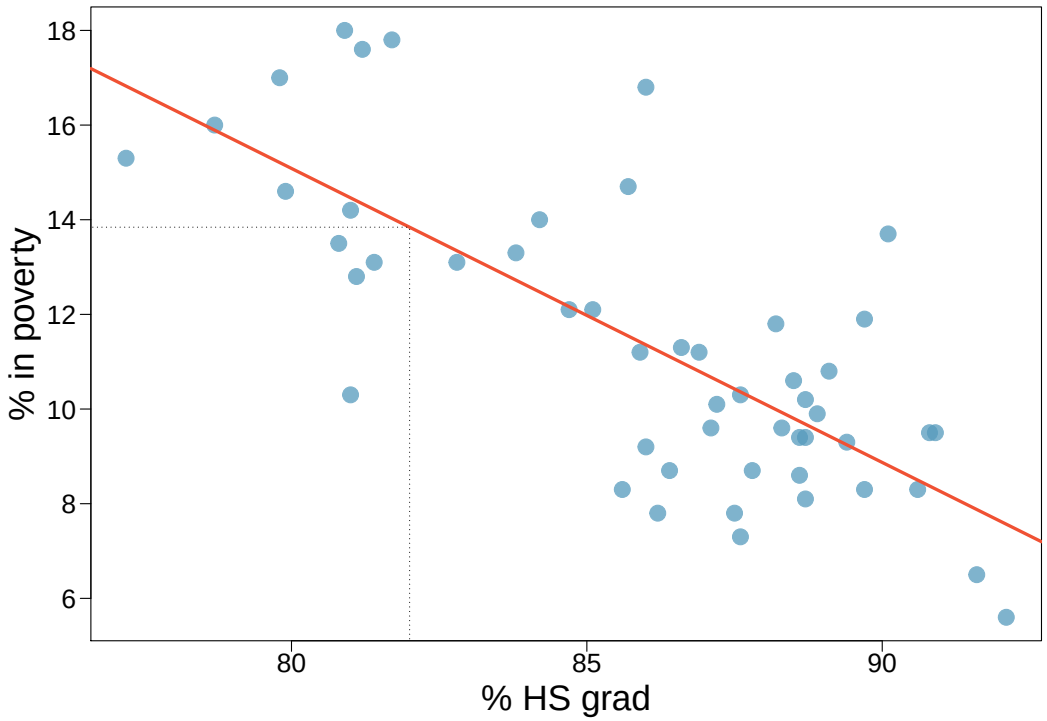
- **Intercept:** When $x = 0$, y is expected to equal the intercept.
- **Slope:** For each unit in x, y is expected to increase / decrease on average by the slope.



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

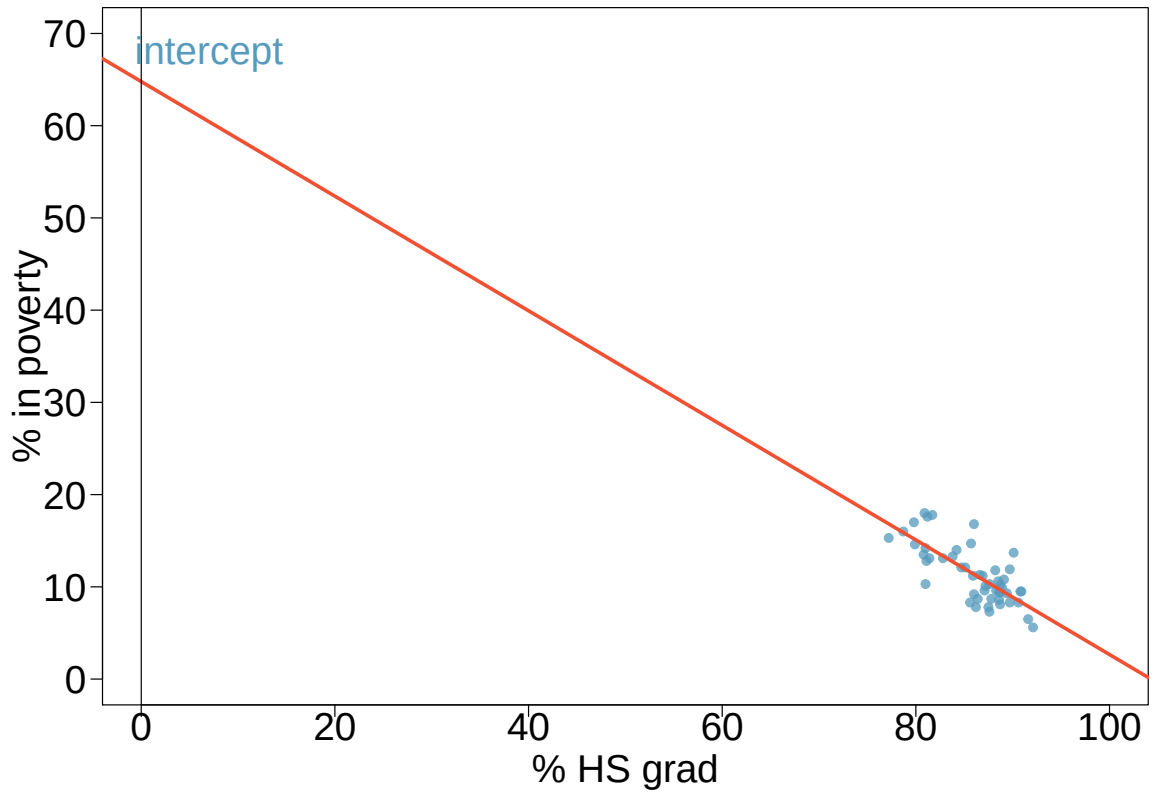
Prediction

- 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.

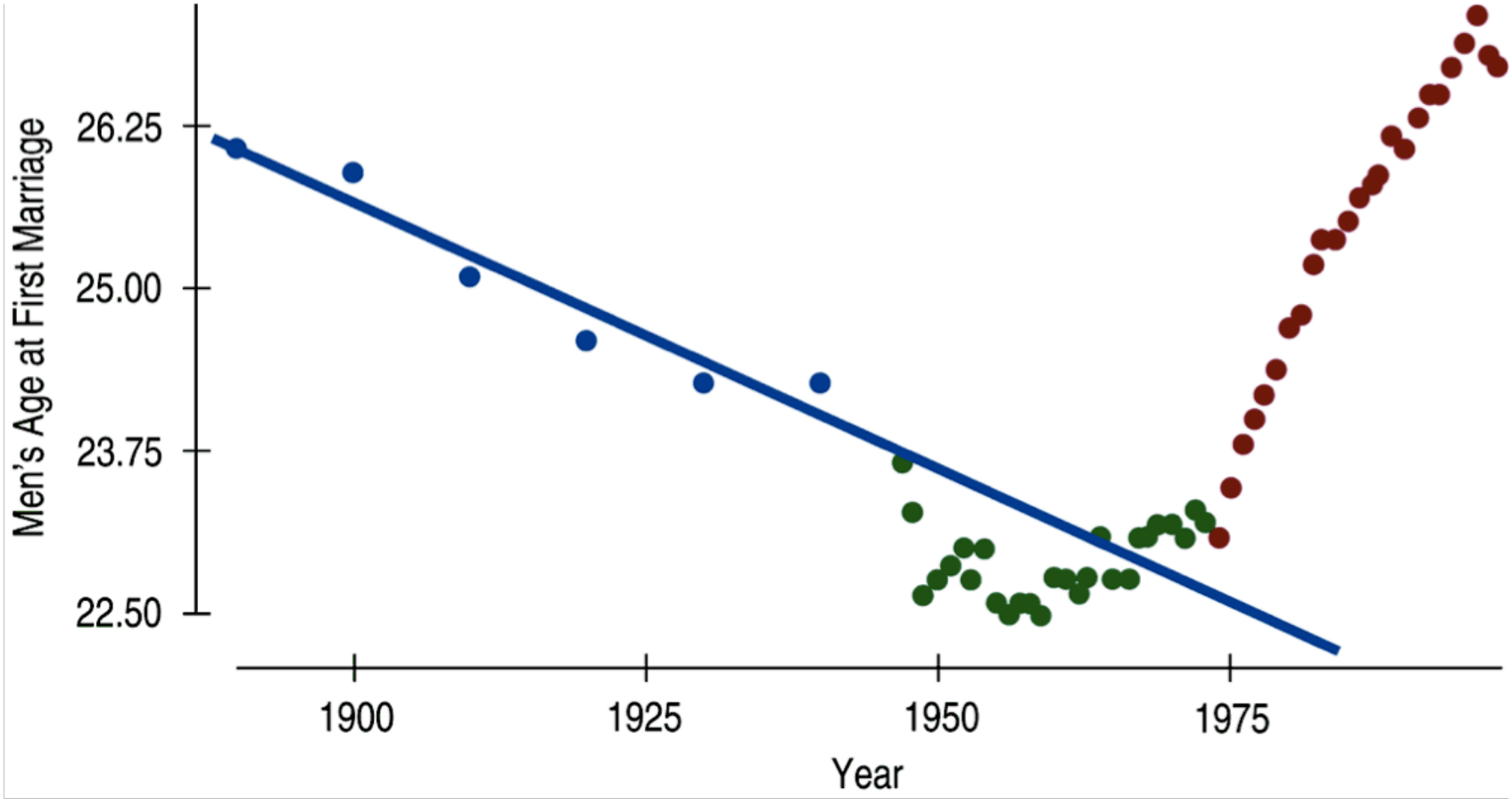


Extrapolation

- 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



Example of extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page

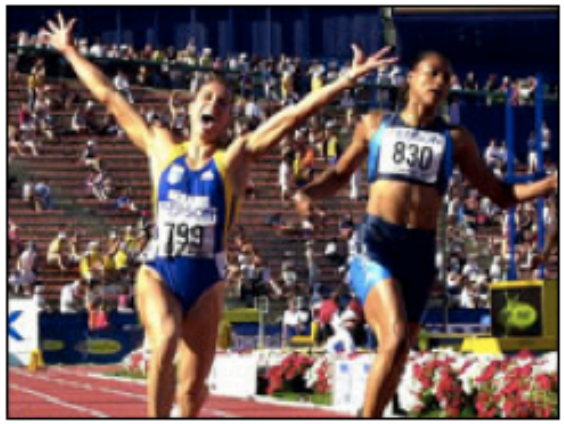
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

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."

Example of extrapolation

Momentous sprint at the 2156 Olympics?

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

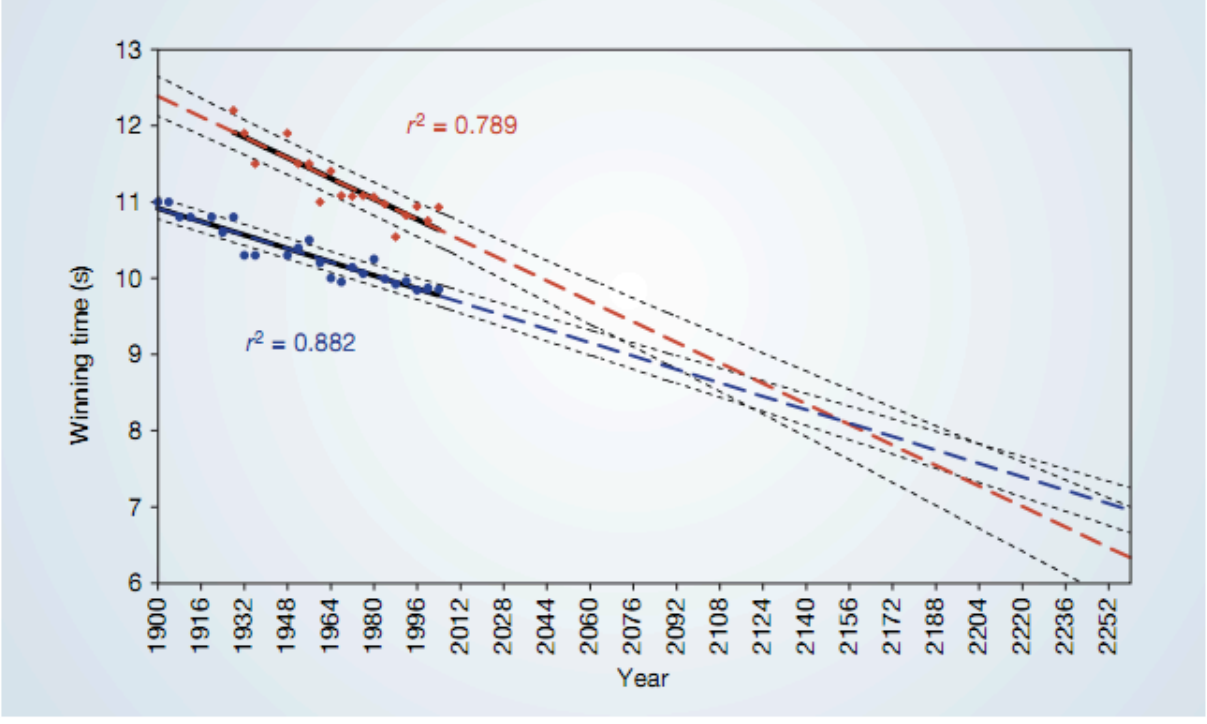


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) 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.

R^2

- The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.

R^2

- The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.
- R^2 is calculated as the square of the correlation coefficient.

R^2

- The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.
- 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.

R^2

- The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.
- 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.

R^2

- The strength of the fit of a linear model is most commonly evaluated using $\mathbf{R^2}$.
- 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, $R^2 = (-0.75)^2 = 0.56$.

Interpretation of R^2

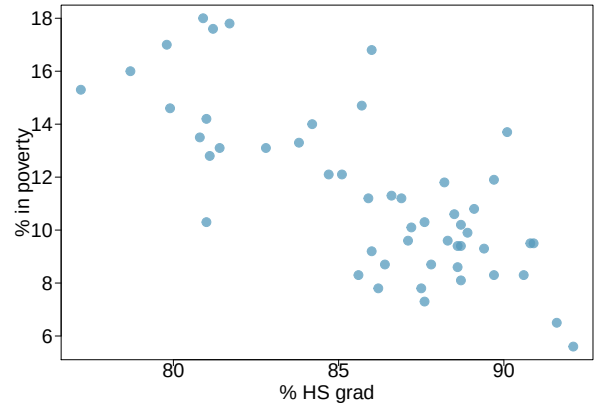
Which of the below is the correct interpretation of $R = -0.75, R^2 = 0.56$?

A) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

B) 56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

C) 56% of the time % HS graduates predict % living in poverty correctly.

D) 75% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Interpretation of R^2

Which of the below is the correct interpretation of $R = -0.75, R^2 = 0.56$?

A) 56% of the variability in the % of HG graduates among the 51 states is explained by the model.

B) 56% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

C) 56% of the time % HS graduates predict % living in poverty correctly.

D) 75% of the variability in the % of residents living in poverty among the 51 states is explained by the model.