Section 001 CDS 101 *Introduction to Computational and Data Sciences*

Class 26: Modeling II

April 26, 2018

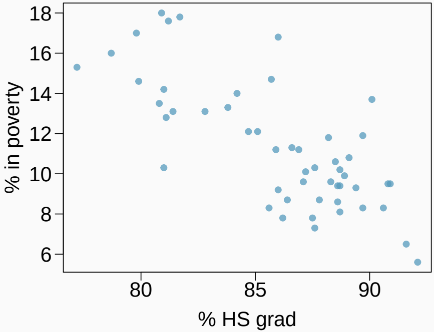
Line fitting, residuals, and correlation

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. HS graduate rate

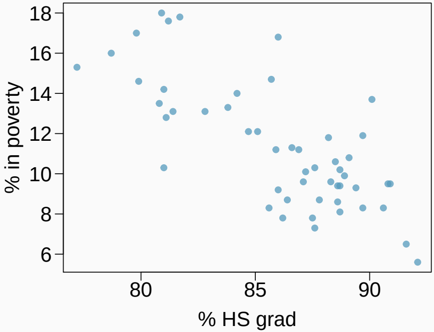
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

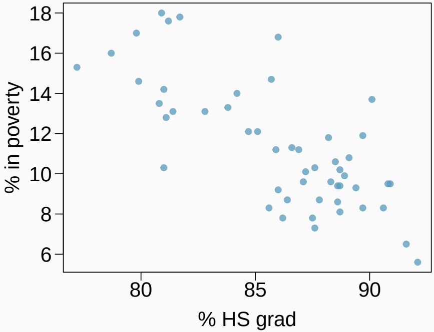


Response variable?

% in poverty

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



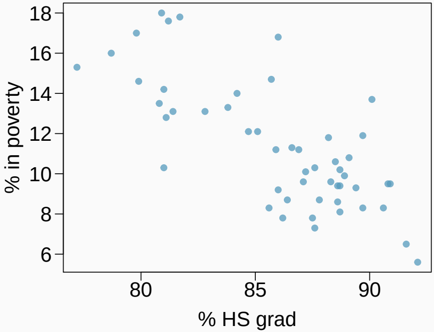
Response variable?

% in poverty

Explanatory variable?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

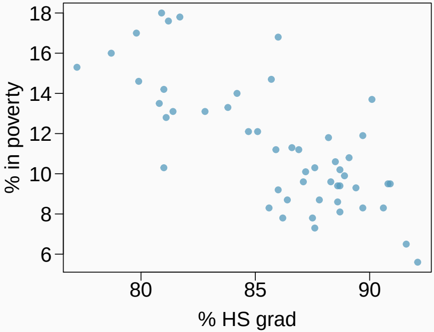
% in poverty

Explanatory variable?

% HS grad

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

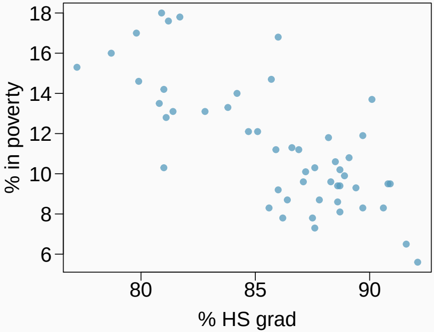
Explanatory variable?

% HS grad

Relationship?

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % 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

Quantifying the relationship

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

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).

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 % in poverty and % HS grad?

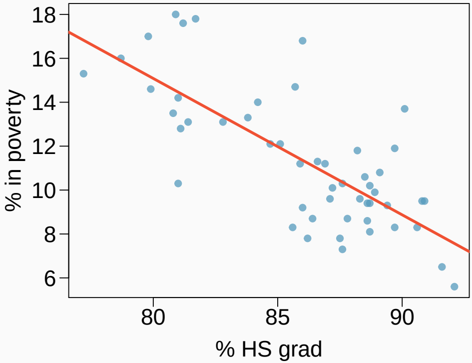
(a) 0.6

(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 % in poverty and % HS grad?

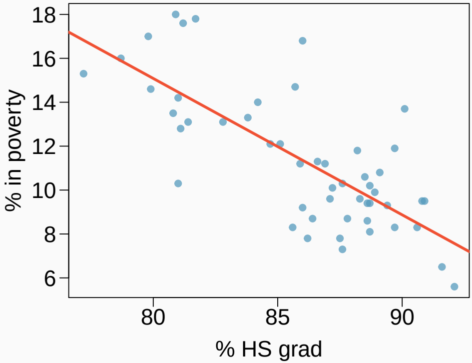
(a) 0.6

(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 % in poverty and % HS grad?

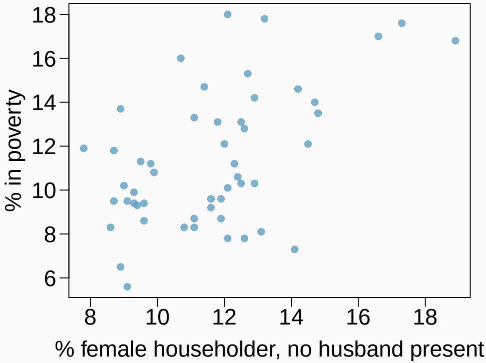
(a) 0.1

(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 % in poverty and % HS grad?

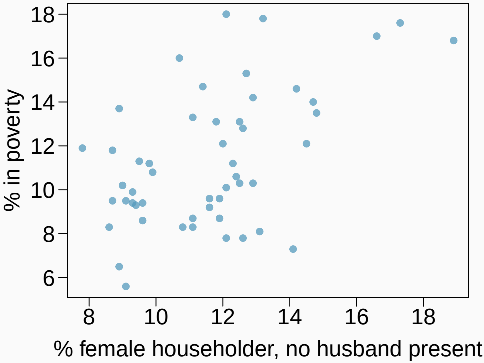
(a) 0.1

(b) -0.6

(c) -0.4

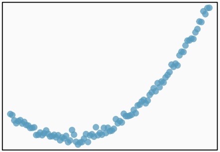
(d) 0.9

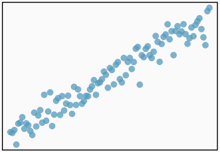
(e) *0.5*

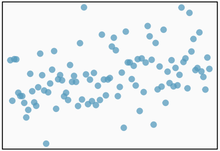


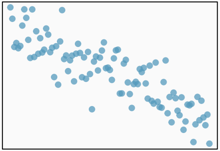
Assessing the correlation

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


(a)

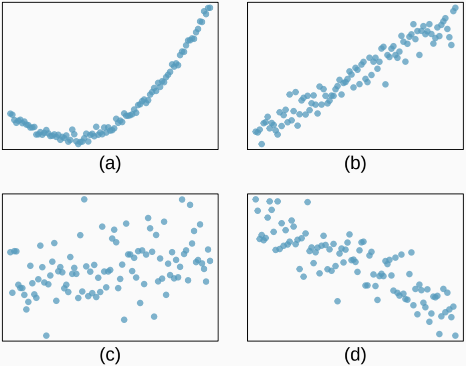

(b)


(c)


(d)

Assessing the correlation

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

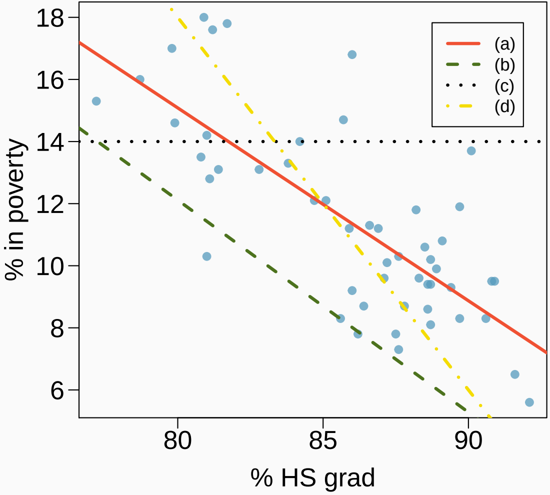


(b) → correlation means <u>linear</u> association

Fitting a line by least squares regression

Eyeballing the line

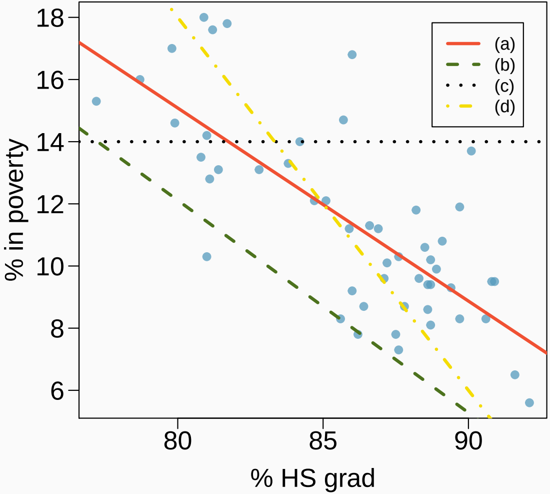
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.



Eyeballing the line

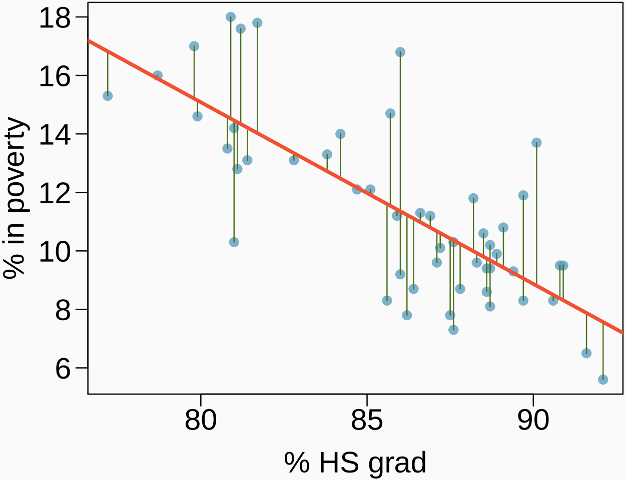
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



Residuals

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

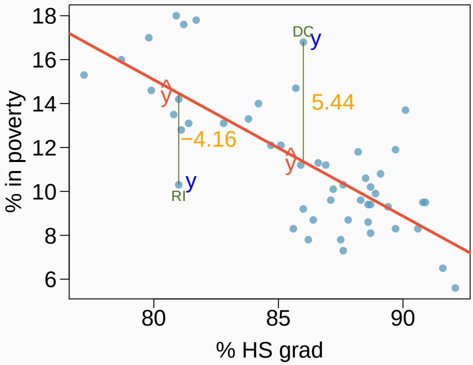


Residuals (cont.)

Residual

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

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

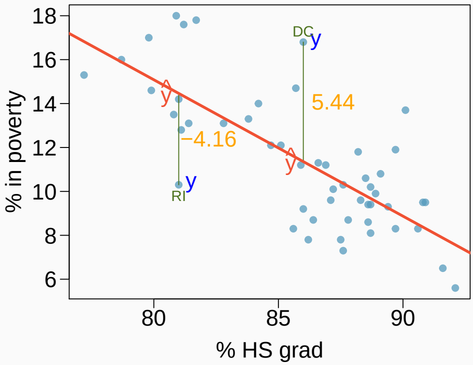


Residuals (cont.)

Residual

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

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



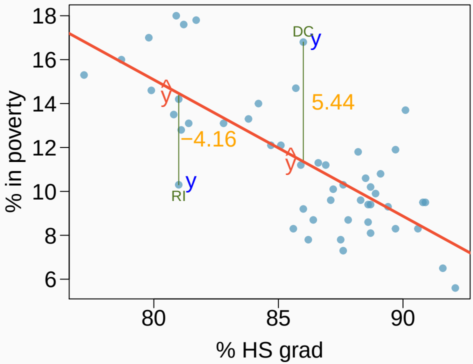
- % living in poverty in DC is 5.44% more than predicted.

Residuals (cont.)

Residual

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.

A measure for the best line

- We want a line that has small residuals:

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

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

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

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

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

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

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

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

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

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

- Why least squares?

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

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

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

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

- Why least squares?
  1. Most commonly used

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

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

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

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

- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software

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

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

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

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

- 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

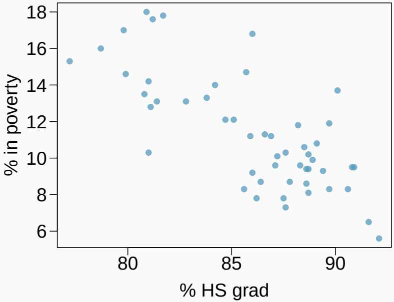
slope

explanatory variable

Notation:

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

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

Slope

The slope of the regression can be calculated as

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

Slope

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

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

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 $(\bar{x}, \bar{y})$.

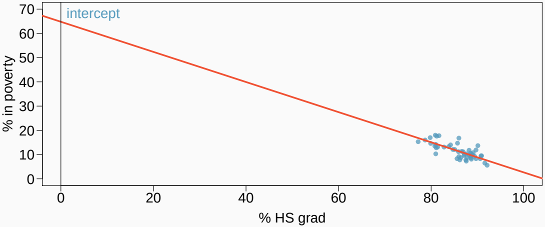
$$b_0 = \bar{y} - b_1\bar{x}$$

Intercept

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 $(\bar{x}, \bar{y})$.

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

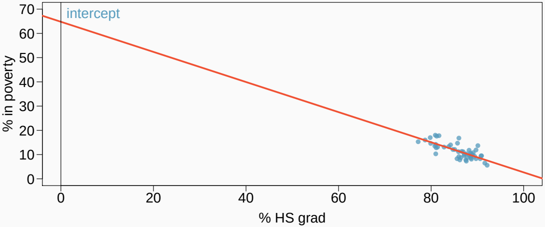


Intercept

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 $(\bar{x}, \bar{y})$.

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



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

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%.

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%.

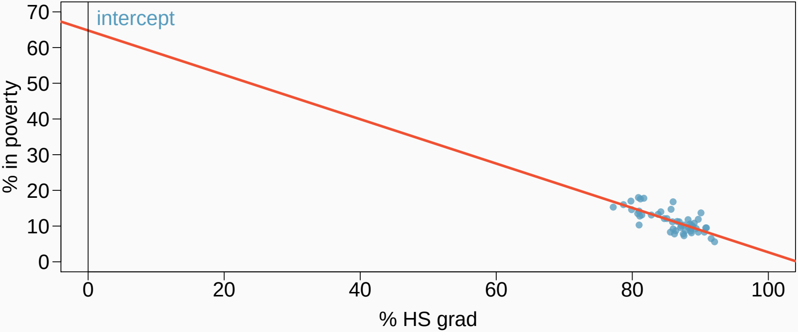
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(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%.

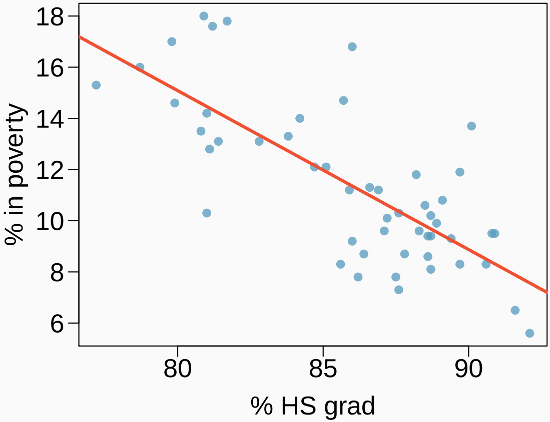
More on the intercept

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

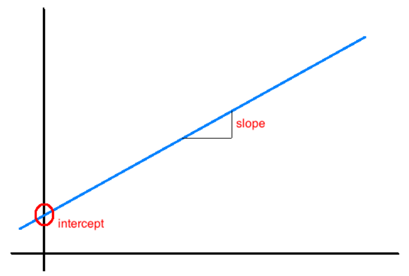
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretation of slope and intercept

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

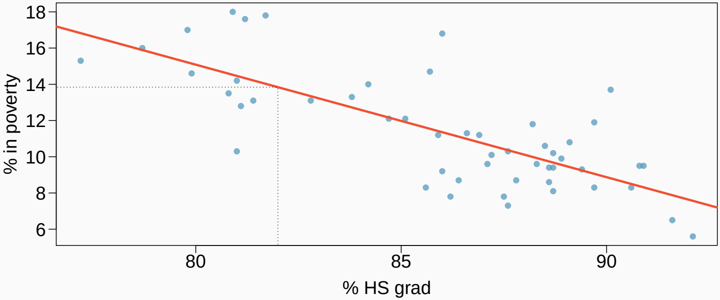
- *Slope:* For each unit in x, y is expected to increase / decrease on average by the slope.



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

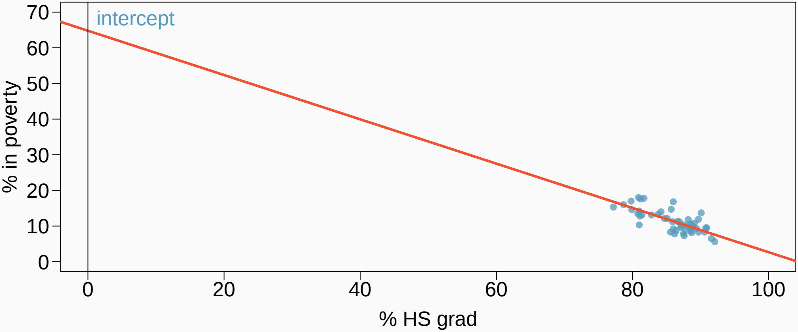
Prediction

- 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.

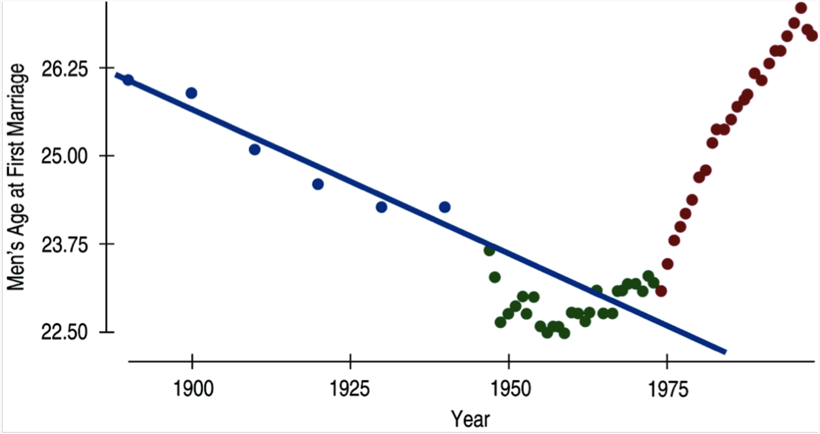


Extrapolation

- 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



Examples of extrapolation

BBC NEWS

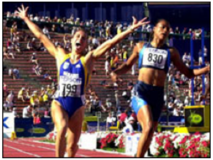
Watch **One-Minute World News**

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."

Examples of extrapolation

Momentous sprint at the 2156 Olympics?

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

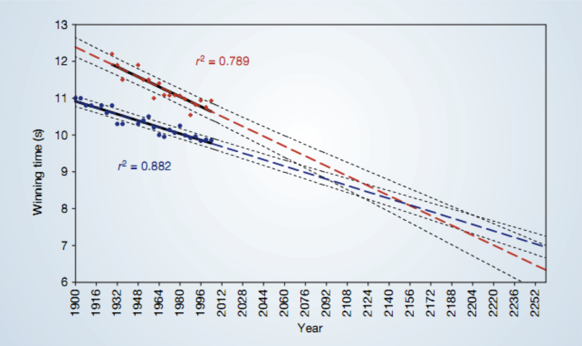


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively)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

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: (1) Linearity

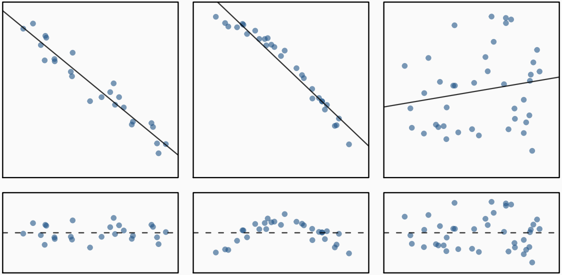
- The relationship between the explanatory and the response variable should be linear.

Conditions: (1) Linearity

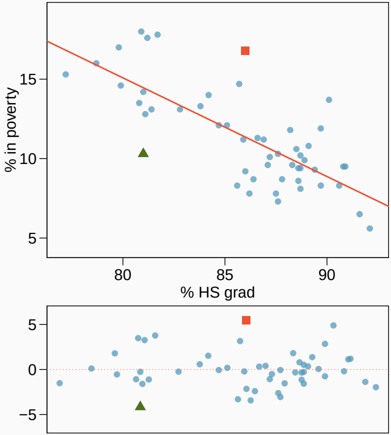
- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.

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:*

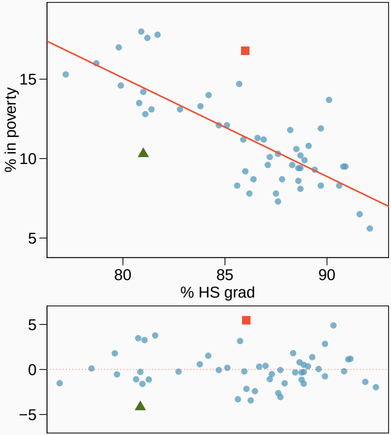
$$\%\ HS\ grad = 81 \qquad \%\ in\ poverty = 10.3$$

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

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

$$= 10.3 - 14.46 = -4.16$$

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

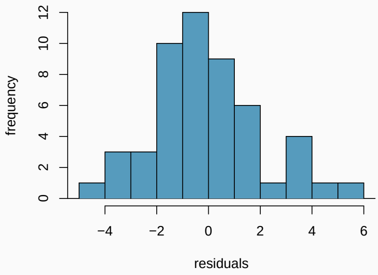
- The residuals should be nearly normal.

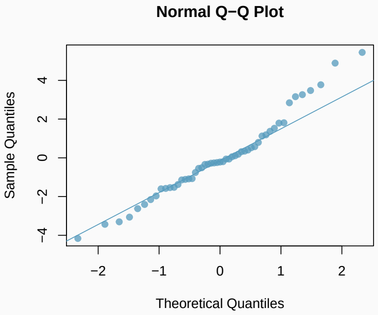
Conditions: (2) Nearly normal residuals

- 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.

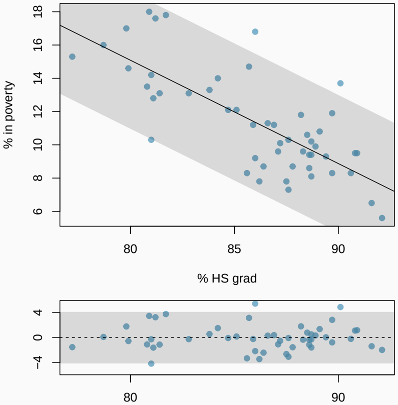
Conditions: (2) Nearly normal residuals

- 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.



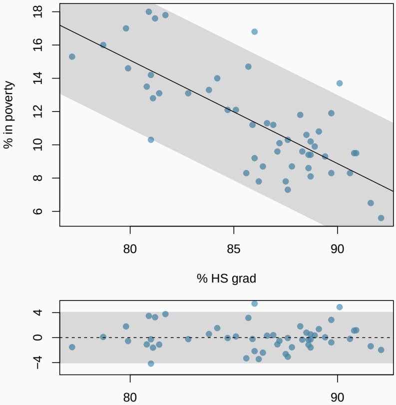


Conditions: (3) Constant variability



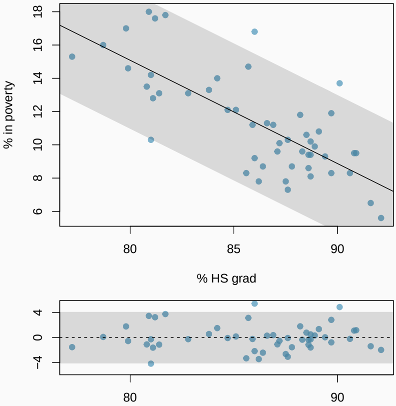
- The variability of points around the least squares line should be roughly constant.

Conditions: (3) Constant variability



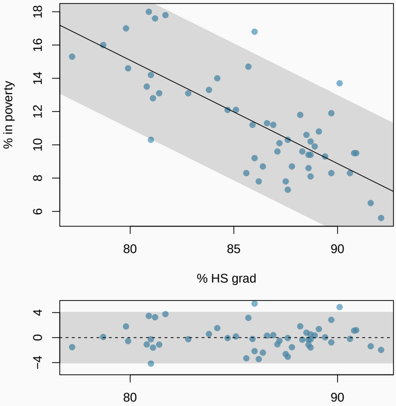
- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.

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*.

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

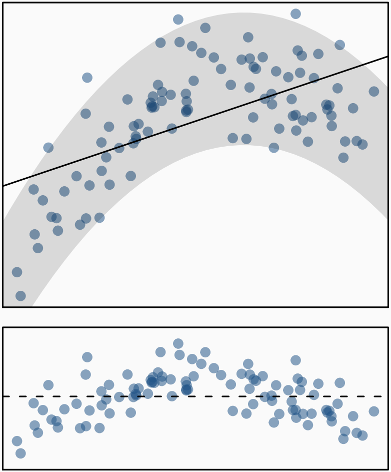
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Checking conditions

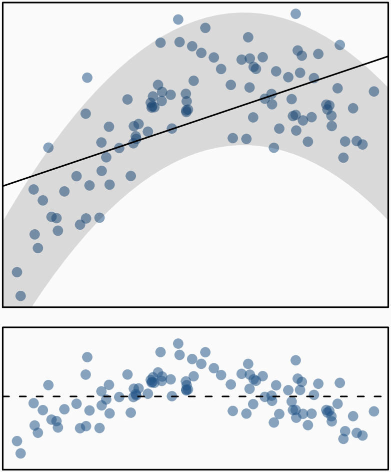
What condition is this linear model obviously violating?

(a) Constant variability

(b) *Linear relationship*

(c) Normal residuals

(d) No extreme outliers



Checking conditions

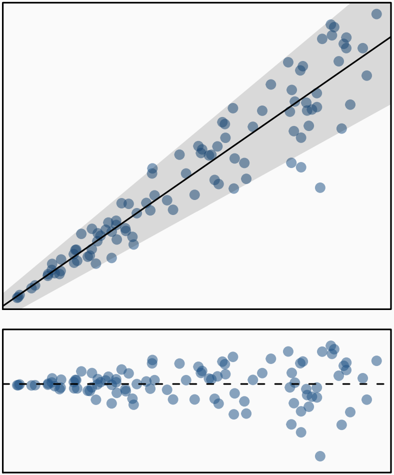
What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

(c) Normal residuals

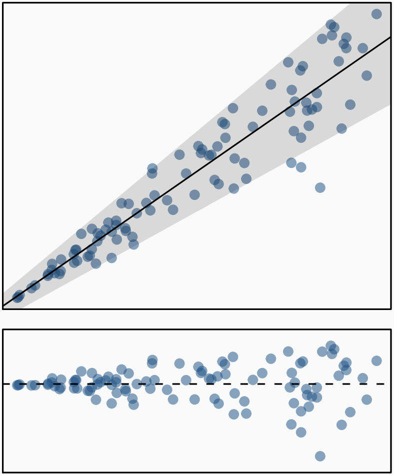
(d) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

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



R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.

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.

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.

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.

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

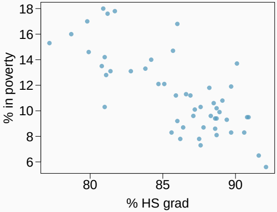
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

(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.



Interpretation of R^2

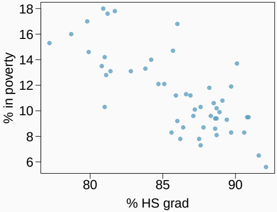
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

(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.



Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
    - This is the value we get if we plug in *0* for the explanatory variable

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* The estimated average poverty percentage in western states is 0.38% higher than eastern states.

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  - Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.

Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  - Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.
  - This is the value we get if we plug in *1* for the explanatory variable

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) *northeast*

(b) midwest

(c) west

(d) south

(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4midwest	0.03	1.15	0.02	0.98
region4west	1.79	1.13	1.59	0.12
region4south	4.16	1.07	3.87	0.00

(a) *northeast*

(b) midwest

(c) west

(d) south

(e) cannot tell