Linear Regression

Recap

- Testing hypotheses allows us to make inferences about population parameters based on data
- E.g., two sample t/z-test tests equality of means of two distributions.
- ANOVA tests equality of means for more than two distributions.
- Linear regression is about predicting the value of a random variable when you know the value of another variable for the same sample.

Covariance

Definition *The covariance of two RVs* X *and* Y *is defined as,*

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Lemma *For any two RVs* X *and* Y*,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

e.g. variance is not a linear operator.

Proof $\mathbf{Var}[X+Y] = \mathbf{E}[(X+Y-\mathbf{E}[X+Y])^2]$

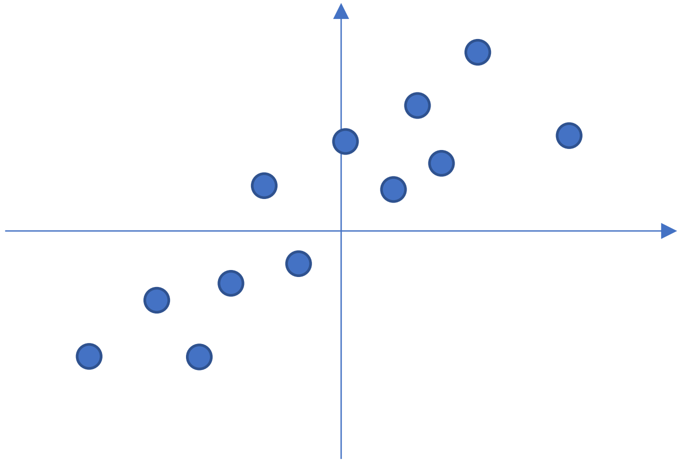
(Linearity of expectation) $= \mathbf{E}[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2]$

(Distributive property) $= \mathbf{E}[(X-\mathbf{E}[X])^2 + (Y-\mathbf{E}[Y])^2 + 2(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Linearity of expectation) $= \mathbf{E}[(X-\mathbf{E}[X])^2] + \mathbf{E}[(Y-\mathbf{E}[Y])^2] + 2\mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]$

(Definition of Var / Cov) $= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$

Covariance



Assume zero means

If $XY > 0 \Rightarrow E(XY) > 0$
If $XY < 0 \Rightarrow E(XY) < 0$

If you have equal number of points in all quarters?

If X, Y are independent:

$$E[(X - E[X])(Y - E[Y])] = E\left[X - E[X]\right]E\left[Y - E[Y]\right] = 0$$

The opposite is not necessarily true

Example:
$X \in \{-1, 0, 1\}, P(X = x) = \frac{1}{3}$ for all x
Find the covariance of X, X^2
Are X, X^2 dependent?

Correlation

- X: Distribution of heights of kids in Greece measured in centimeters.
- Y: Distribution of heights of kids in Greece measured in meters.
- Z: Distribution of head circumference in Greece measured in centimeters

- Which pair has larger covariance: X, Z or Y, Z?

- $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

Covariance and correlation

$$Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = E(XY) - E(X)E(Y)$$

- Captures the degree to which X and Y vary together
- Large covariance: X’s and Y’s are far away from their mean together.
- Are influenced by the magnitude of X and Y
    - Consider heights measured in inches vs centimeters vs meters.

Correlation: $\mathrm{r} = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

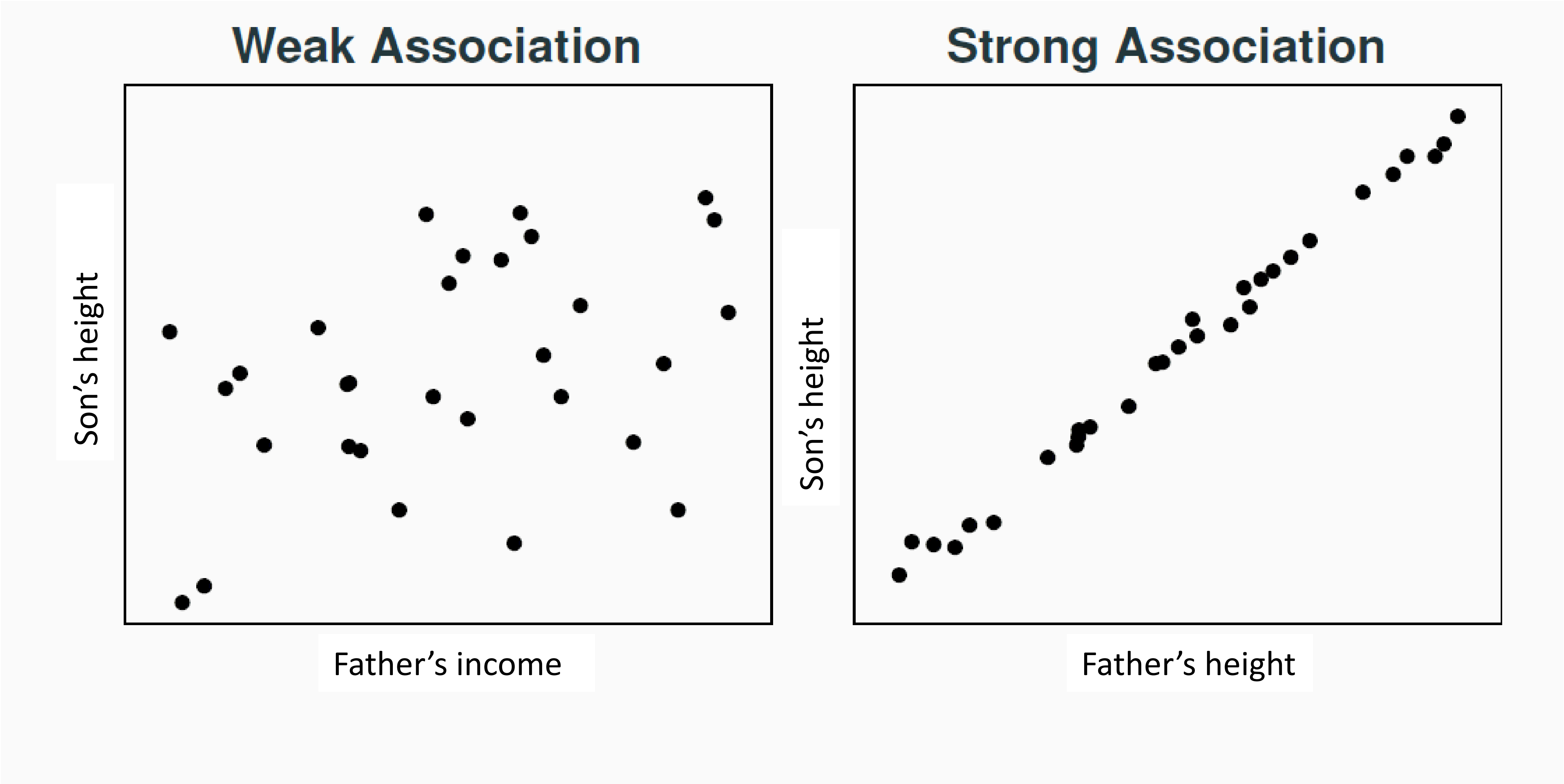
Covariance and correlation

$$Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = E(XY) - E(X)E(Y)$$

- Captures the degree to which X and Y vary together
- Large covariance: X's and Y's are far away from their mean together.
- Are influenced by the magnitude of X and Y
  - Consider heights measured in inches vs centimeters vs meters.

Correlation: $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

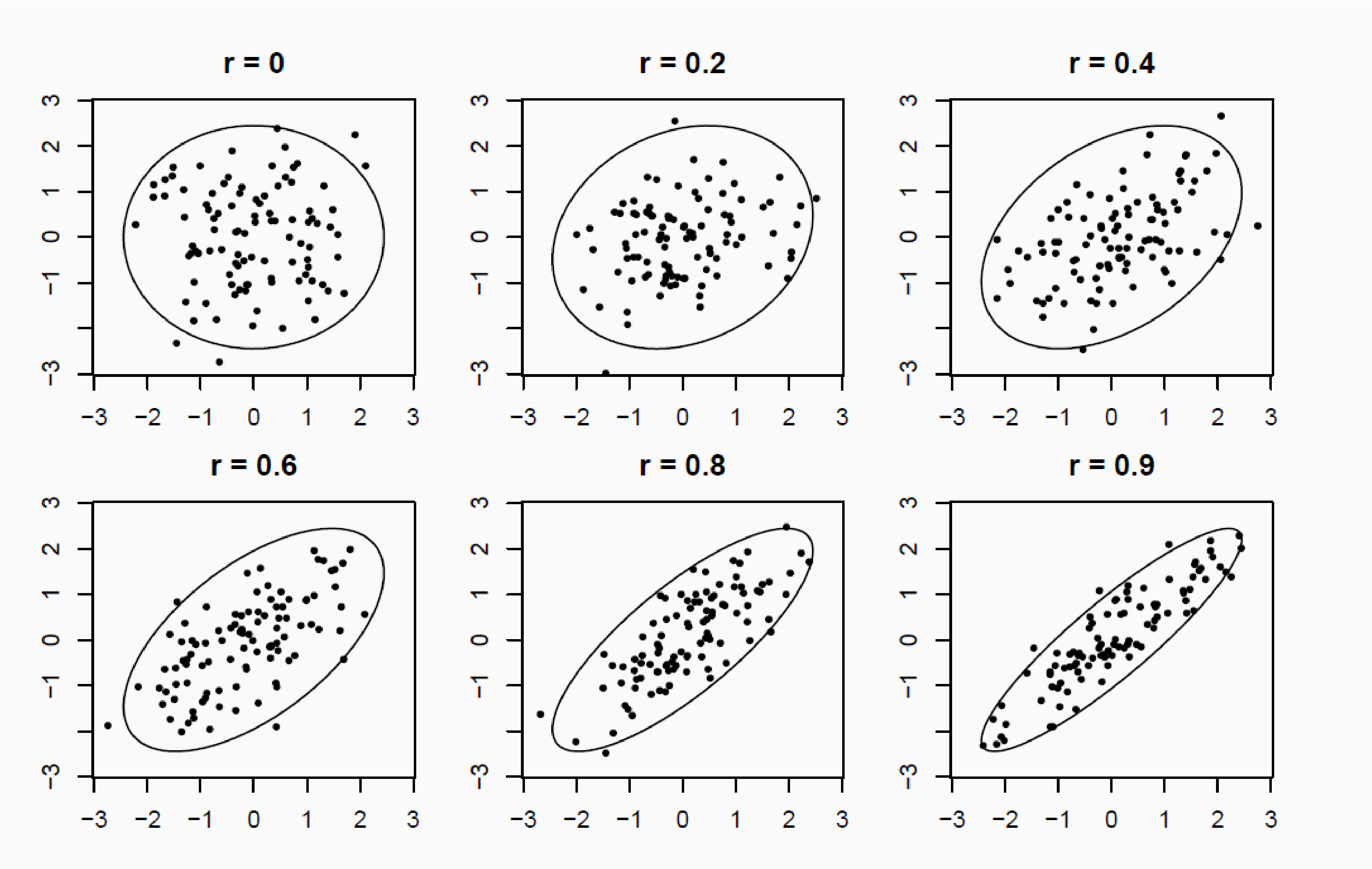
Measuring sample correlation



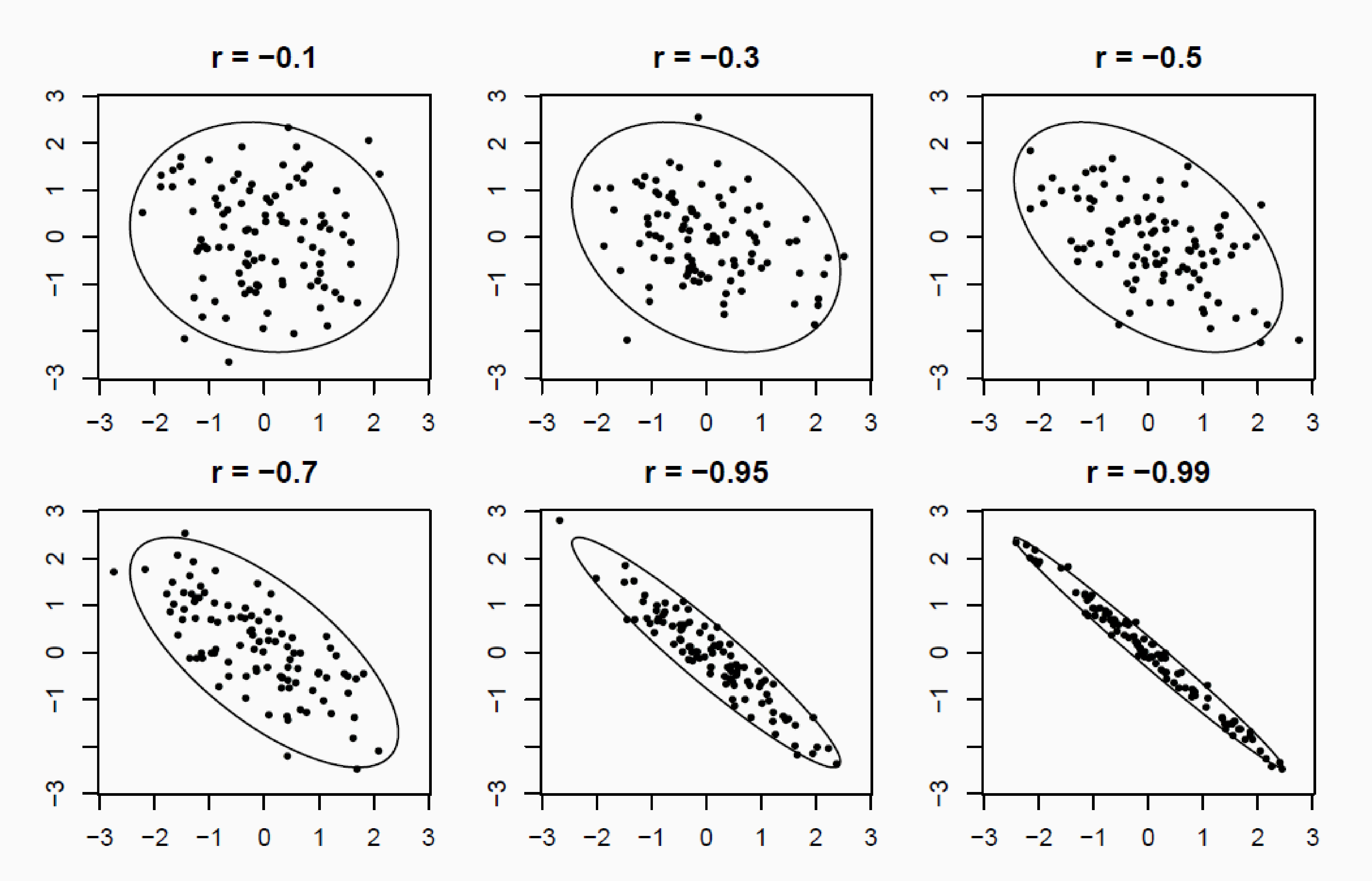
Correlation

- Correlation r is a numerical measure of the direction and strength of the **linear** relationship between two numerical variables.
- “r” always lies between -1 and 1; the strength increases as you move away from 0 to either -1 or 1.
  - $r > 0$: positive association
  - $r < 0$: negative association
  - $r \approx 0$: very weak linear relationship
  - large $|r|$: strong linear relationship
  - $r = -1$ or $r = 1$: only when all the data points on the scatterplot lie exactly along a straight line

Positive Correlations



Negative Correlations

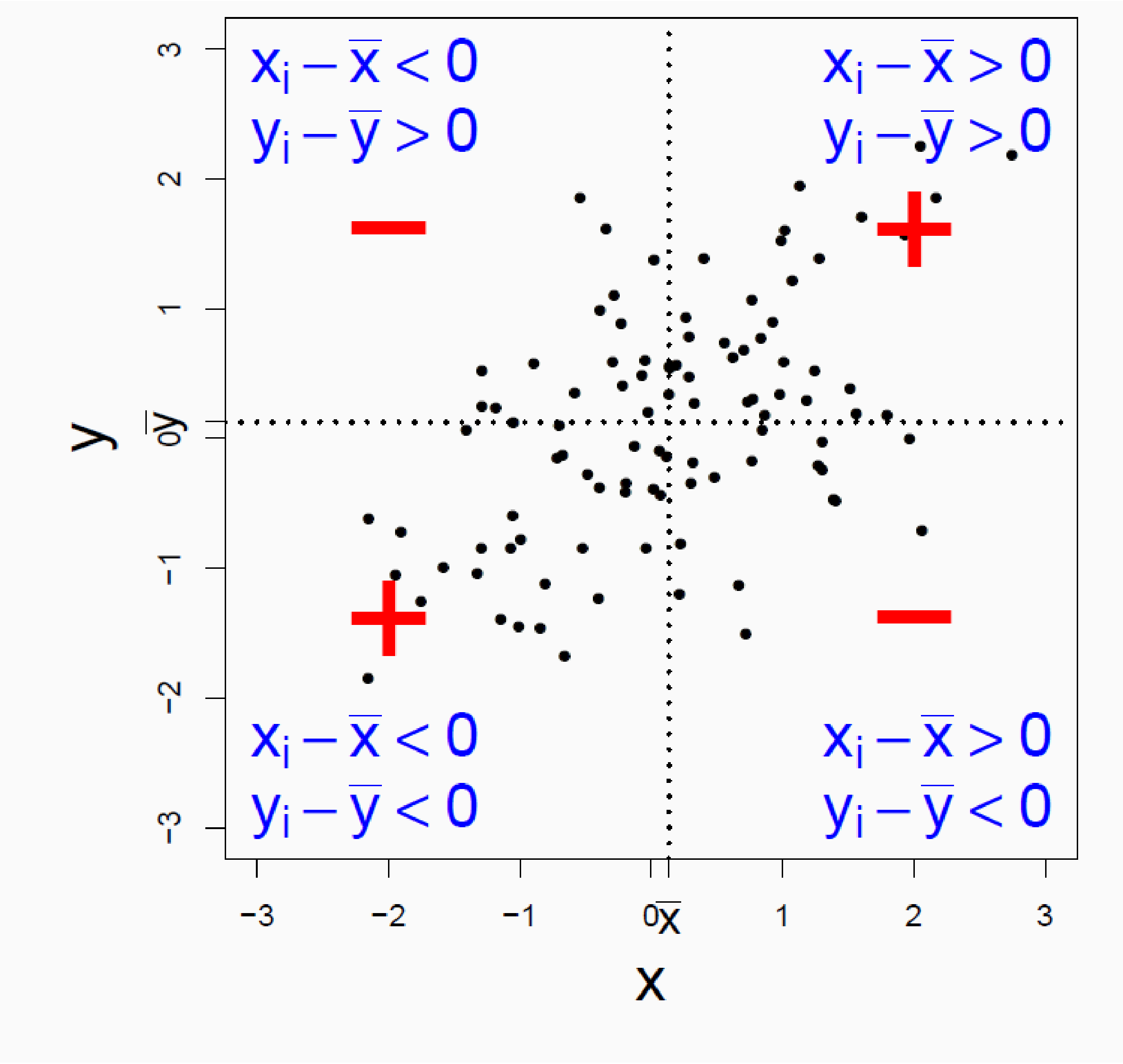


Correlation Coefficient

$$\begin{matrix}(x_1, y_1)\\(x_2, y_2)\\ \vdots \\(x_n, y_n)\end{matrix}$$

- $r = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$

$$\frac{1}{n-1} \sum_i \frac{(x_i - \bar{x})}{\underbrace{s_X}_{\text{sample standard deviation of X}}} \frac{(y_i - \bar{y})}{\underbrace{s_Y}_{\text{sample standard deviation of Y}}}$$



What is the sign of

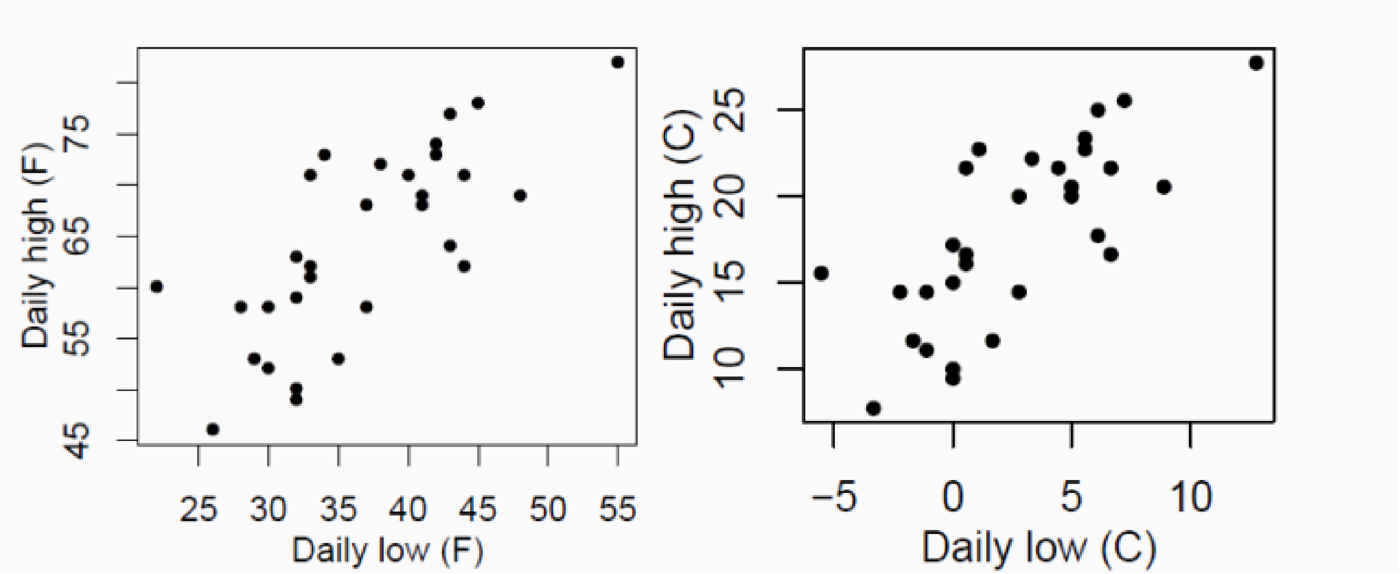
$$\frac{(x_i - \bar{x})}{s_X}\frac{(y_i - \bar{y})}{s_Y}$$

Here r > 0;
more positive contributions than negative.

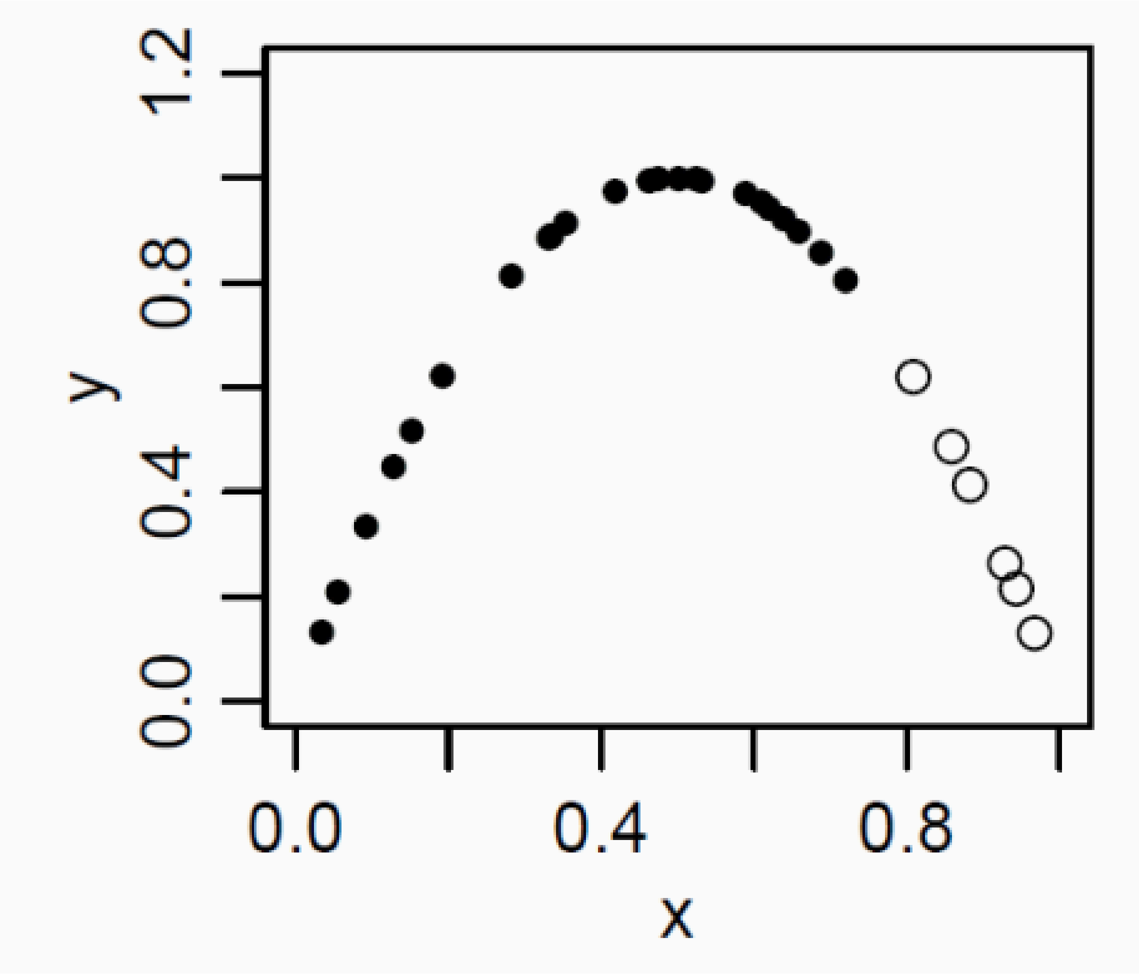
What kind of points have large contributions to the correlation?

Correlation has no unit

- After standardization, the z-score of neither x_i nor y_i has a unit.
- r is unit-free.
- we can compare r between data sets, where variables are measured in different units or when variables are different.



Correlation only measures linear dependence

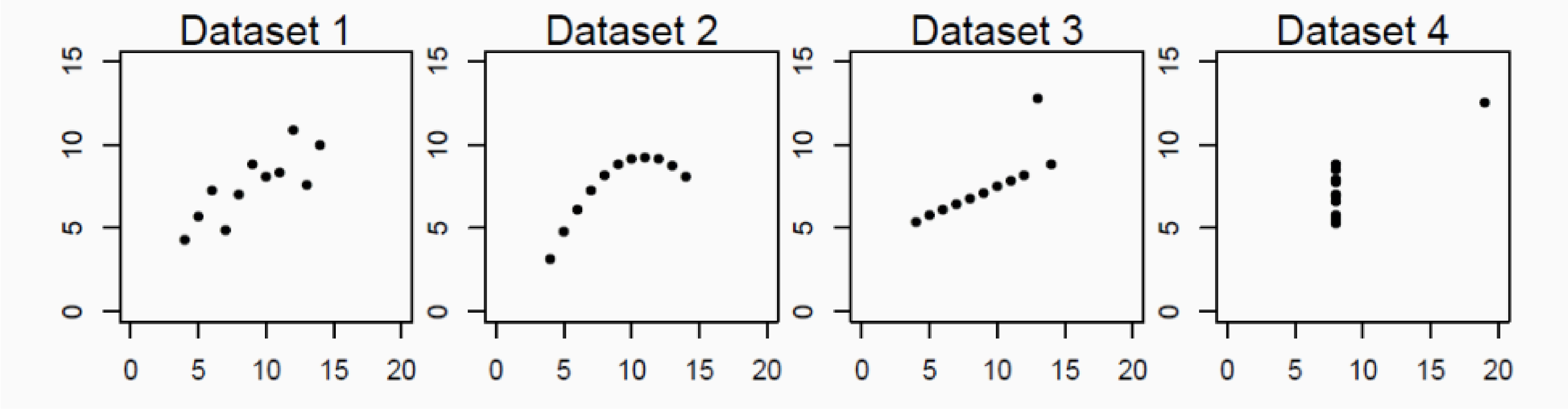


The scatter plot below shows a perfect nonlinear association. All points fall on the quadratic curve

$$y = 1 - 4(x - 0.5)^2$$

r of all black dots = 0.803,
r of all dots = −0.019:
(black + white)

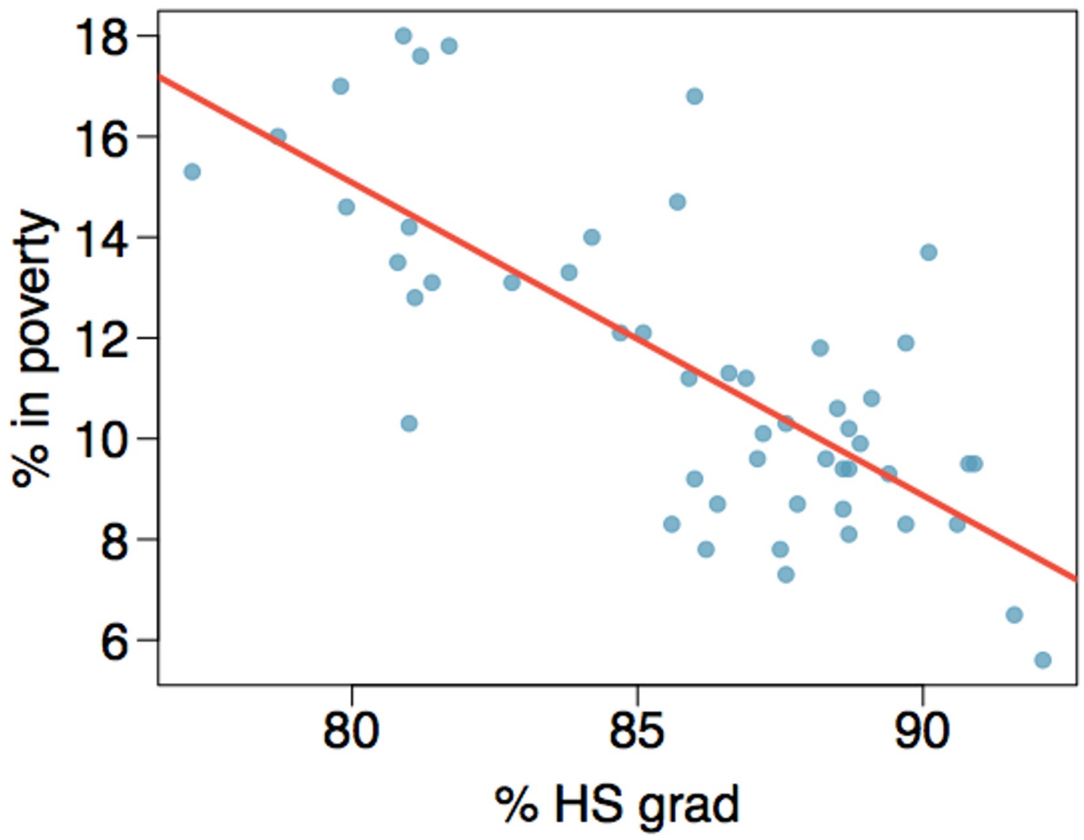
Correlation can be misleading



In all these data sets, r =0.82!

Poverty vs. HS graduate rate

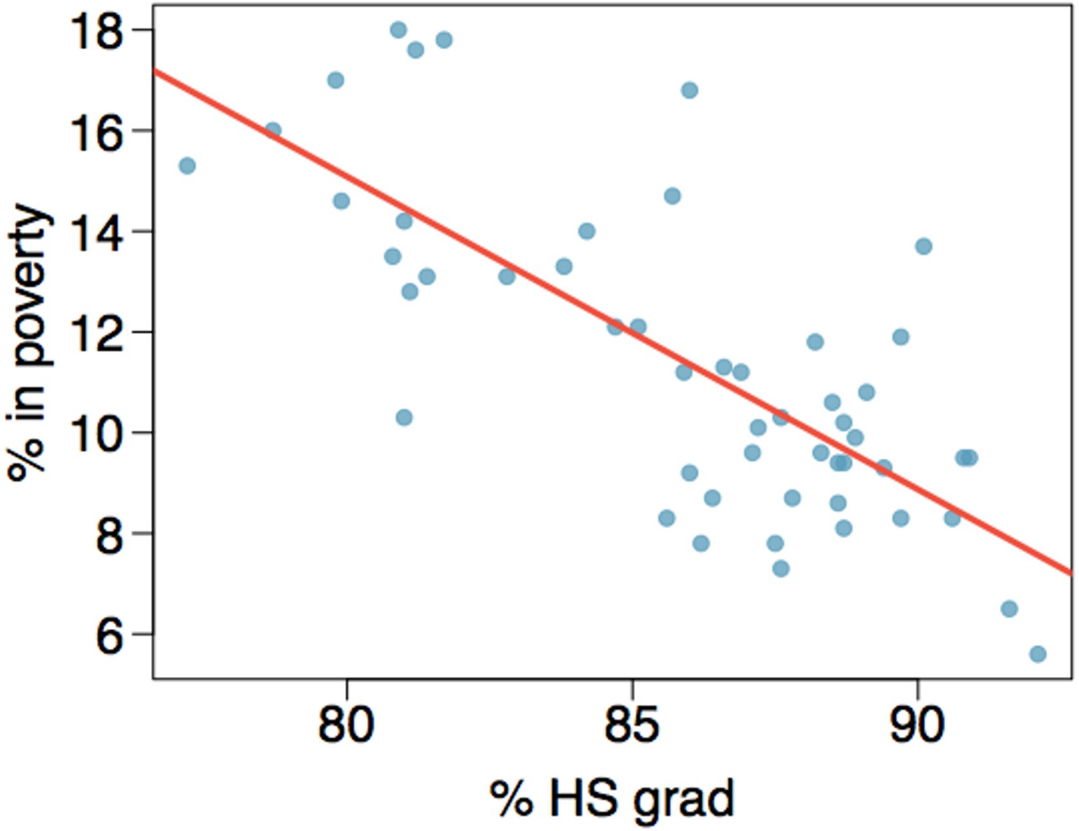
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



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?

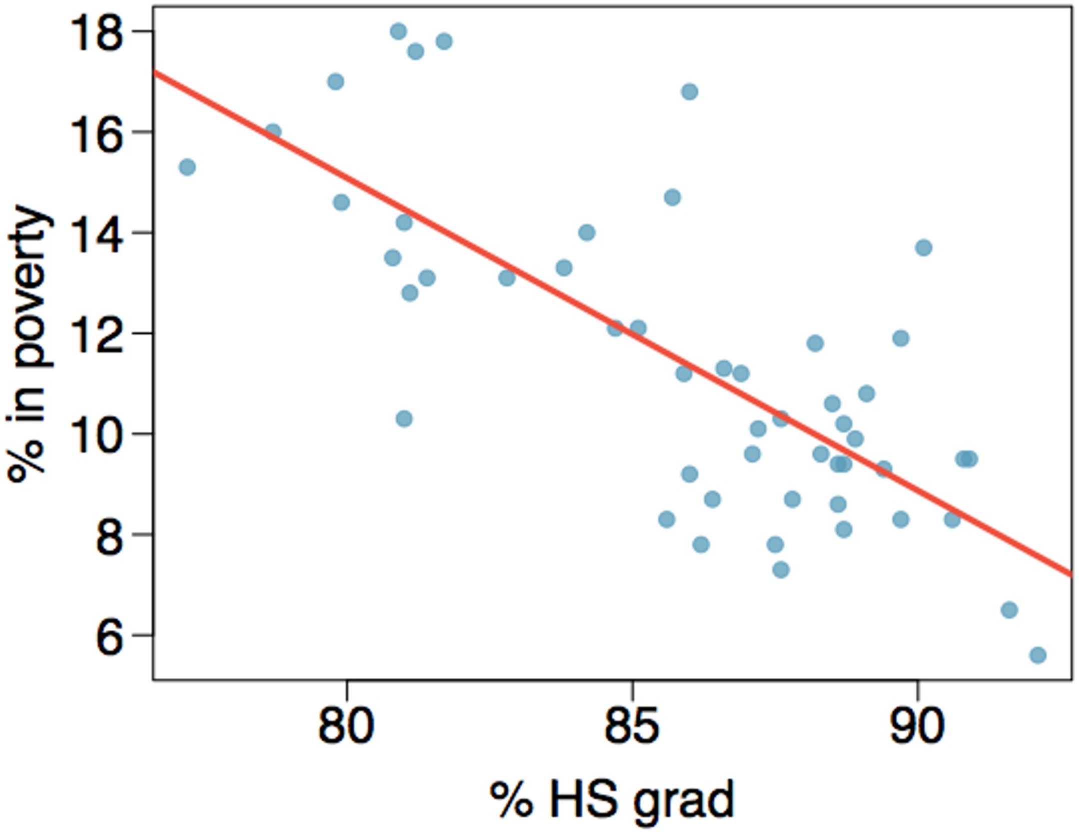
(a) 0.6

(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?

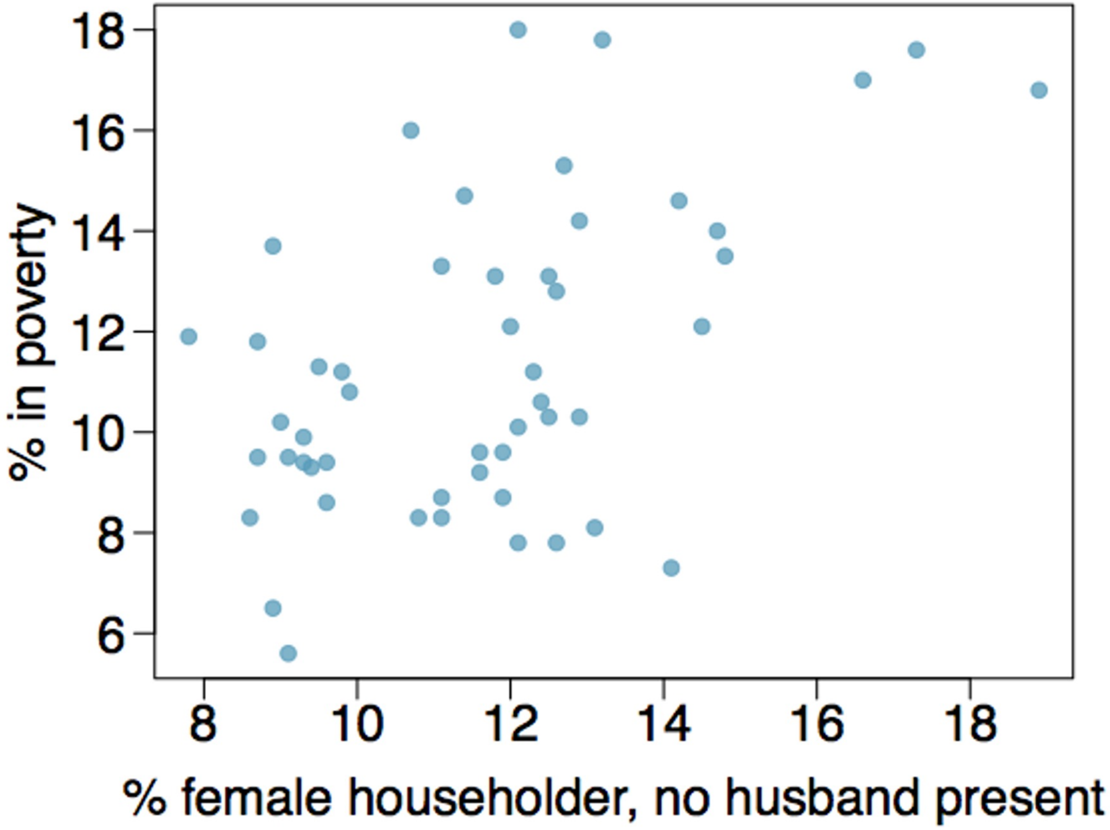
(a) 0.1

(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?

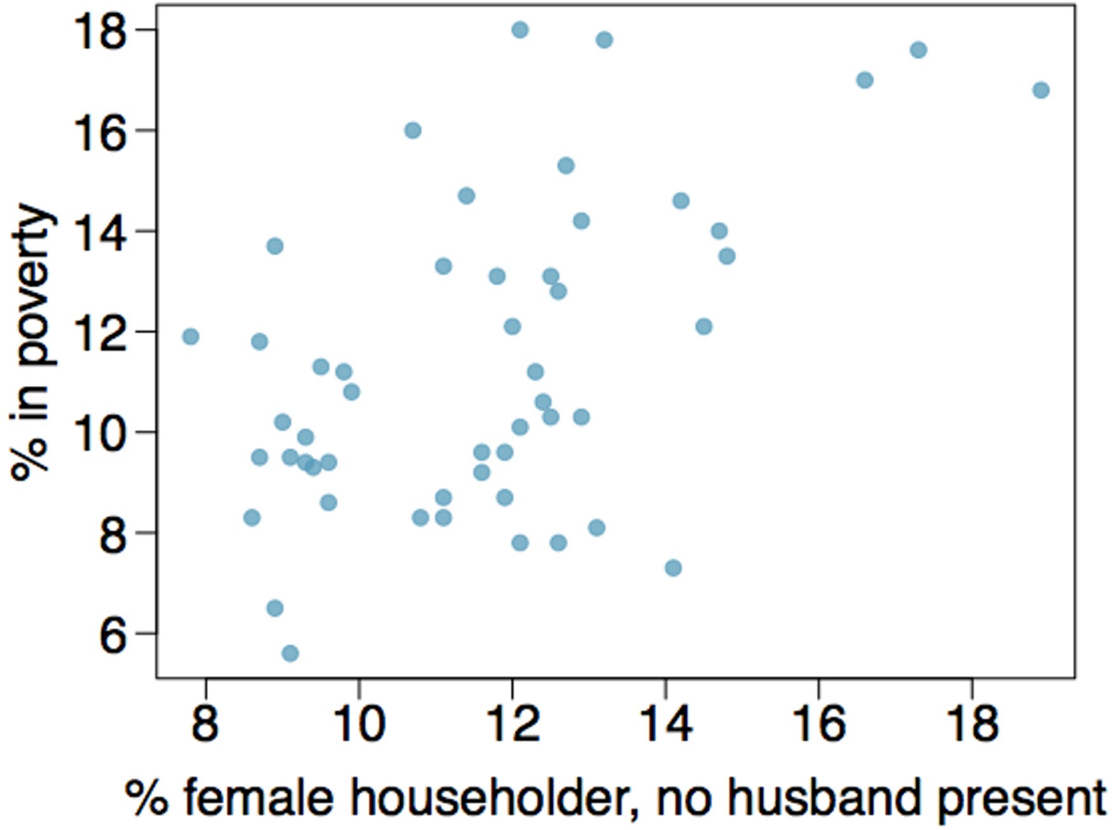
(a) 0.1

(b) -0.6

(c) -0.4

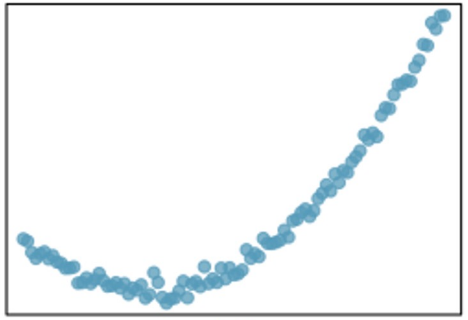
(d) 0.9

(e) 0.5

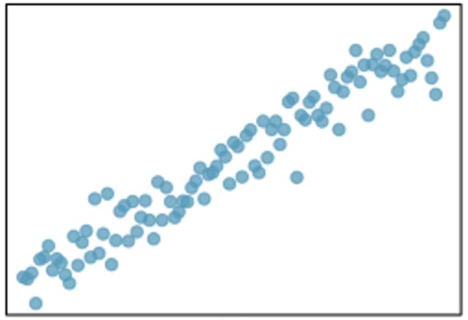


Assessing the correlation

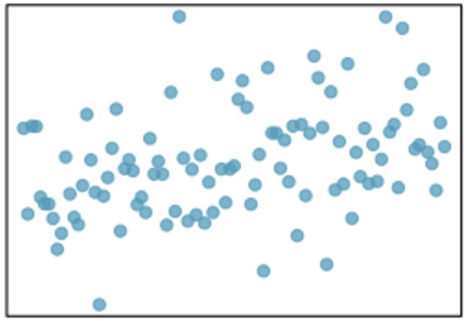
Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



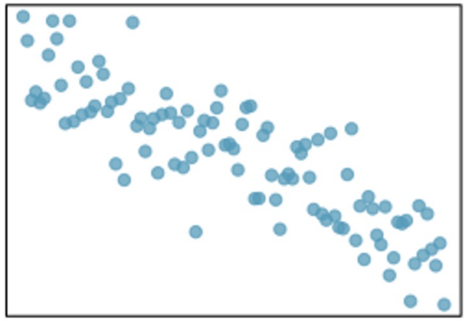
(a)



(b)



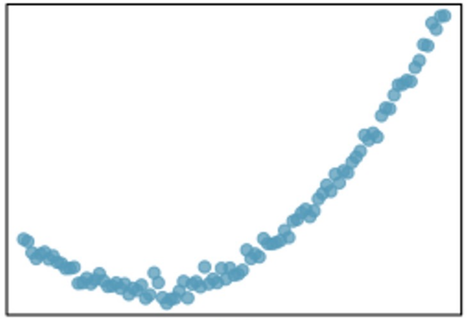
(c)



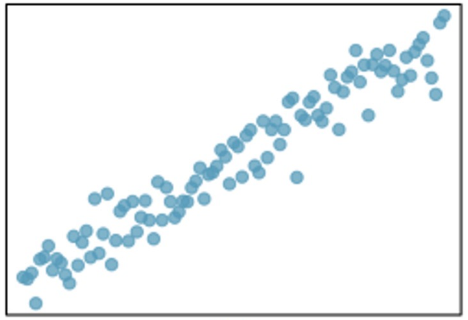
(d)

Assessing the correlation

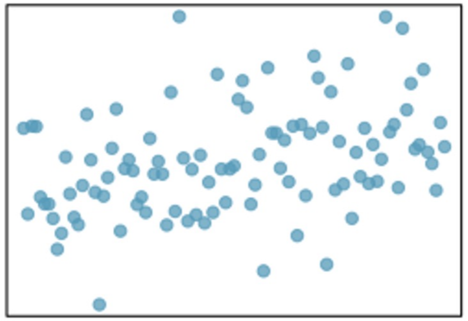
Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



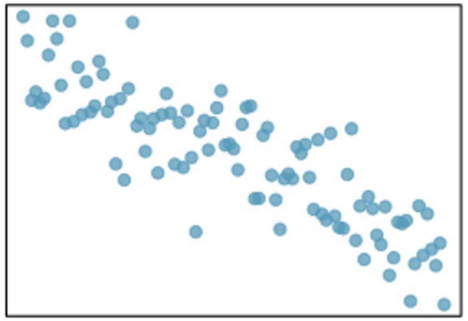
(a)



(b)



(c)



(d)

(b) → correlation means linear association

Line Fitting and Residuals

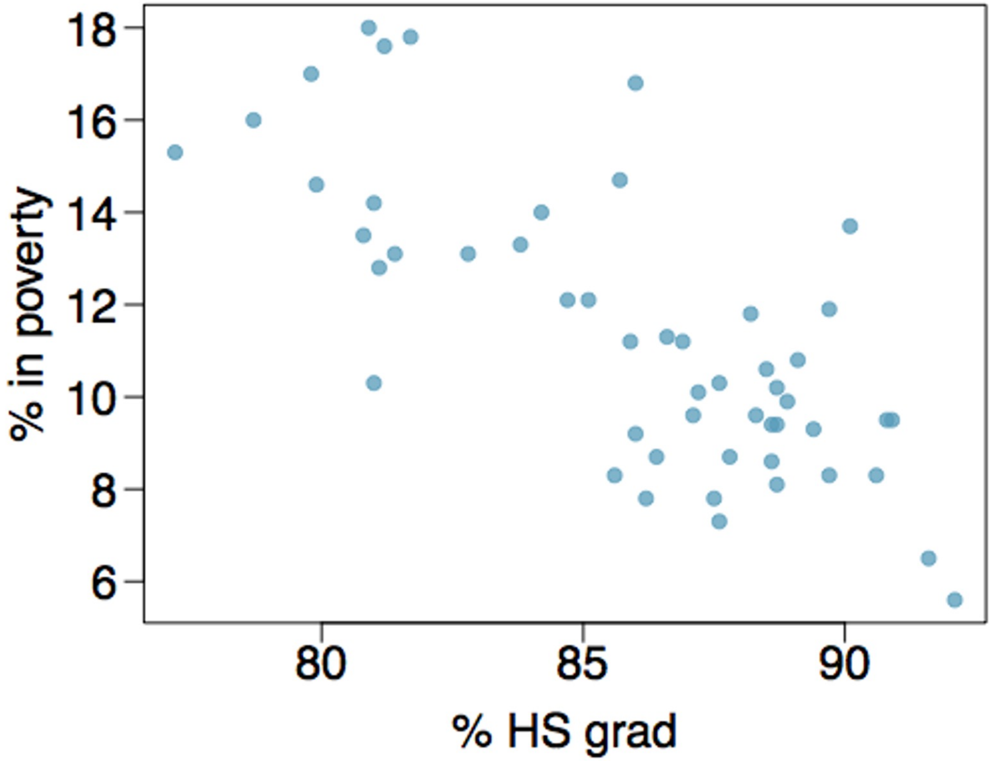
Predicting a numeric variable

Correlation is symmetric: corr(X,Y) = corr(Y, X)

In this unit we will learn to model numerical response variables using a numerical or categorical explanatory variable.

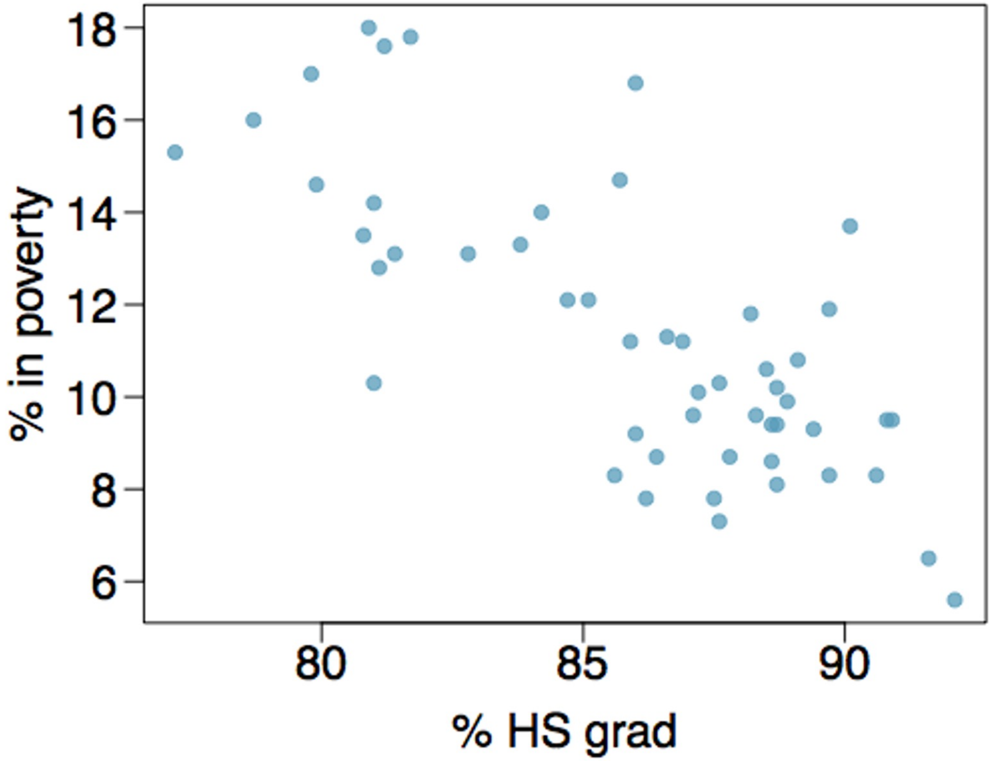
Poverty vs. HS graduate rate

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



Poverty vs. 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).

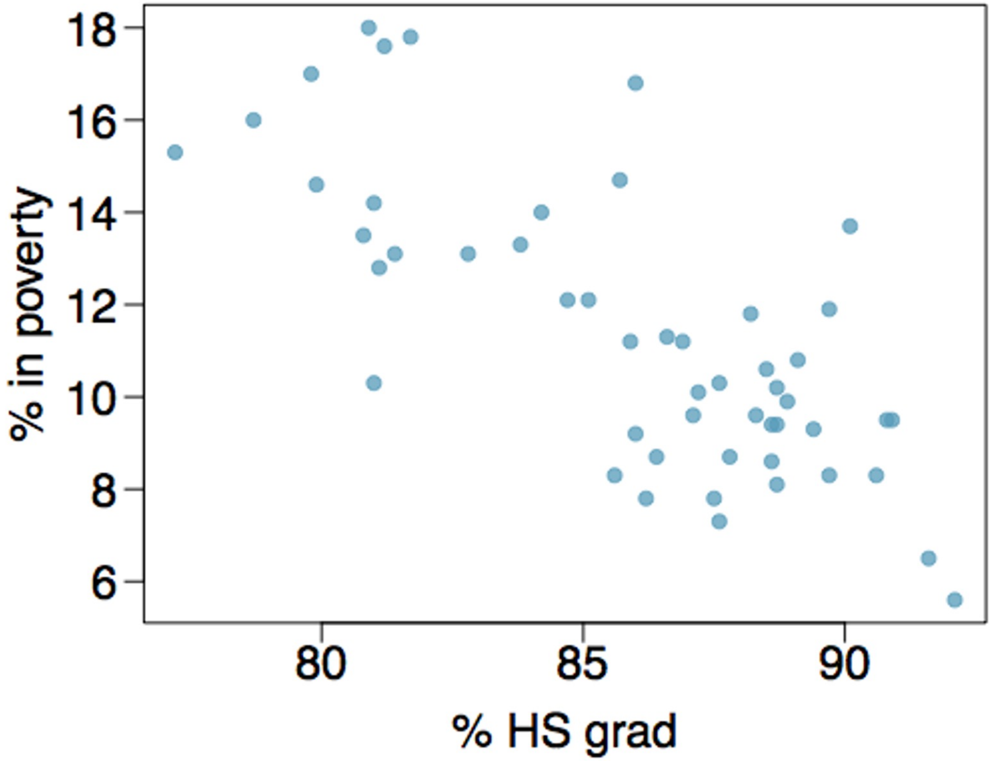


You want to model the relationship of % in poverty given %HS graduates.

i.e., predict the % in poverty given %HS graduates.

Poverty vs. 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).

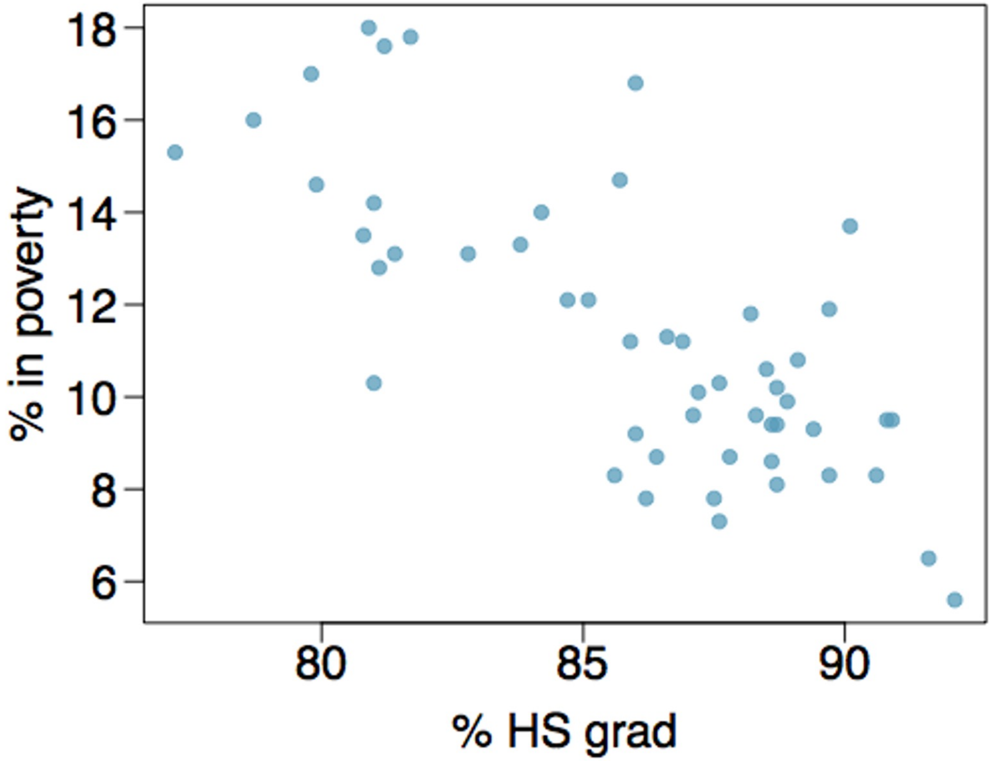


Response variable?

% in poverty

Poverty vs. 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).



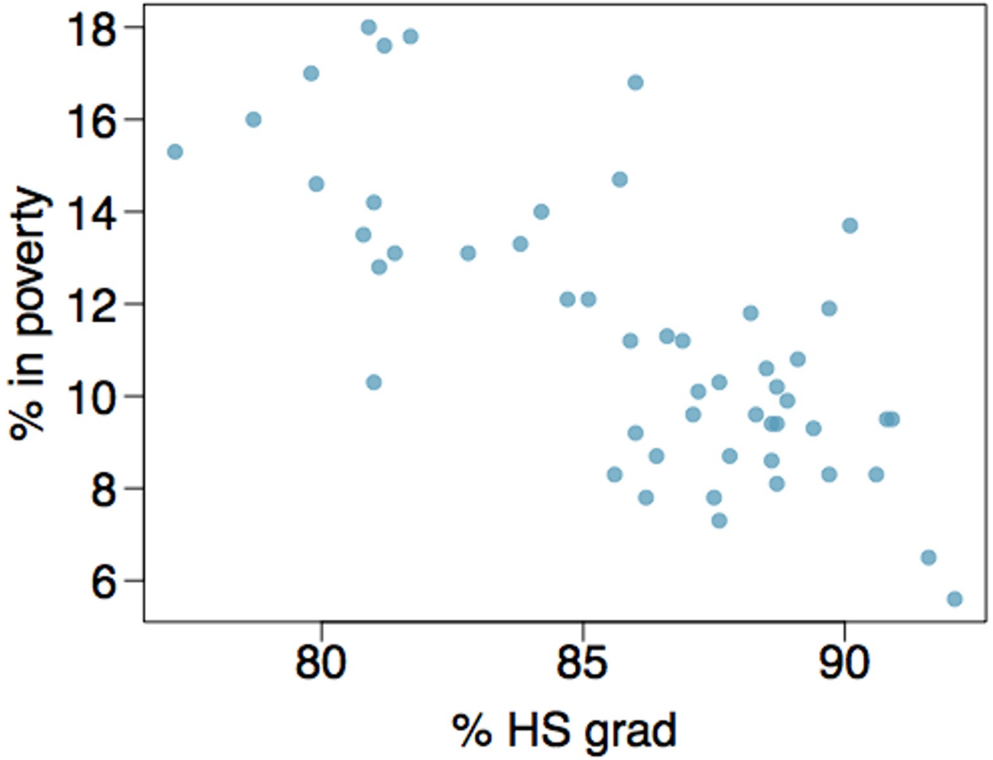
Response variable?

% in poverty

Explanatory variable?

Poverty vs. 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?

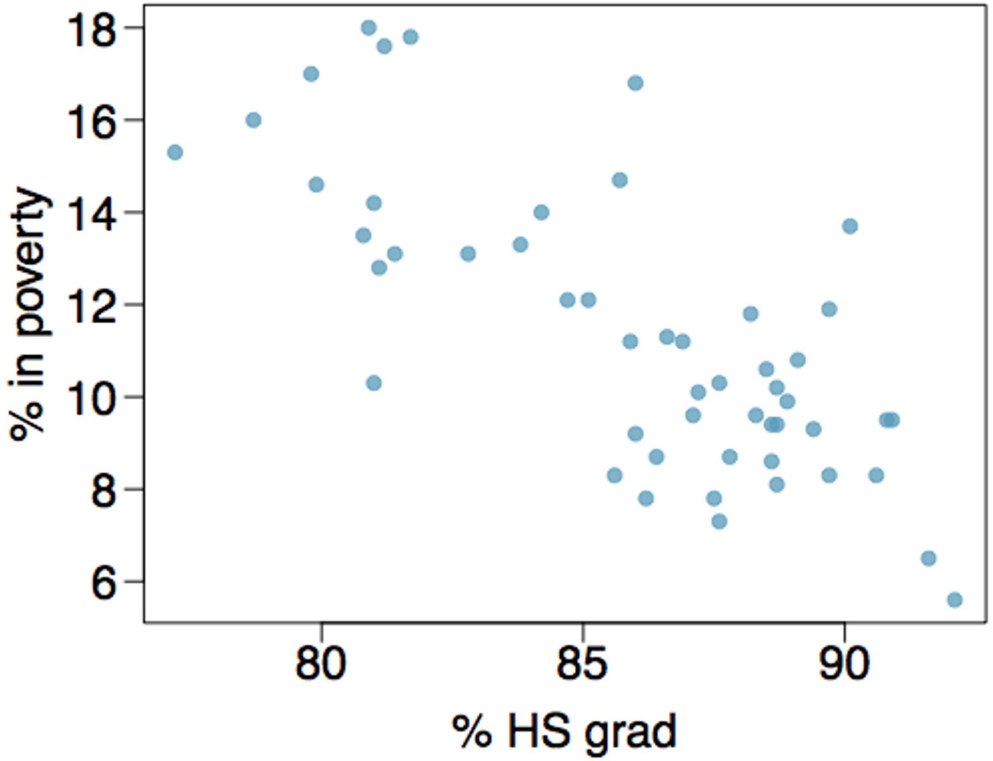
% in poverty

Explanatory variable?

% HS grad

Poverty vs. 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

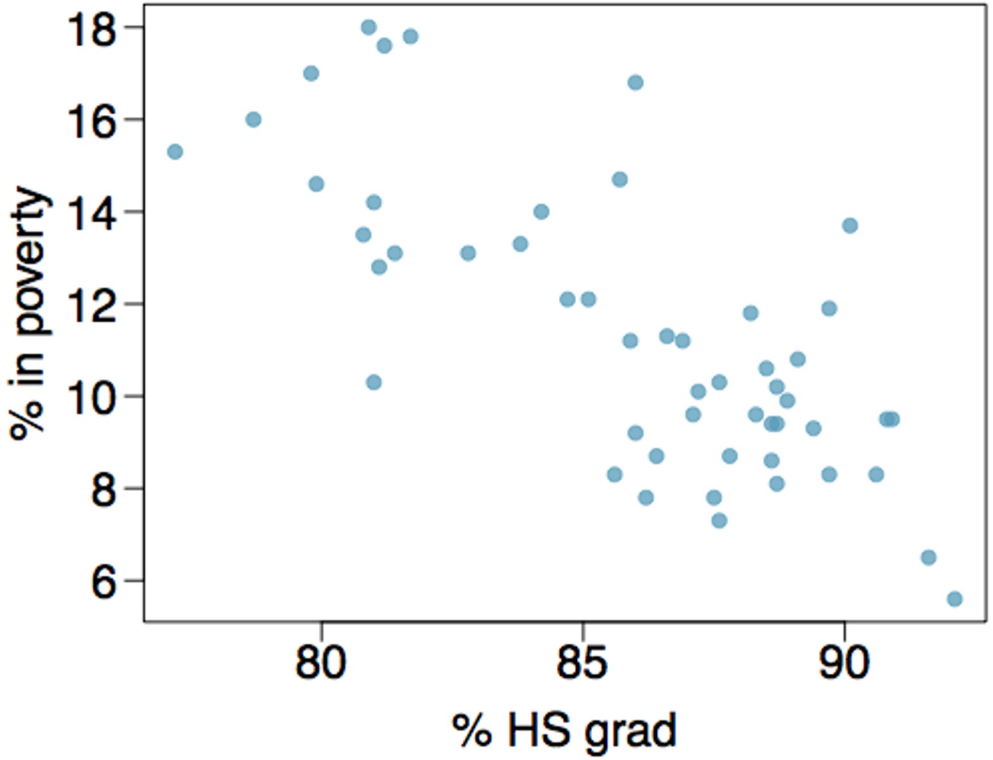
Explanatory variable?

% HS grad

Relationship?

Poverty vs. 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

Poverty vs. HS graduate rate

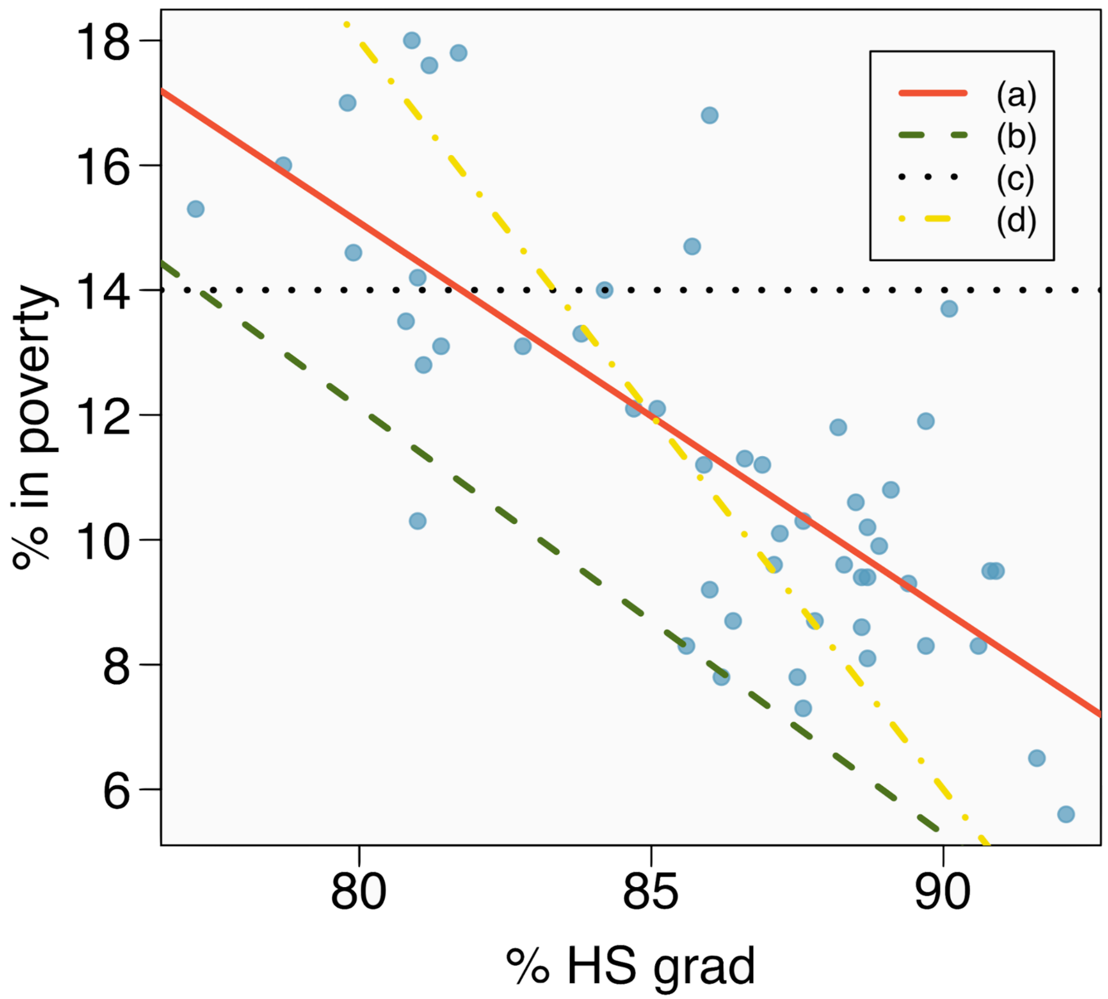
The linear model for predicting poverty from high school graduation rate in the US is

$$\widehat{poverty} = \beta_0 + \beta_1\, HS_{grad}$$

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

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.



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.

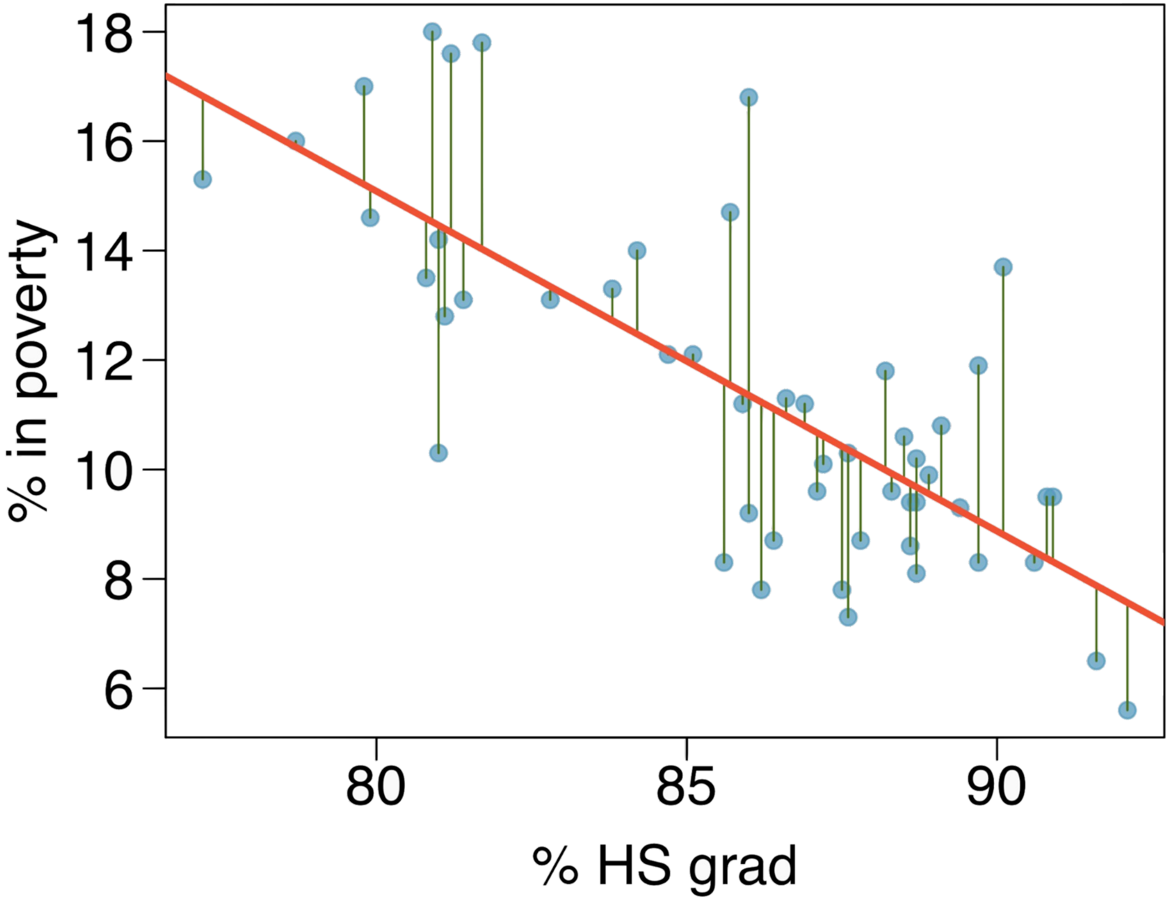
The high school graduate rate in Georgia is 85.1%. What poverty level does the model predict for this state?

64.78 − 0.62 x 85.1 = 12.018

Residuals

Residuals are the leftovers from the model fit:

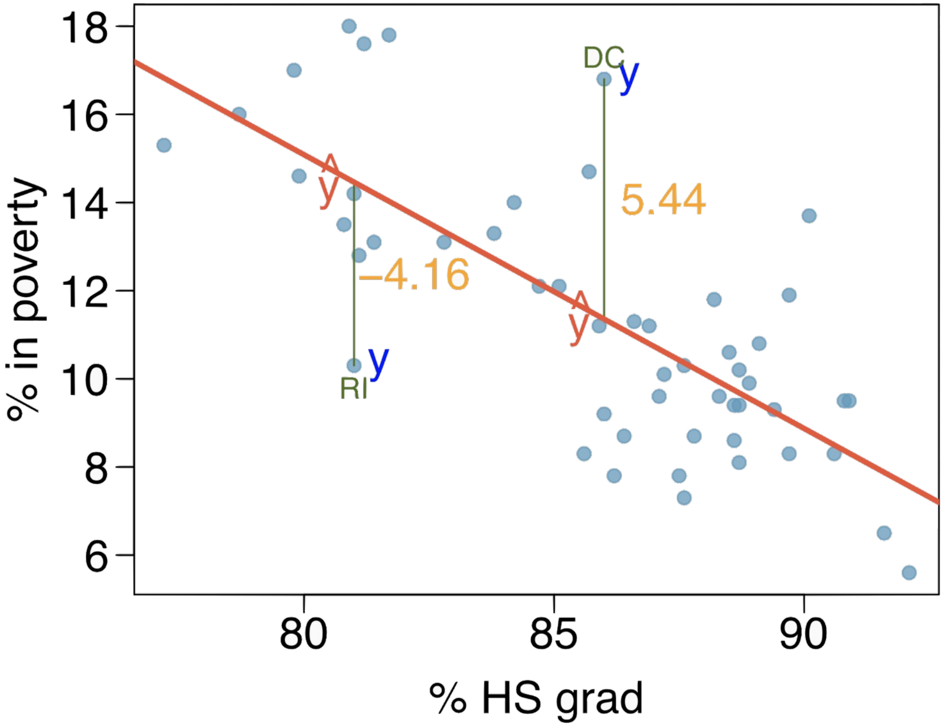
Data = Fit + Residual



Residuals (cont.)

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

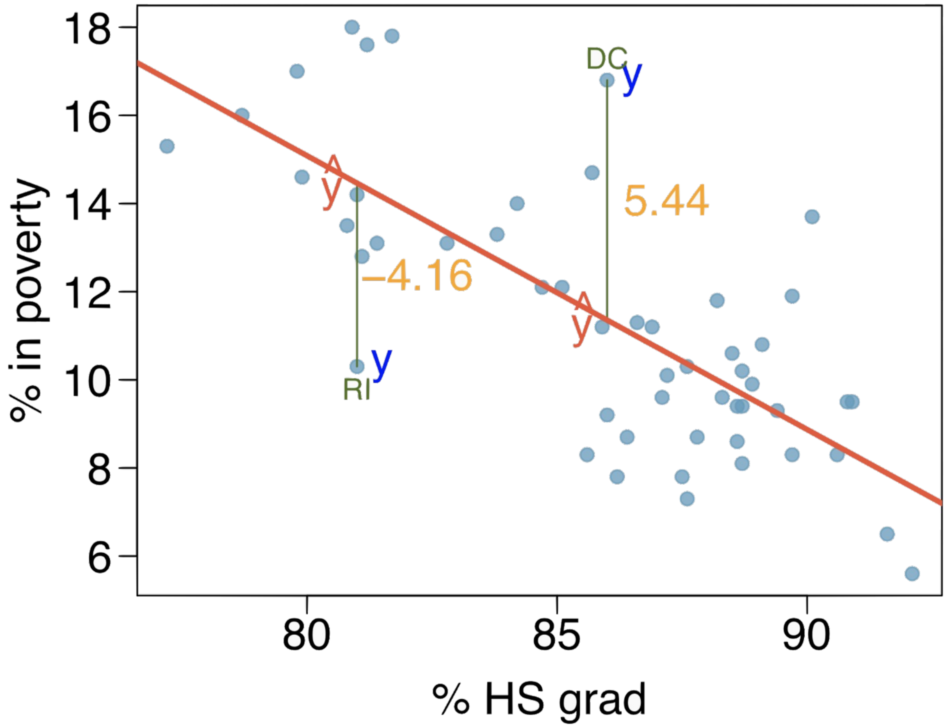
$$e_i = y_i - \hat{y}_i$$



Residuals (cont.)

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

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

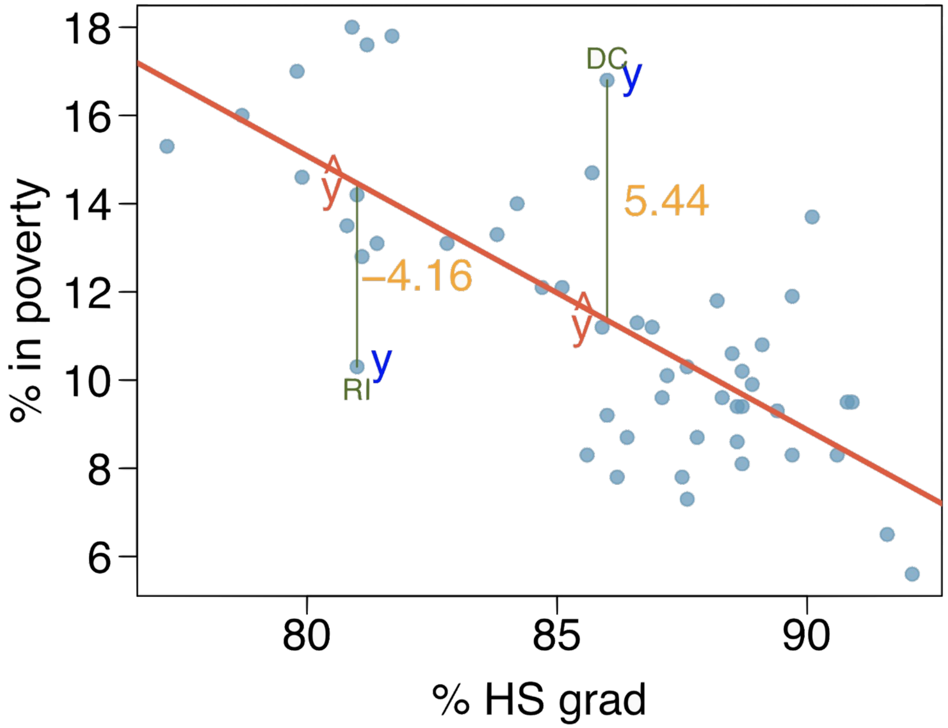


% living in poverty in DC is 5.44% more than predicted.

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.

Fitting a line by least squares regression

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

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

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

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

    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?

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

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*

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

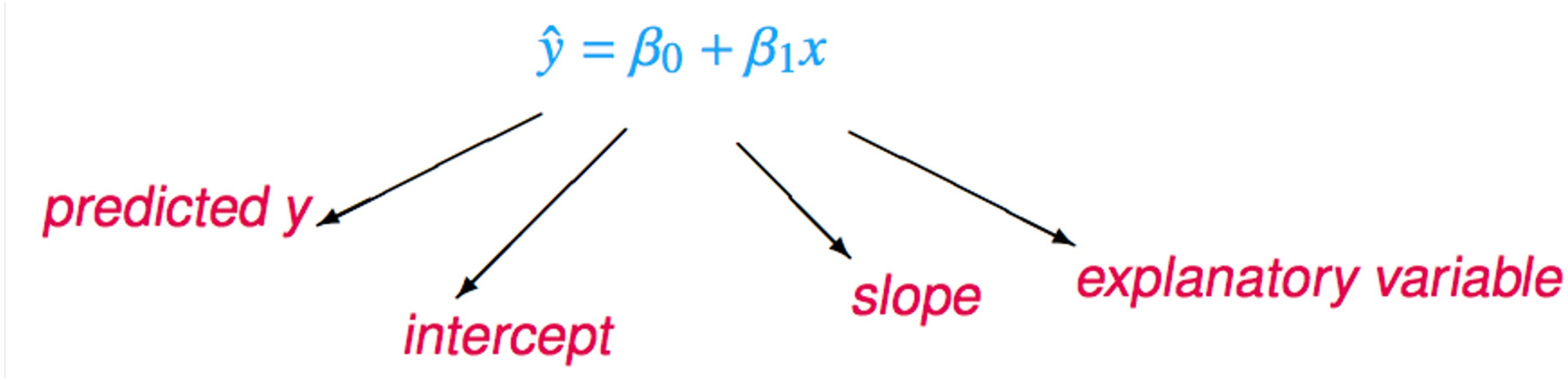
- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software

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
     $$|e_1| + |e_2| + \dots + |e_n|$$
  2. 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

The least squares line



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

Notation:

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

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

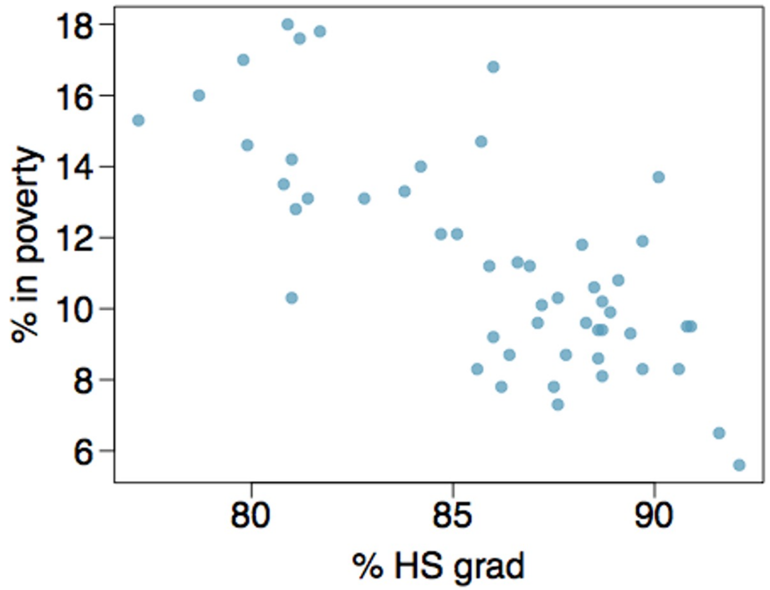
Finding the least squares line

Find b_0, b_1 that minimize the sum of squared residuals

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

To compute the distribution of the estimators $b_0 = \widehat{\beta_0}, b_1 = \widehat{\beta_1}$ we need to make some assumptions.

Given...



	% HS grad	% in poverty
	(x)	(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

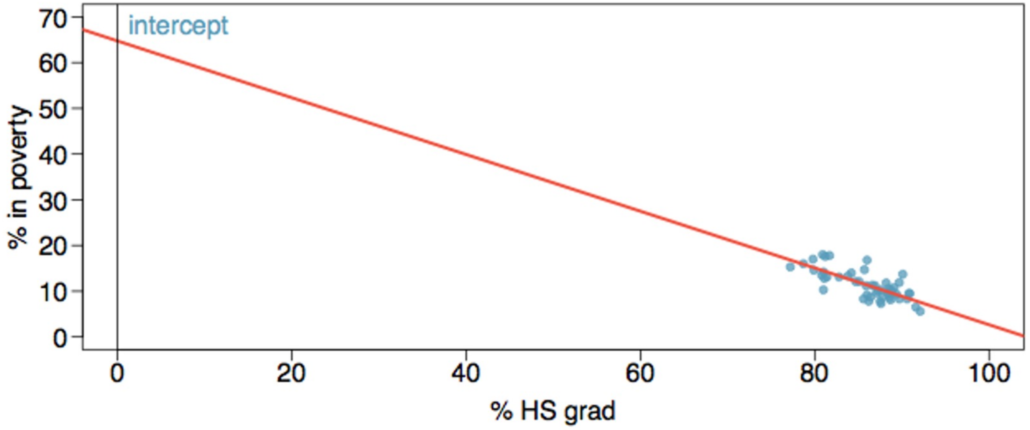
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through ($\bar{x}$, $\bar{y}$).

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

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

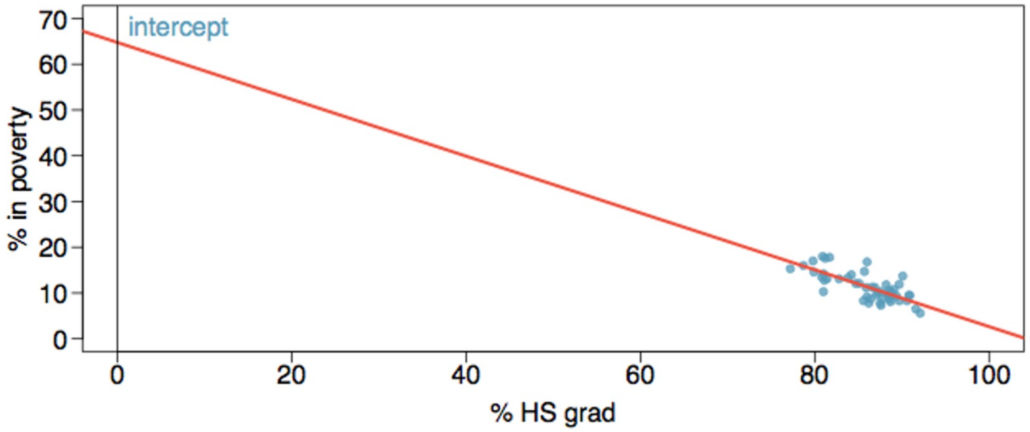
$$b_0 = \bar{y} - b_1 \bar{x}$$



Intercept

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

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



$$b_0 = 11.35 - (-0.62) \times 86.01$$
$$= 64.68$$

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%.

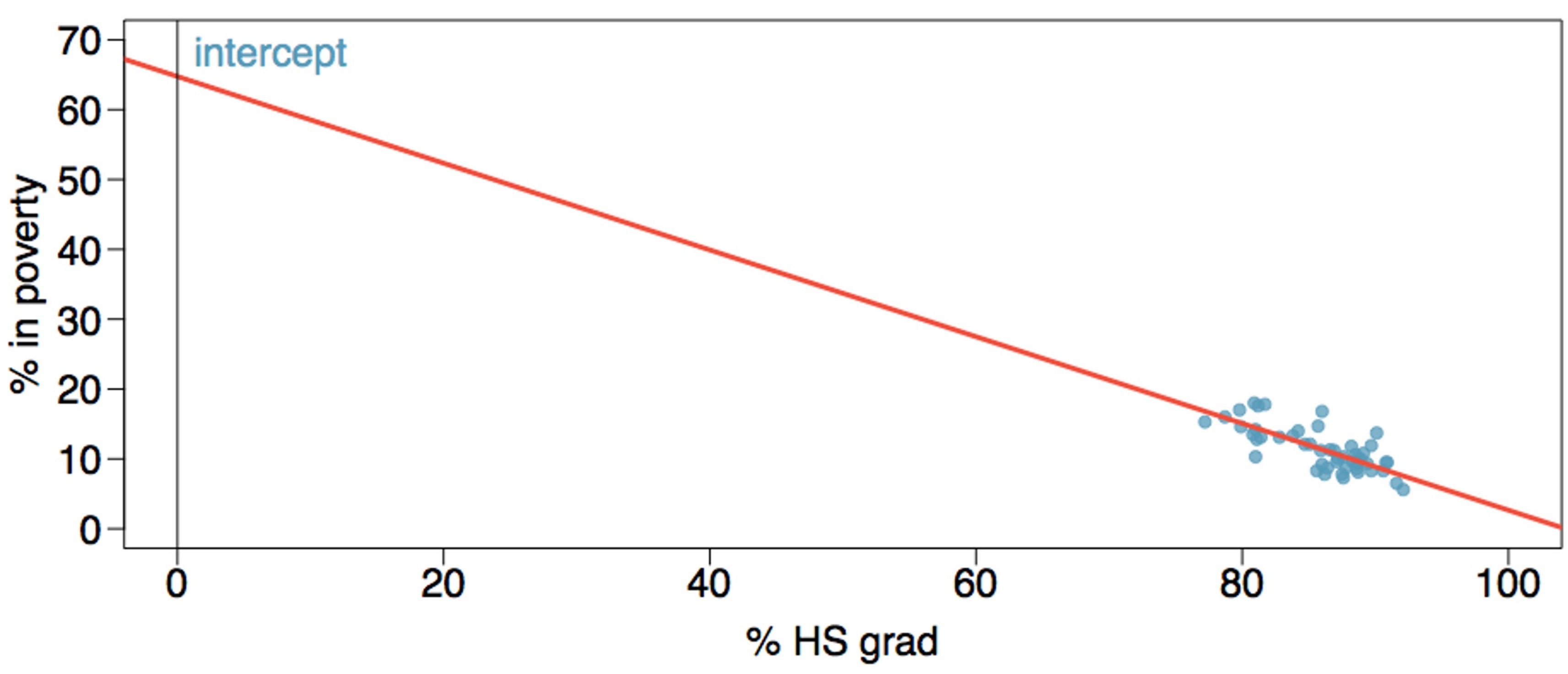
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line.

(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%.

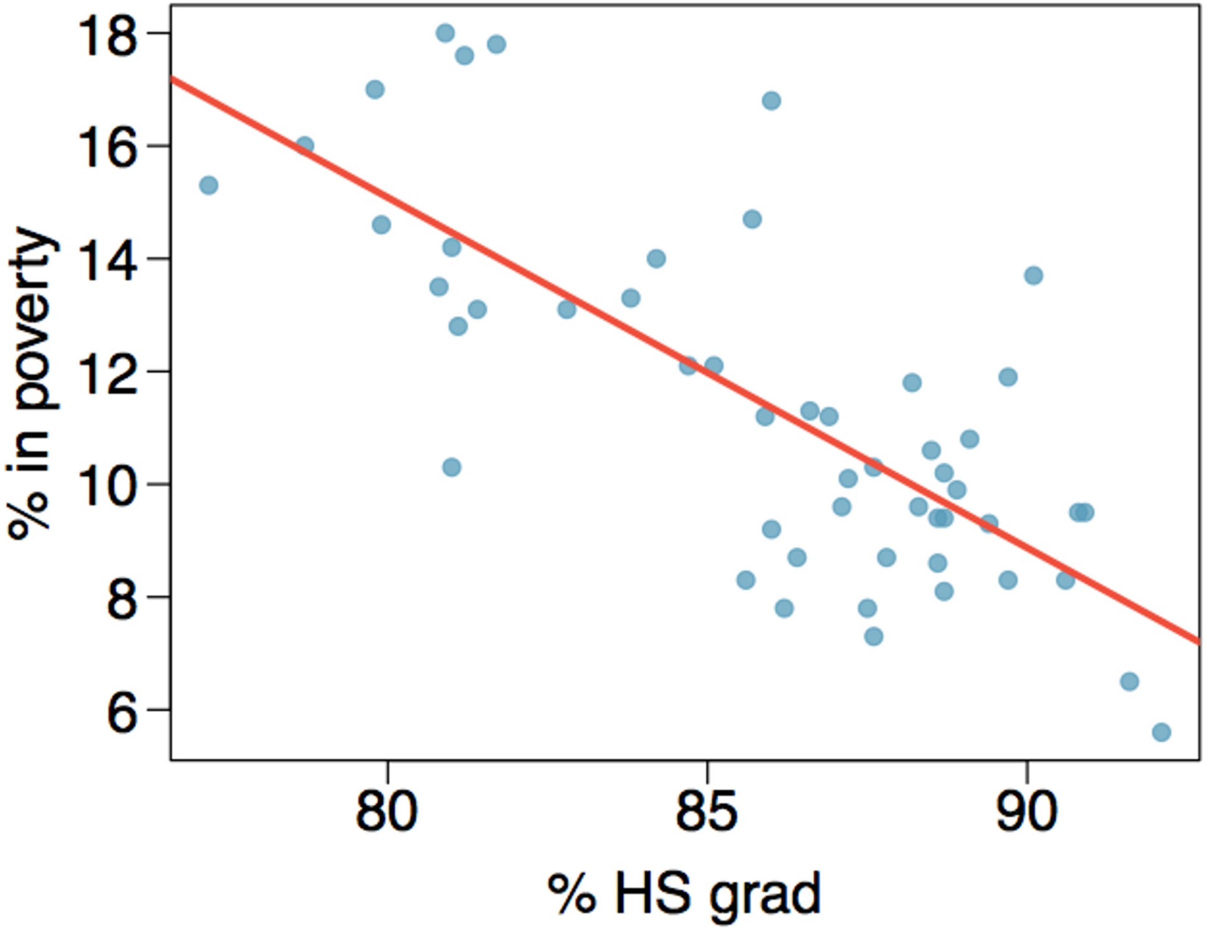
More on the intercept

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

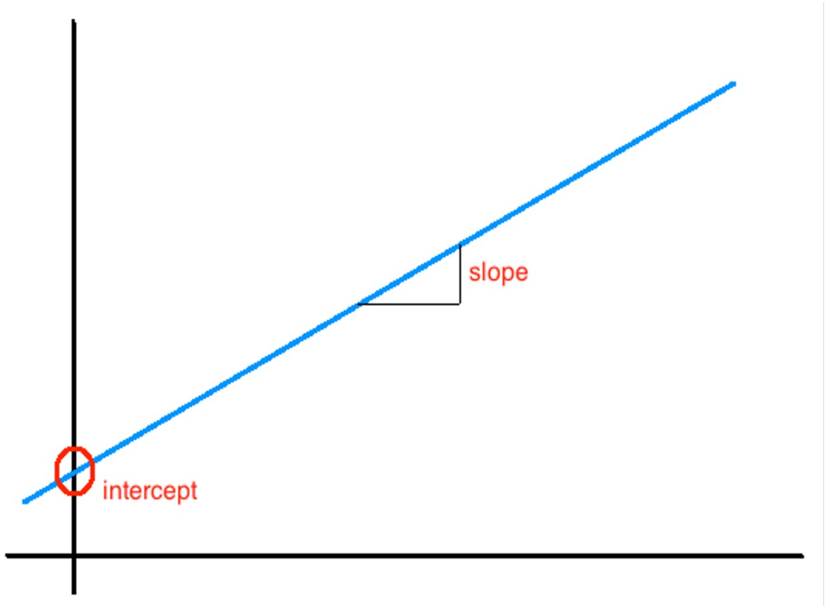
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretation of slope and intercept

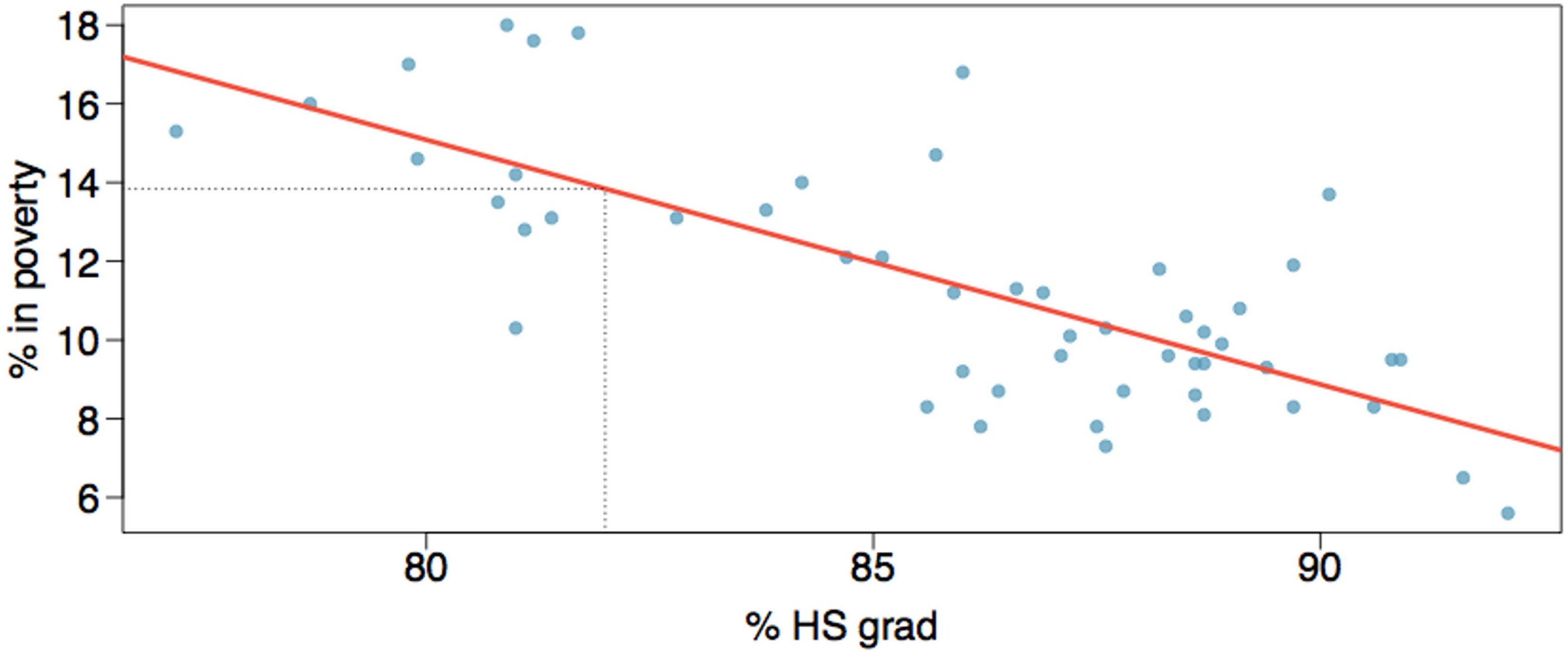
- *Intercept*: When *x* = 0, *y* is expected to equal the intercept.

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



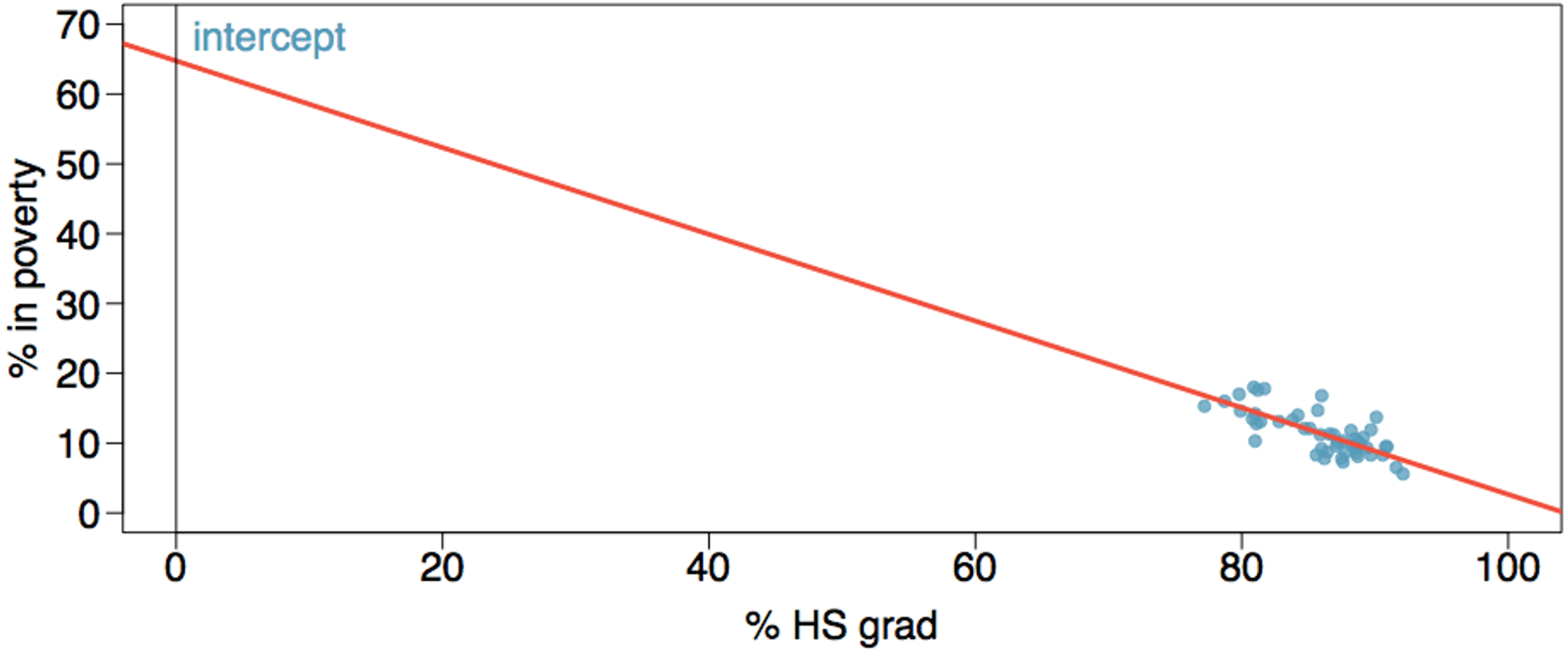
Prediction

- 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.

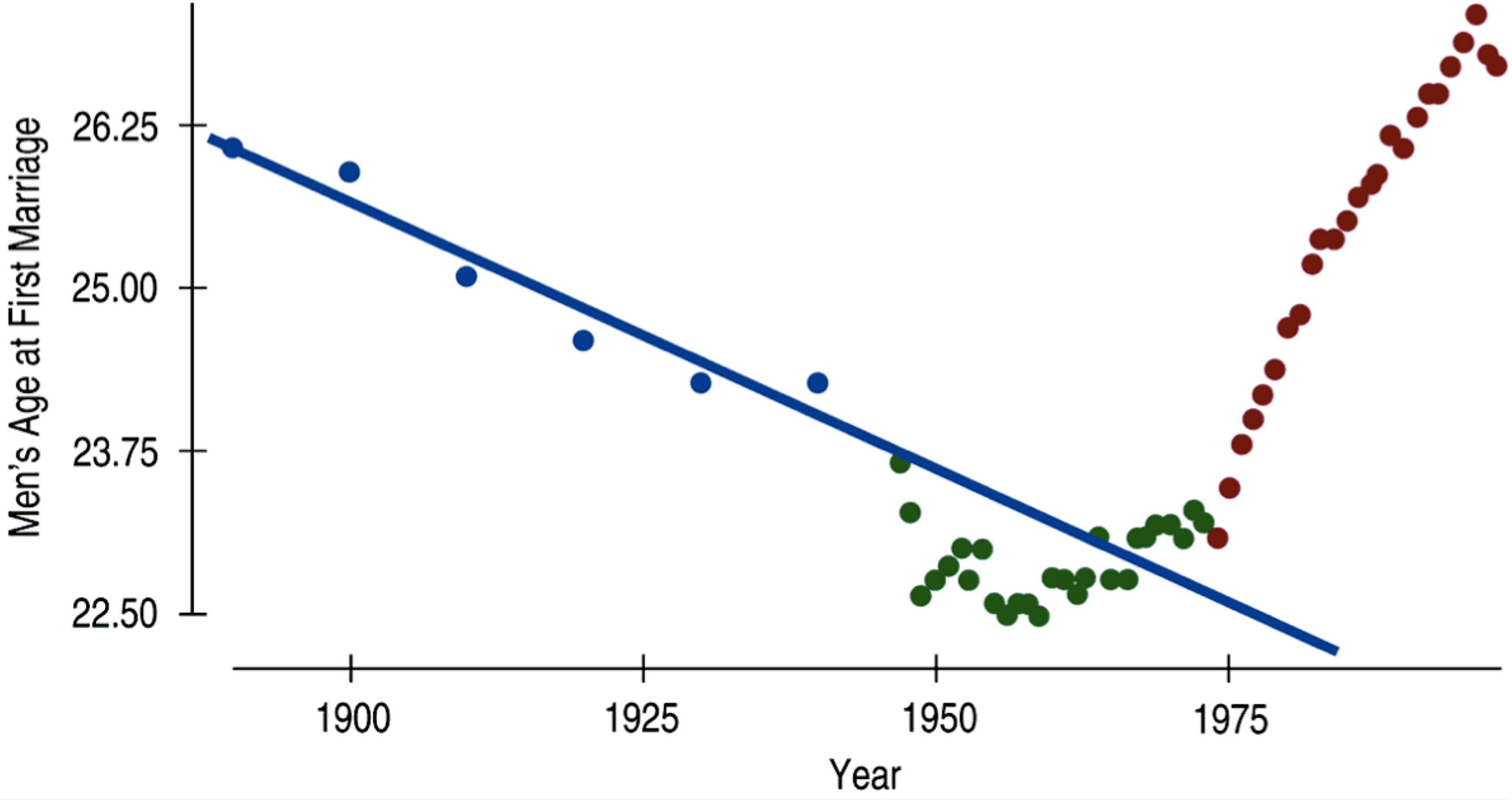


Extrapolation

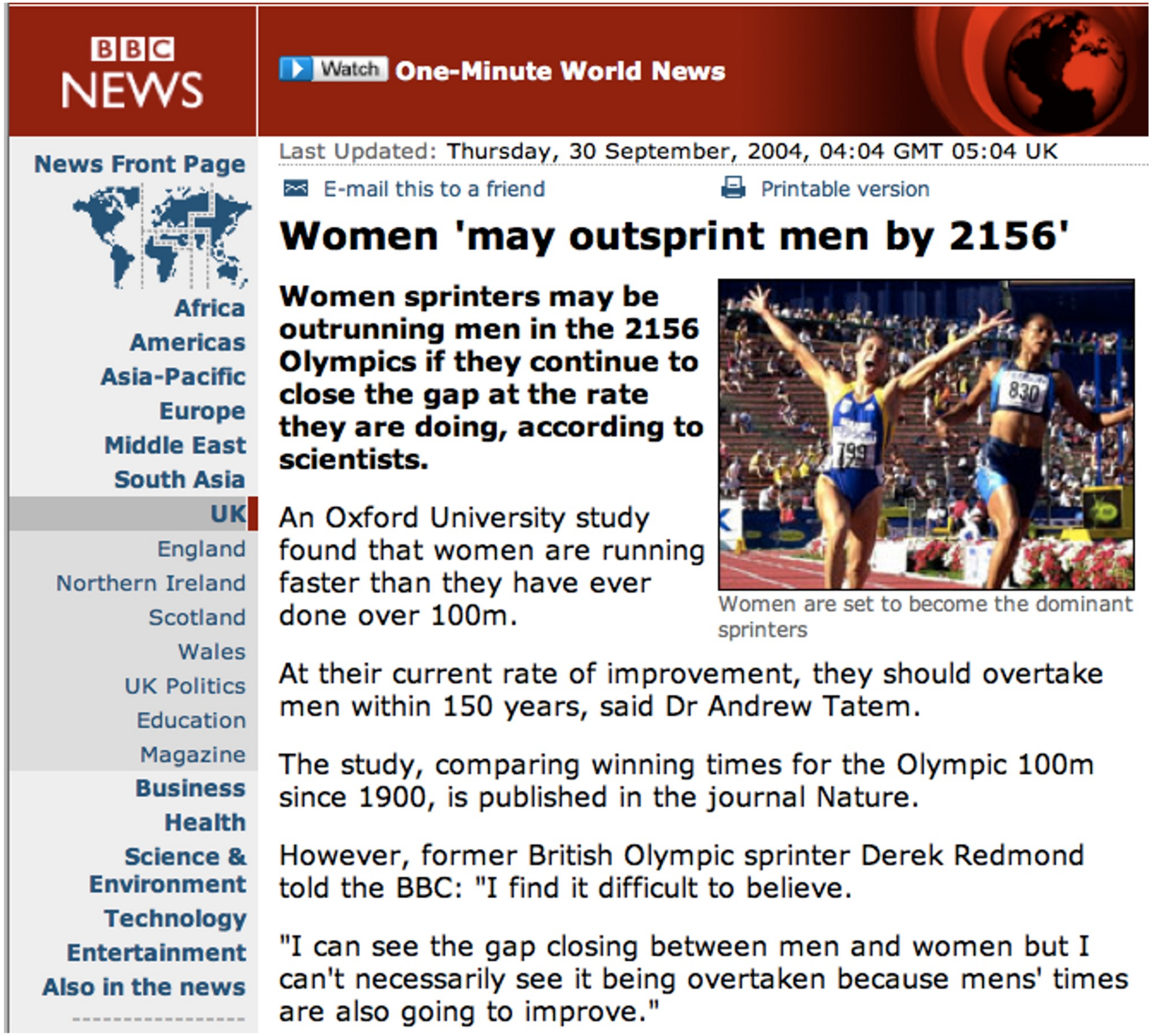
- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation



Examples of extrapolation



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.

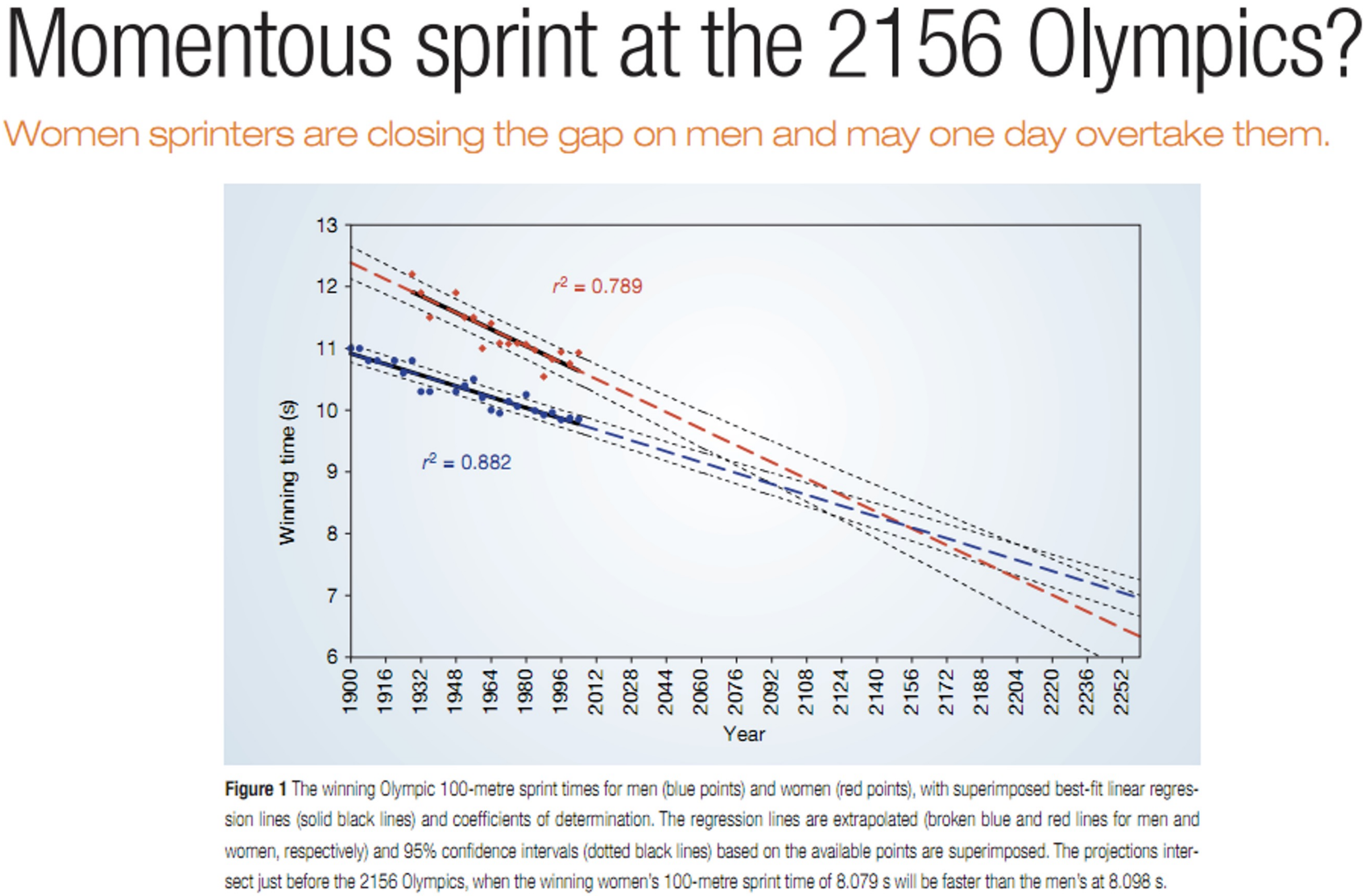
At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

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."

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.