Section 8.1
Line Fitting, Residuals, and Correlation

Stats 7 Summer Session II 2022

Modeling numerical variables

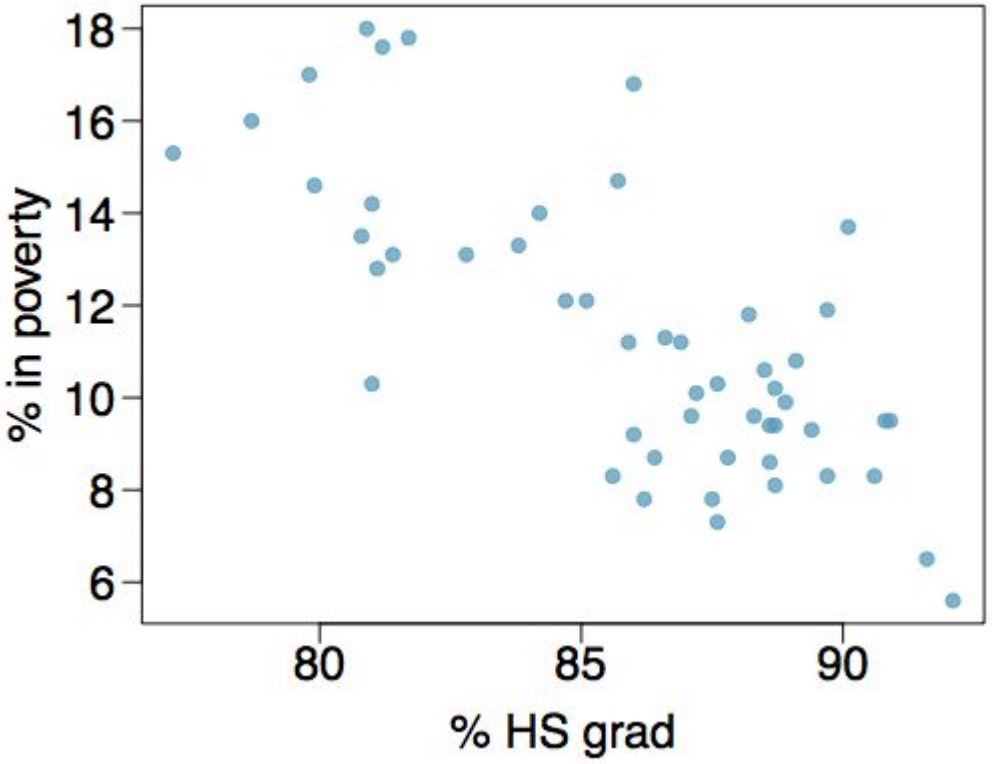
In this unit we will learn to:

- quantify the relationship between two numerical variables
- model numerical response variables using a numerical or categorical explanatory variable
- predict the value of the one variable given the value of the other

We will do this by fitting a line to data and evaluating how well the line represents the trend of the data

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, we'll call this y

Explanatory variable?

% HS grad, we'll call this x

Relationship?

linear, negative, moderately strong

Fitting a line

Recall the formula for a line is

$$y = mx + b \text{ or } y = b + mx,$$

where b is the y-intercept (value of y when x is 0) and m is the slope (i.e. the change in y for a one unit increase in x)

In statistics we will denote b with β_0 and m with β_1, so we have

$$y = \beta_0 + \beta_1 x.$$

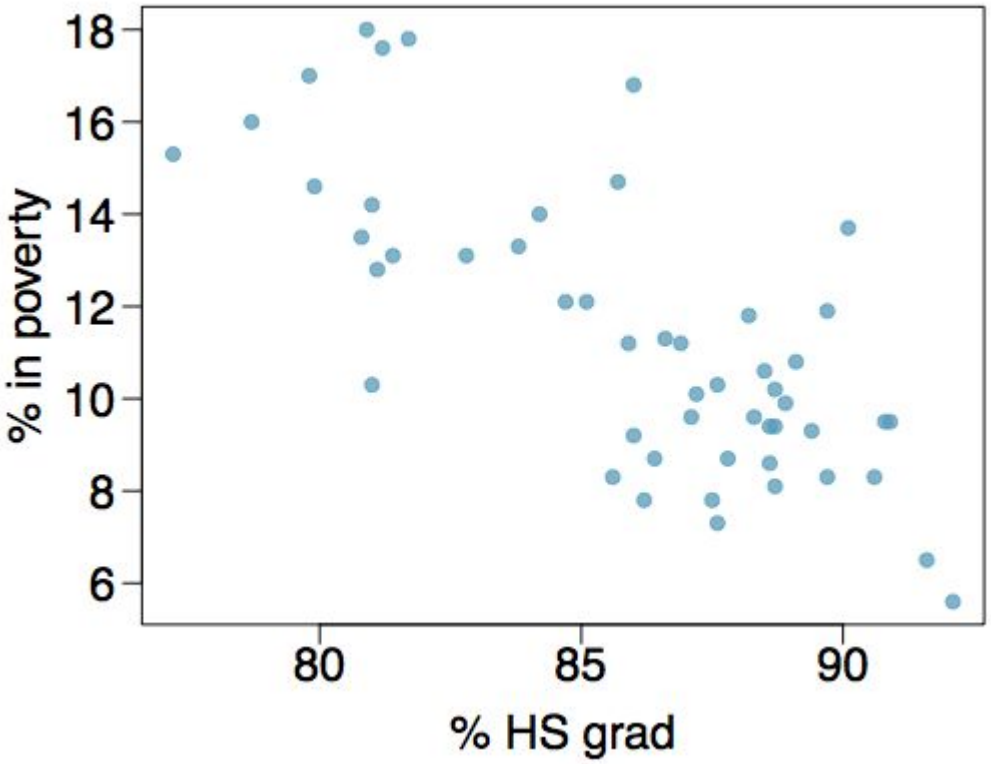
So:

β_0 = value of y when x is 0

β_1 = change in y for a one unit increase in x

Poverty vs. HS graduate rate

Consider if we wanted to fit a line to the data.



What is your guess of the y-intercept, β_0?

It is hard to tell with the x-axis concatenated, but it'll be a high value maybe around 70.

What is your guess of the slope, β_1?

It is negative and not too extreme, maybe close to -1.

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.

So our estimate for β_0 is 64.78 and our estimate for β_1 is -0.62.

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

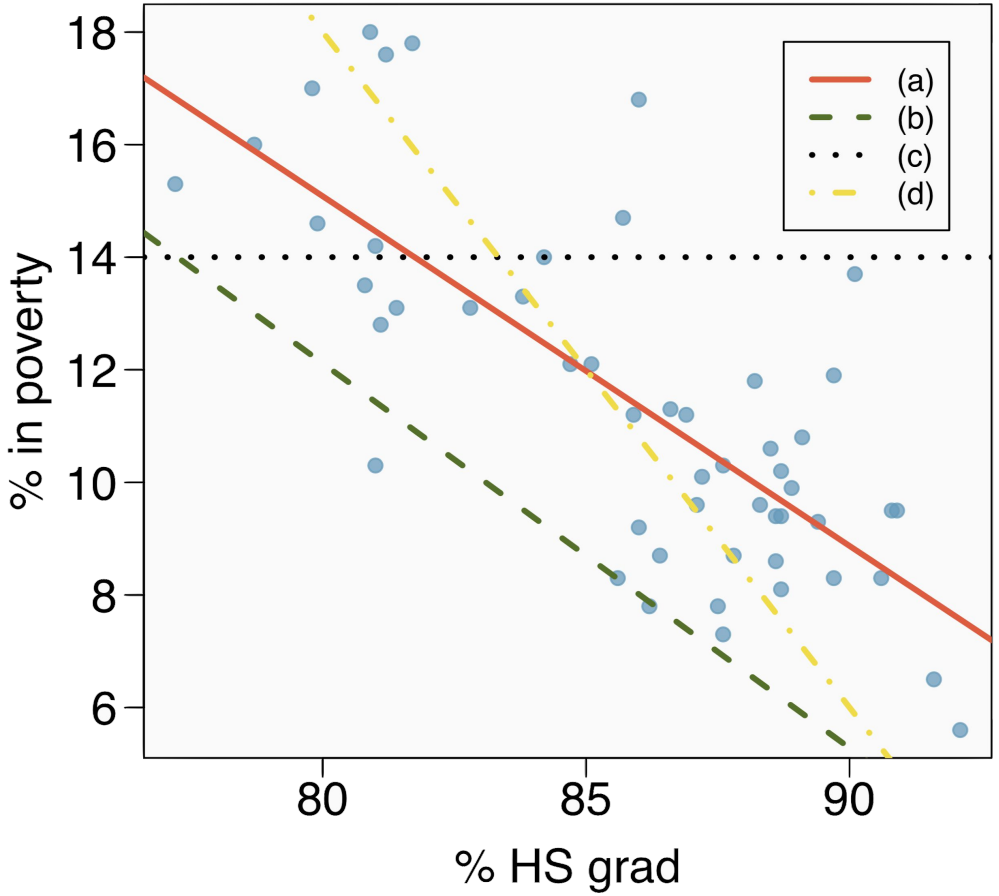
64.78 − 0.62 x 85.1 = 12.018

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)

Note that the data do not lie perfectly along our best-fit line.



Error when fitting a line

Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The values β_0 and β_1 represent the model's parameters, and the error is represented by ε (the Greek letter epsilon), these are unknown values.

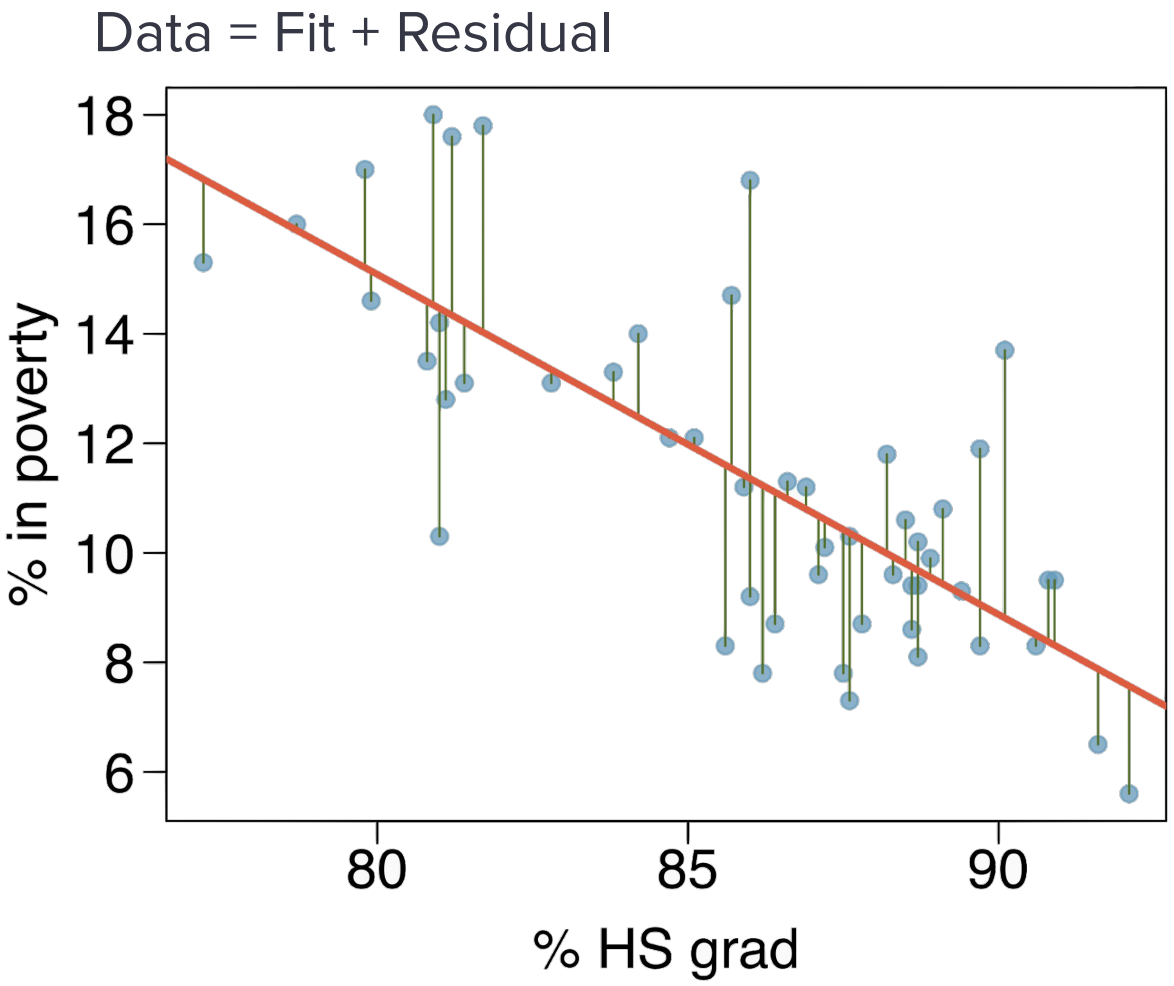
The parameters are estimated using data, and we write their point estimates as b_0 and b_1.

Consider what a perfect linear relationship means: ($\varepsilon = 0$) we know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process.

We often drop the ε term when writing down the model since our main focus is often on the prediction of the average outcome.

Residuals

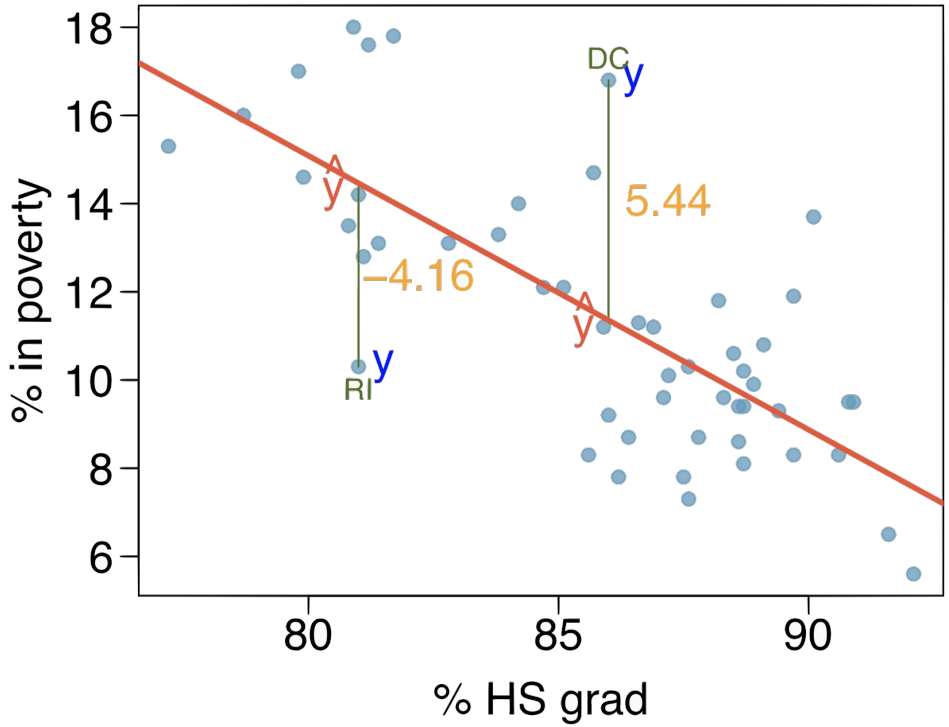
Residuals are the leftovers from the model fit:



Residuals

Residual (e_i) is the difference between the observed (y_i) and predicted ($\hat{y}_i$), for observation i.

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



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

So the residual value of that point is 5.44%.

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

So the residual value of that point is -4.16%.

Residuals

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a residual plot.

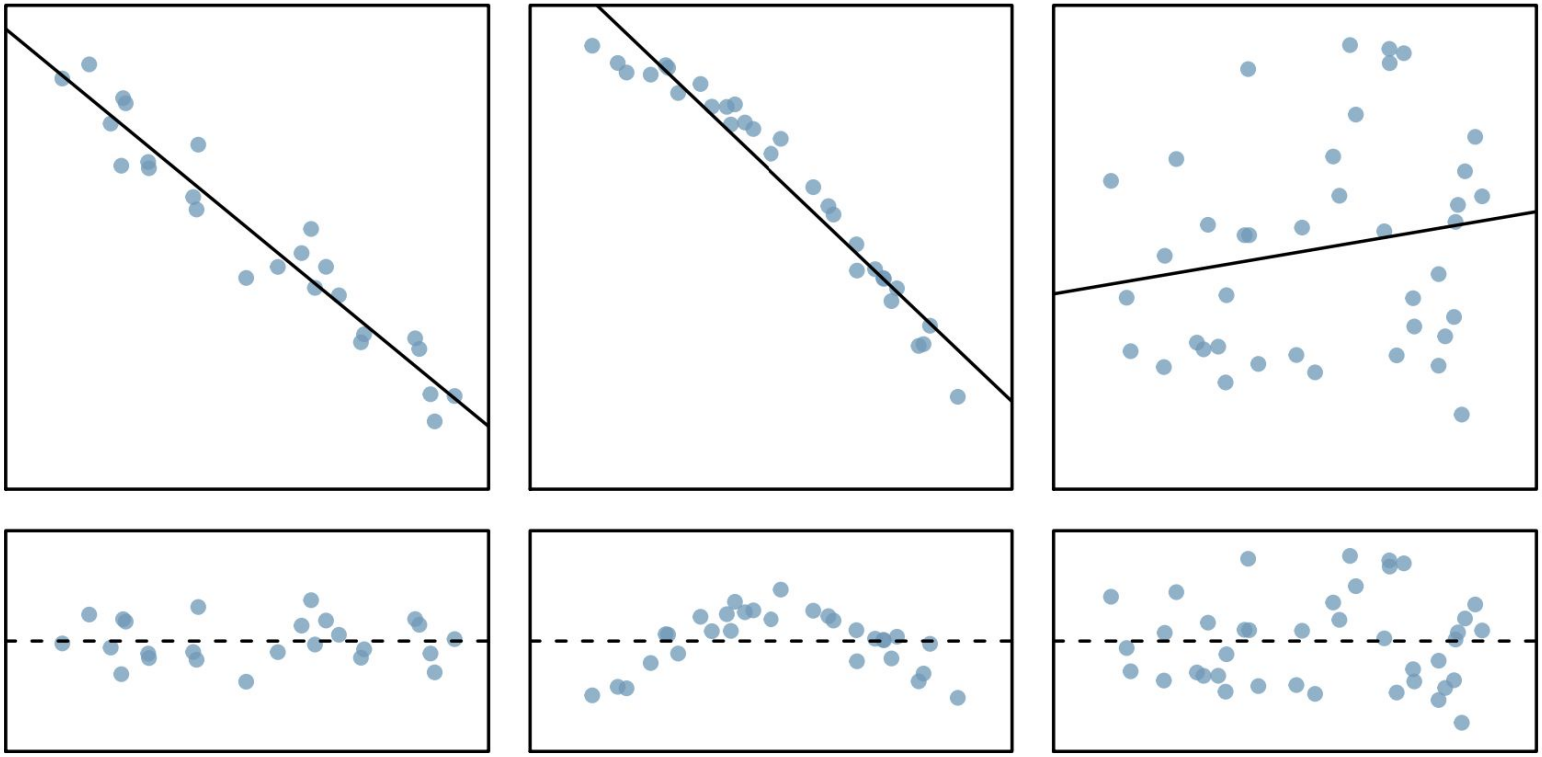
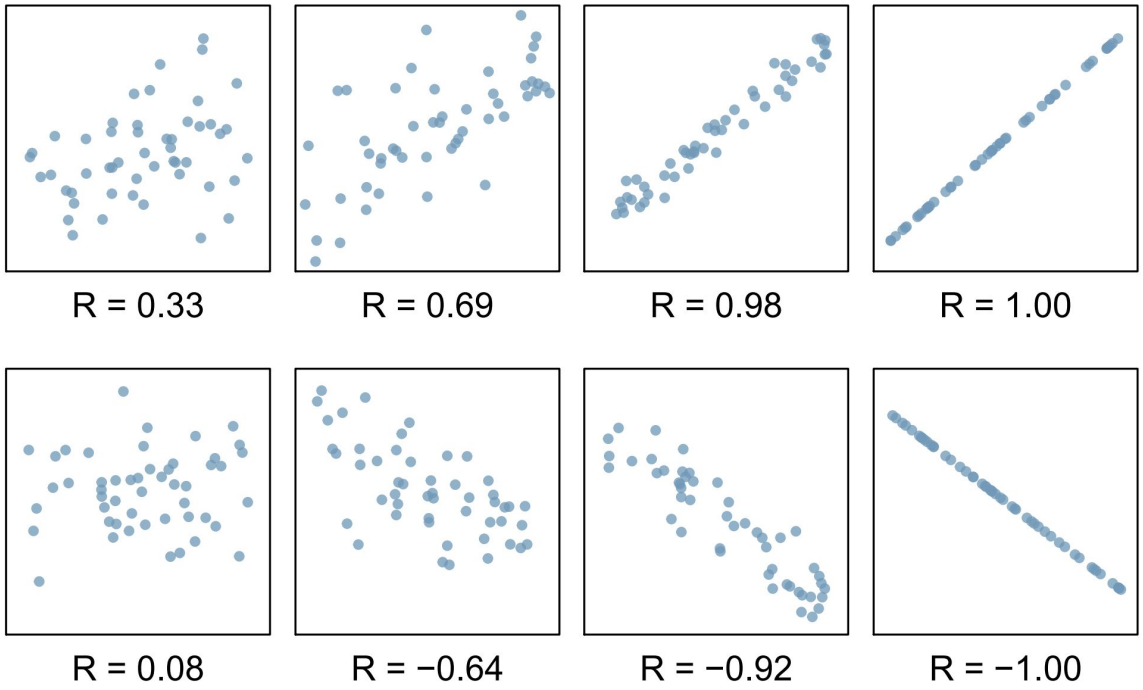


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

Quantifying the relationship

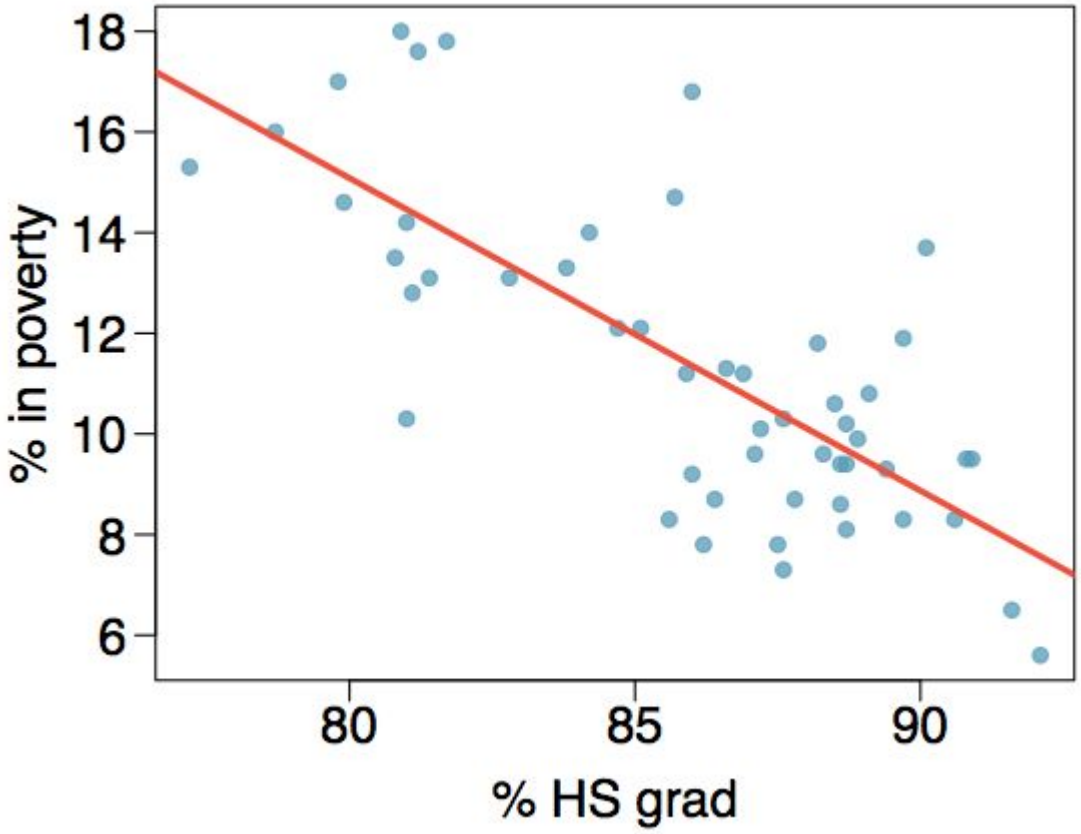
- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.



Guessing the correlation

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

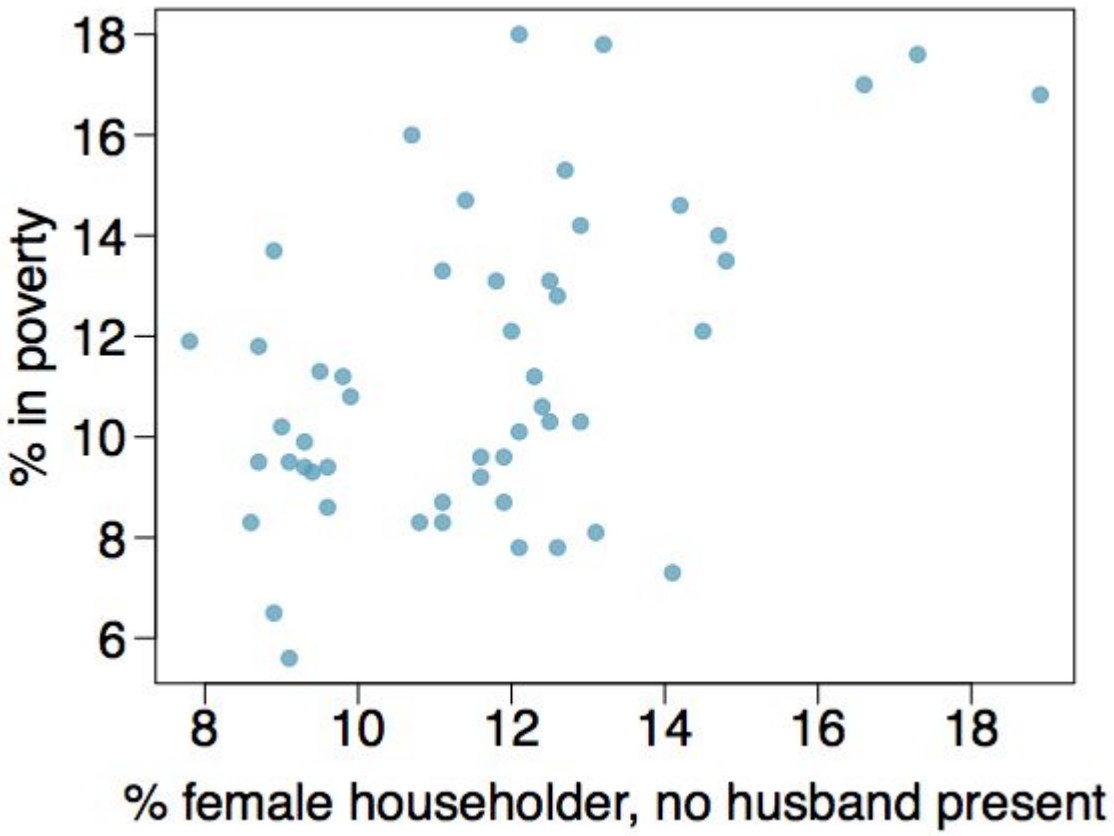
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Guessing the correlation

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

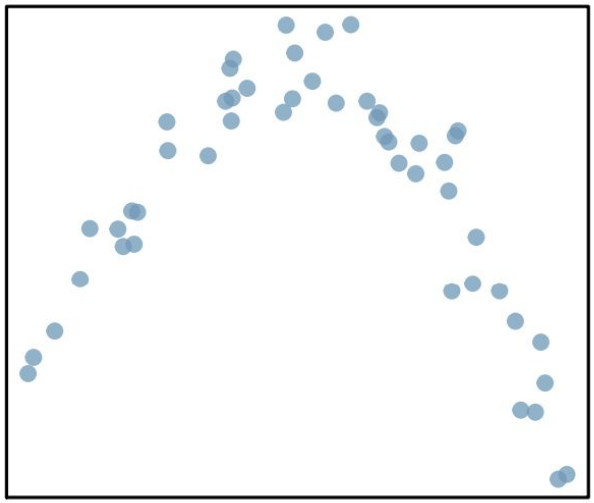
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



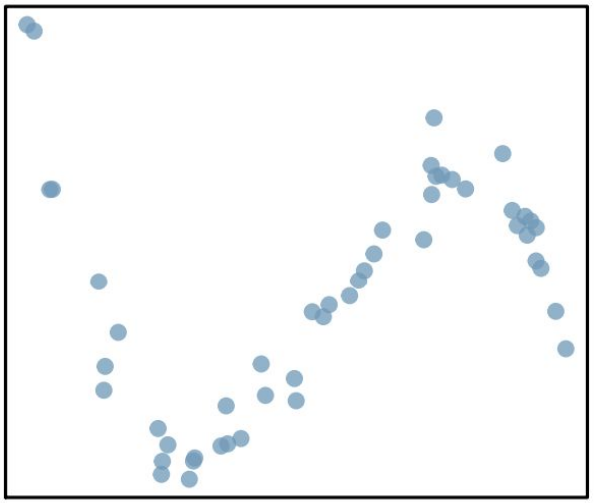
Linear correlation of non linear relationships

The correlation is intended to quantify the strength of a linear trend.

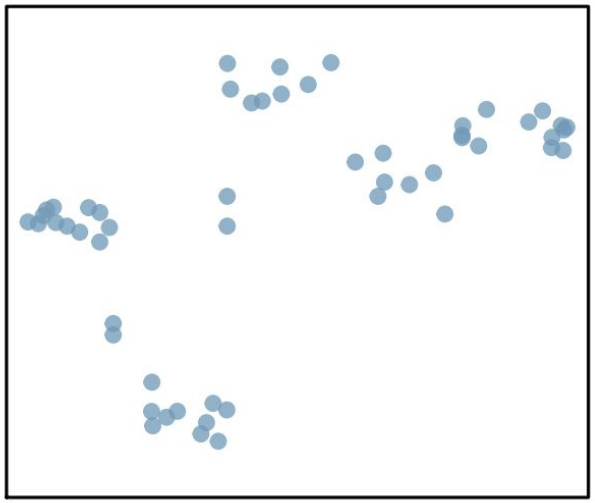
Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship; see three such examples



R = −0.23



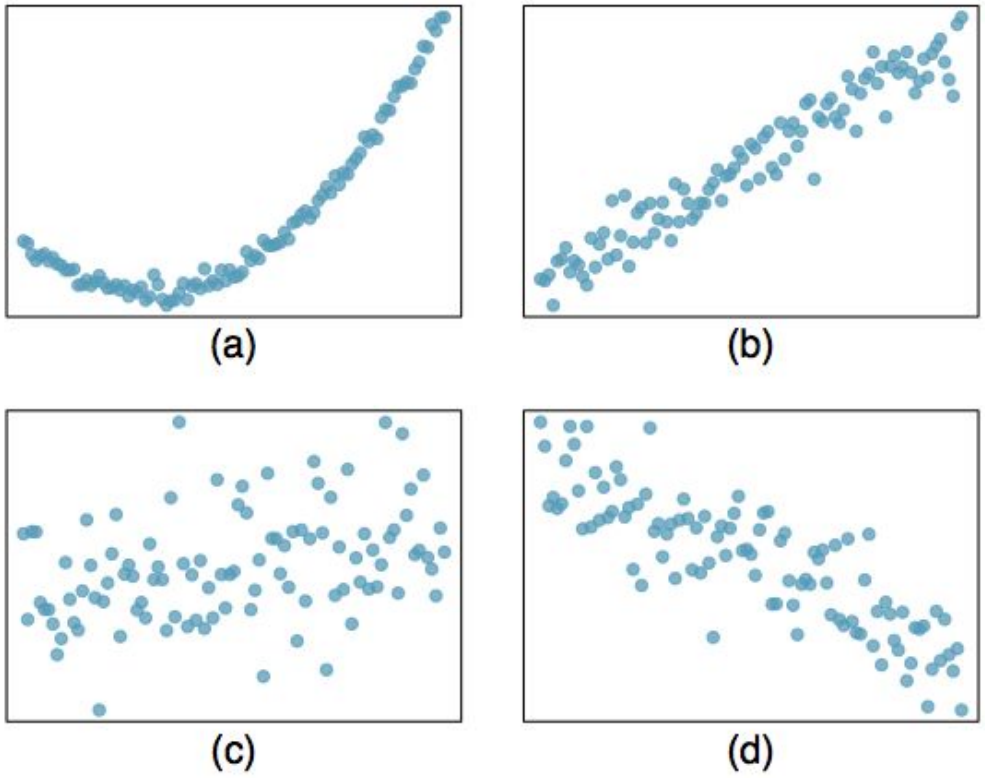
R = 0.31



R = 0.50

Assessing the correlation

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



(b) ➜ correlation means linear association

Derivative of slides developed by Mine Çetinkaya-Rundel of OpenIntro.
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.
The slides may be copied, edited, and/or shared via the
CC BY-SA license