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.

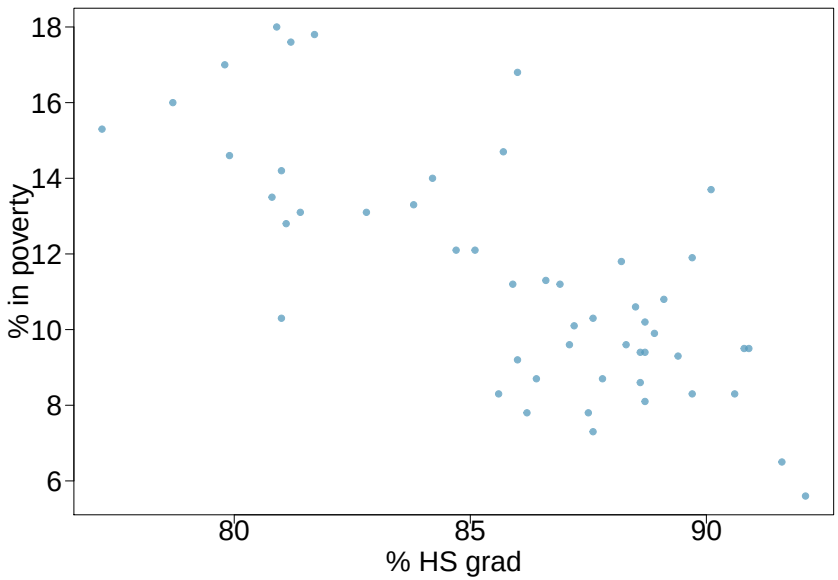
Line fitting, residuls, and correlation

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.

Powerty vs. HS graduate rate

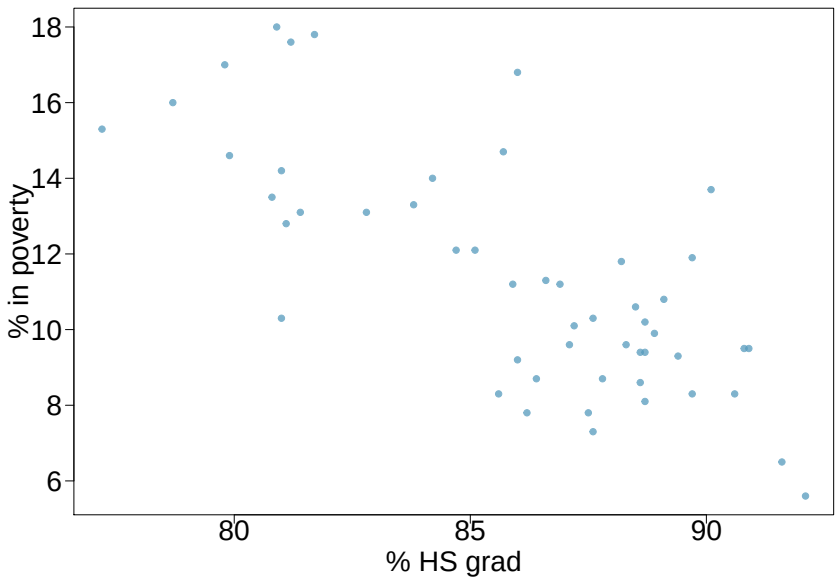
The **scatterplot** below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

Powerty 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).

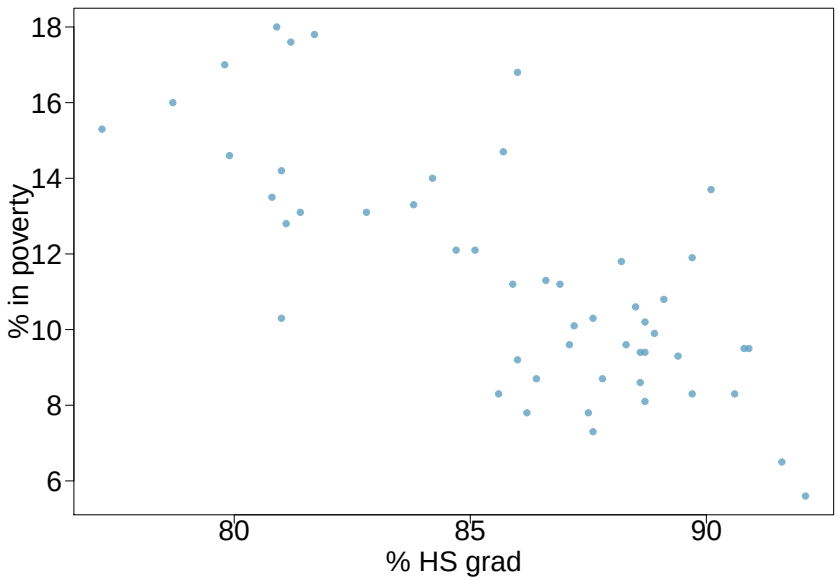


Response variable?

% in poverty

Powerty 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).



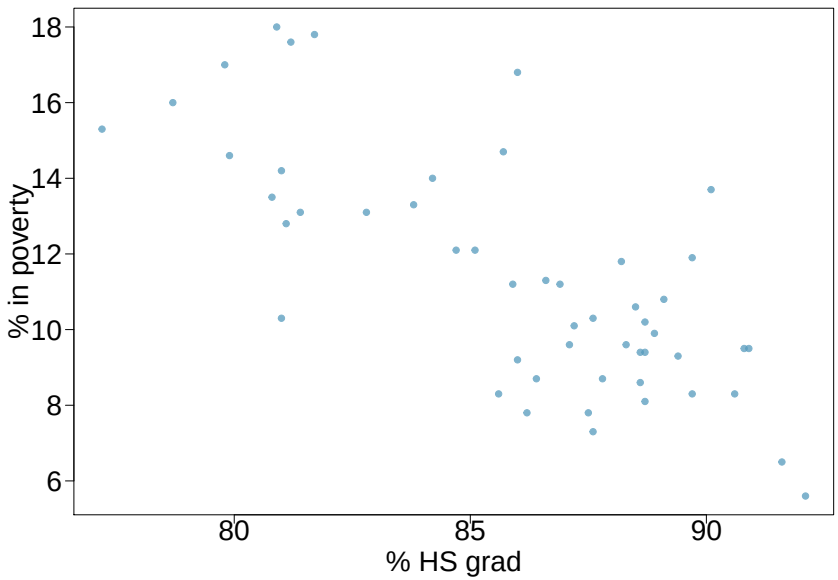
Response variable?

% in poverty

Explanatory variable?

Powerty 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?

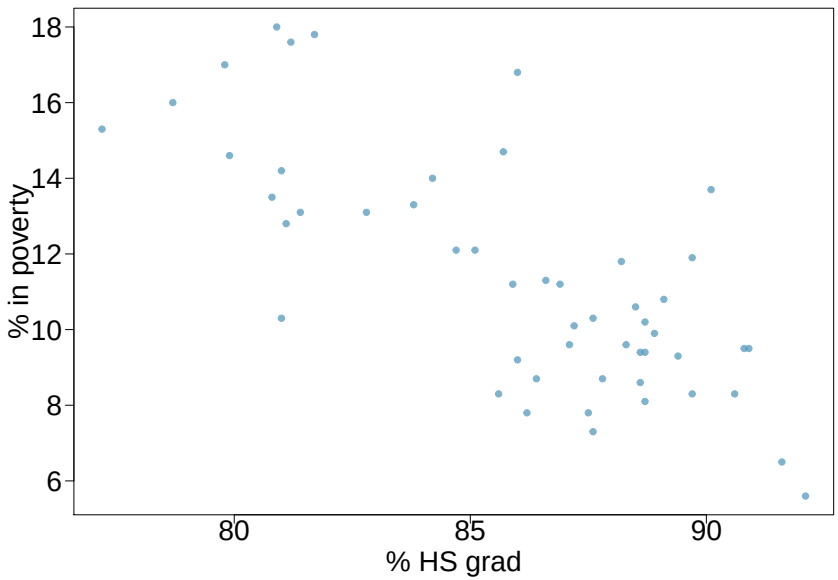
% in poverty

Explanatory variable?

% HS grad

Powerty 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

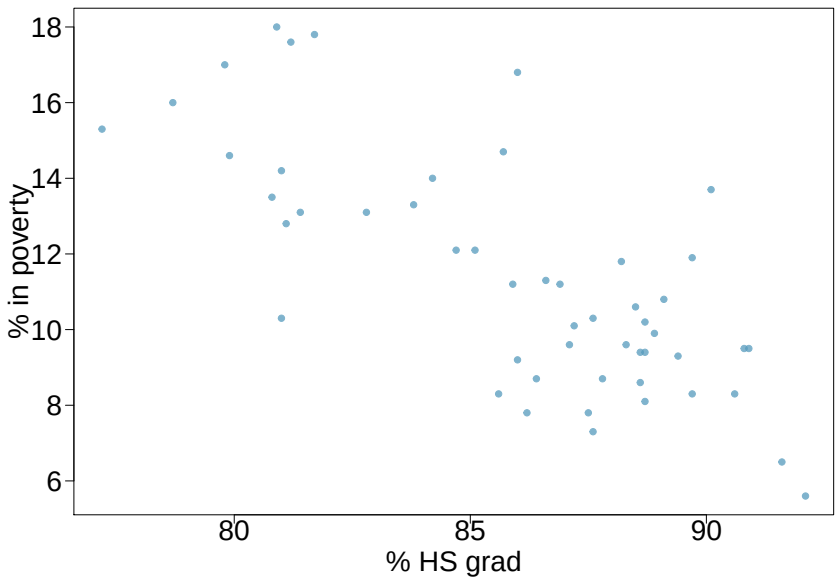
Explanatory variable?

% HS grad

Relationship?

Powerty 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

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

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

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

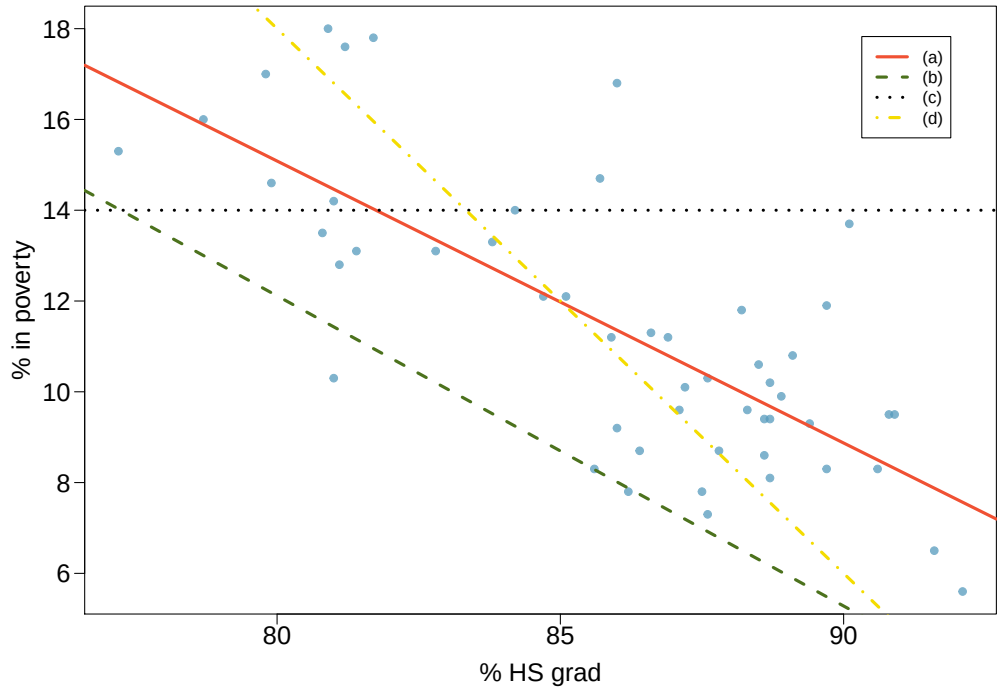
The high school graduate rate in Georgia is 85.1%. What poverty level does the model predict for this state?

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

$$64.78 - 0.62 \times 85.1 = 12.018$$

Eyeballing the line

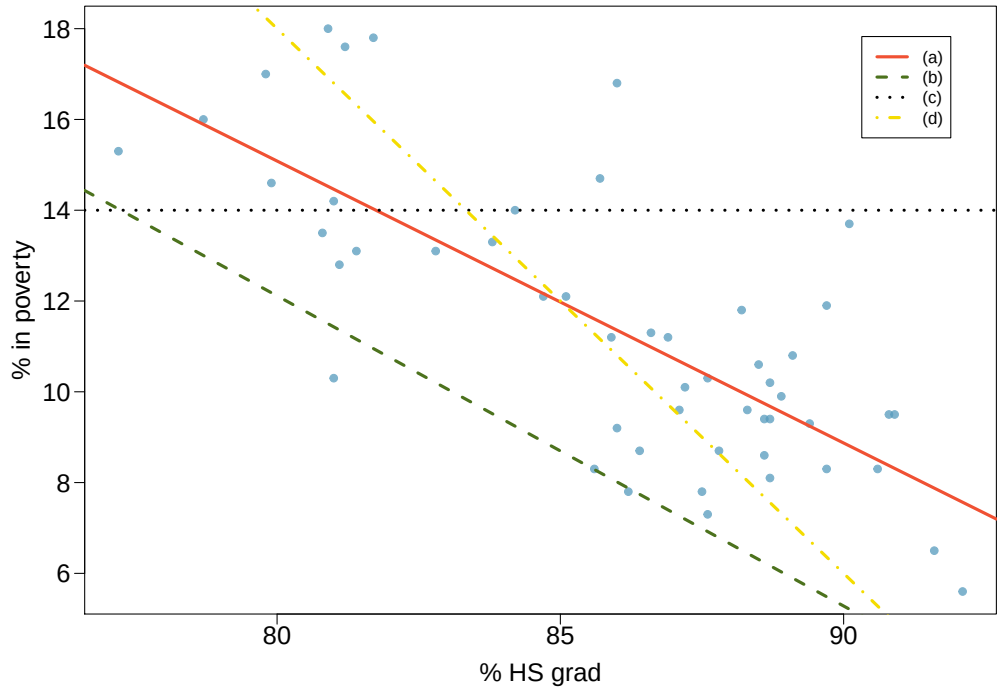
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one



Eyeballing the line

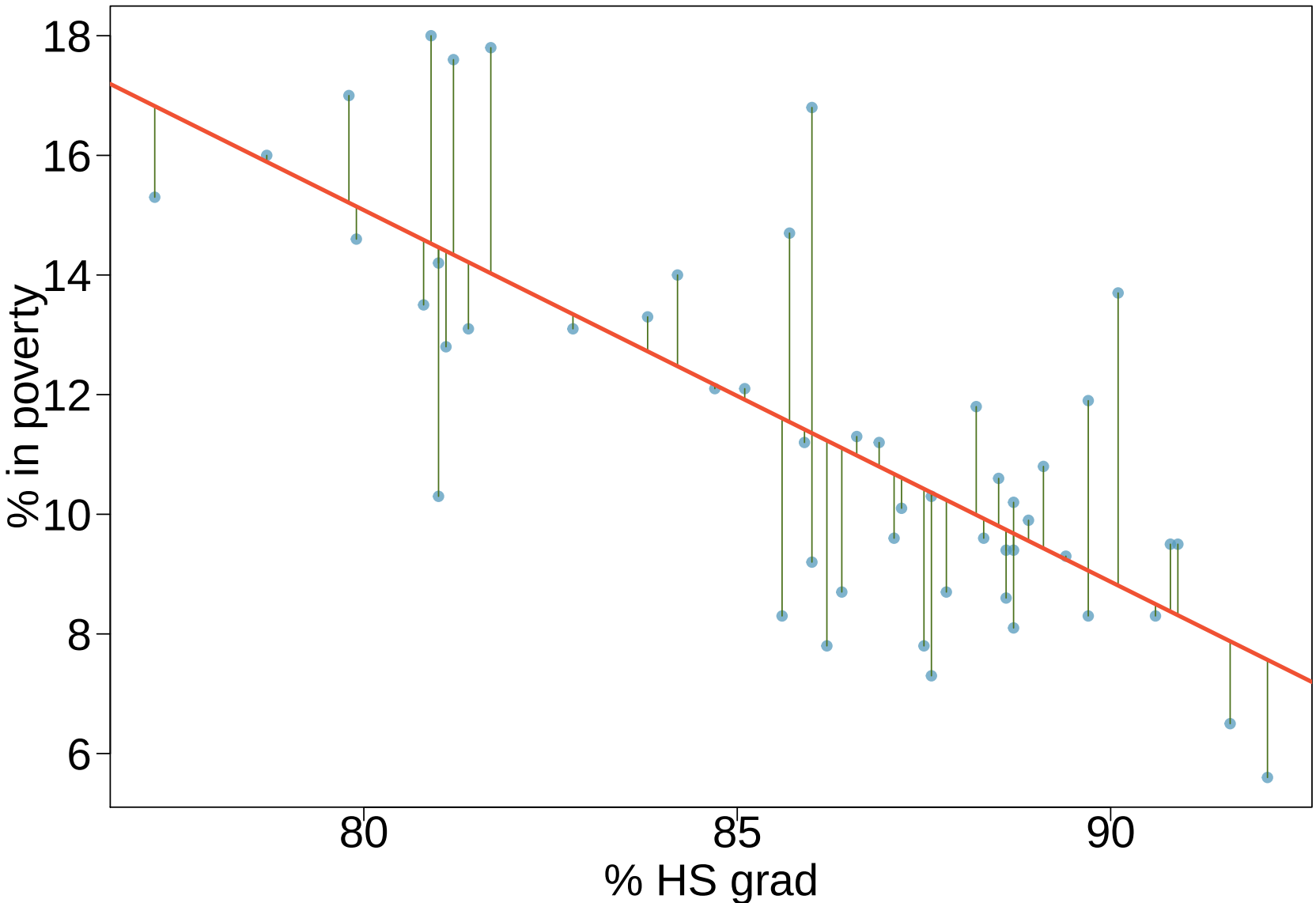
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one

Answer: (a) Solid Red Line



Residuals

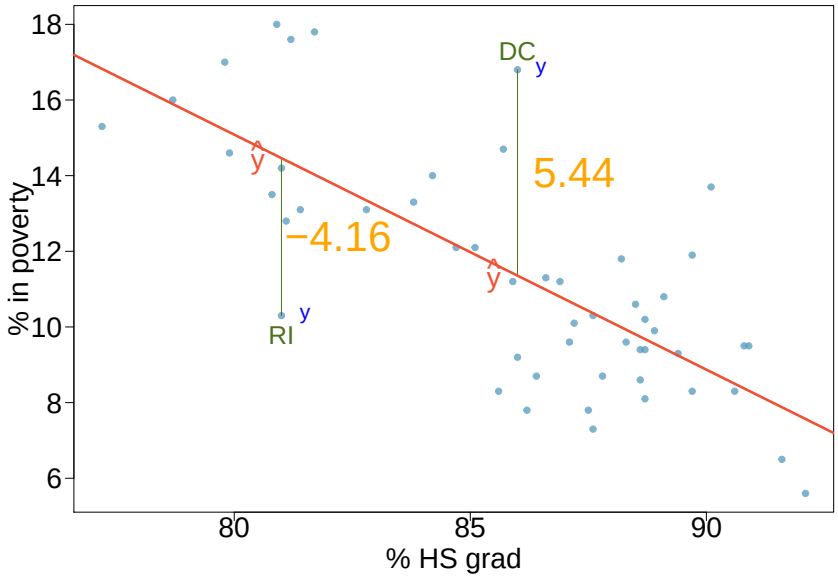
Residuals are the leftovers from the model fit: Data = Fit + Residual



Residuals

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

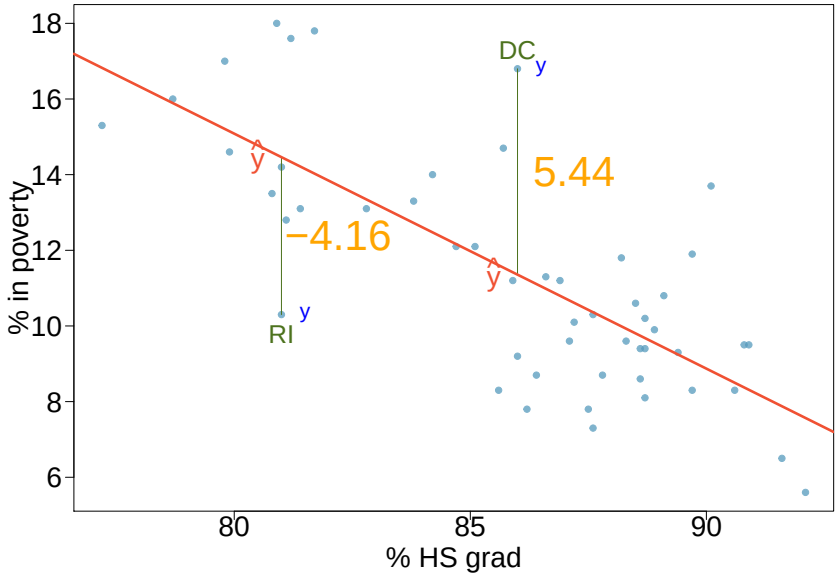
$$e_i = y_i - \hat{y}_i$$



Residuals

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

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

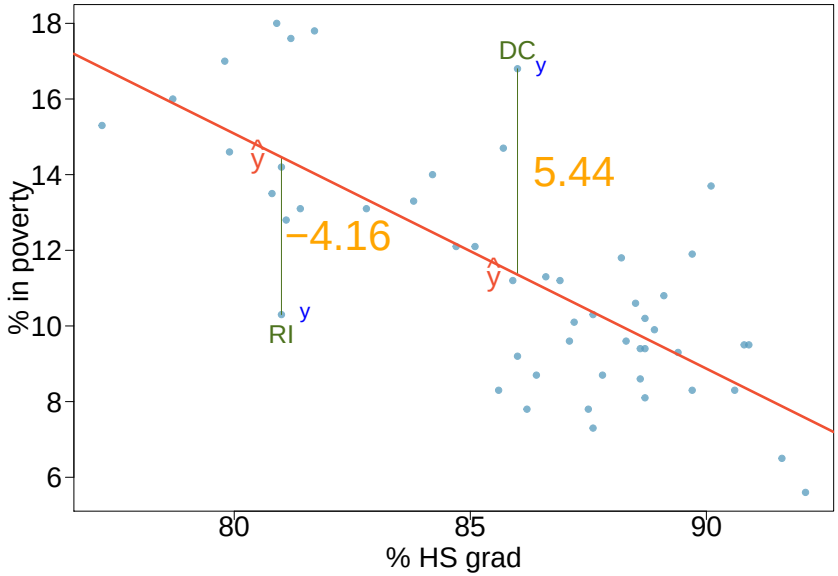


- % living in poverty in DC is 5.44% more than predicted.

Residuals

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.

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).

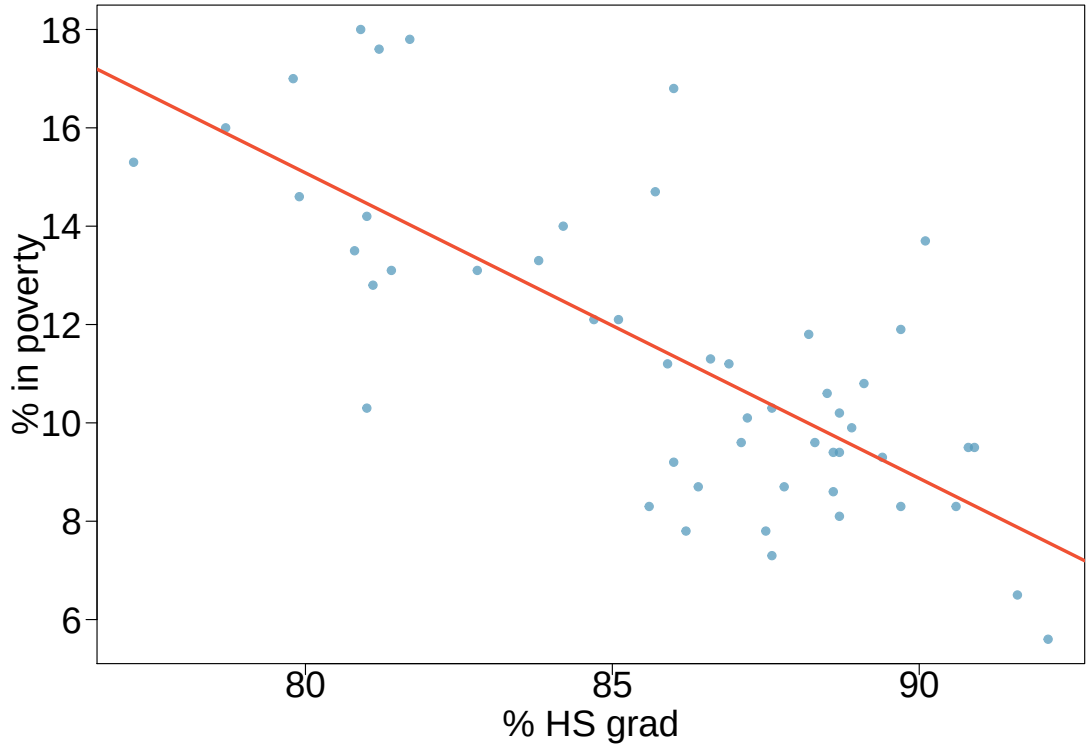
Quantifying the relationship

- **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.

Guessting the correlation

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

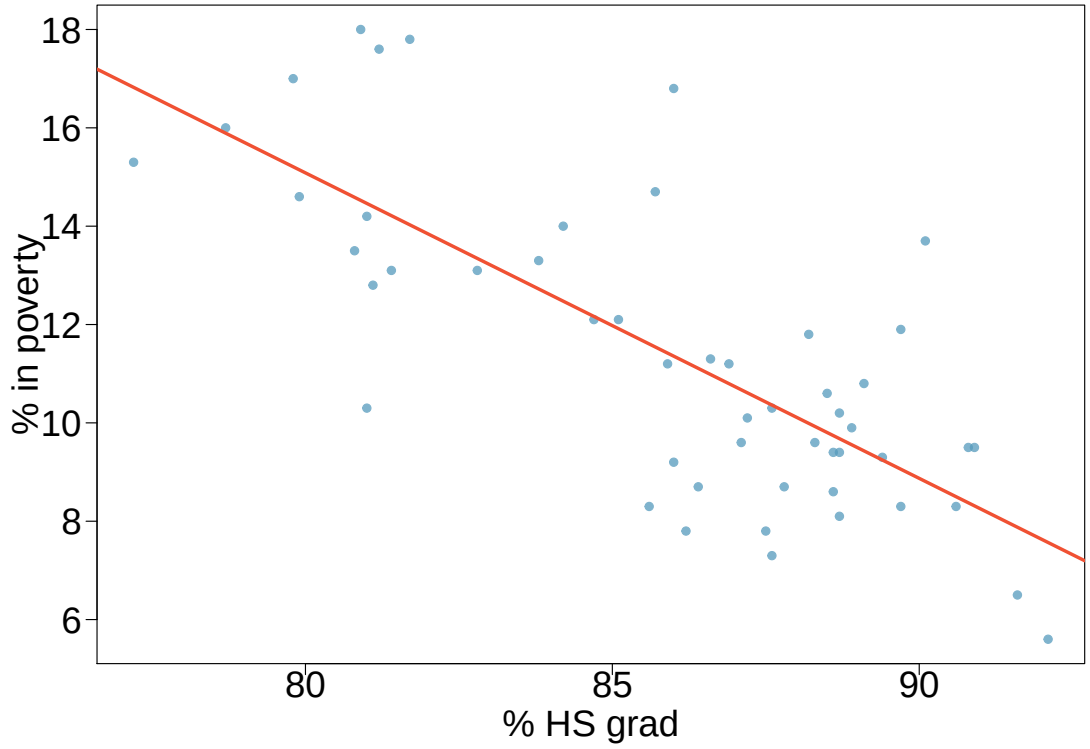
A) 0.6
B) -0.75
C) -0.1
D) 0.02
C) -1.5



Guessting the correlation

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

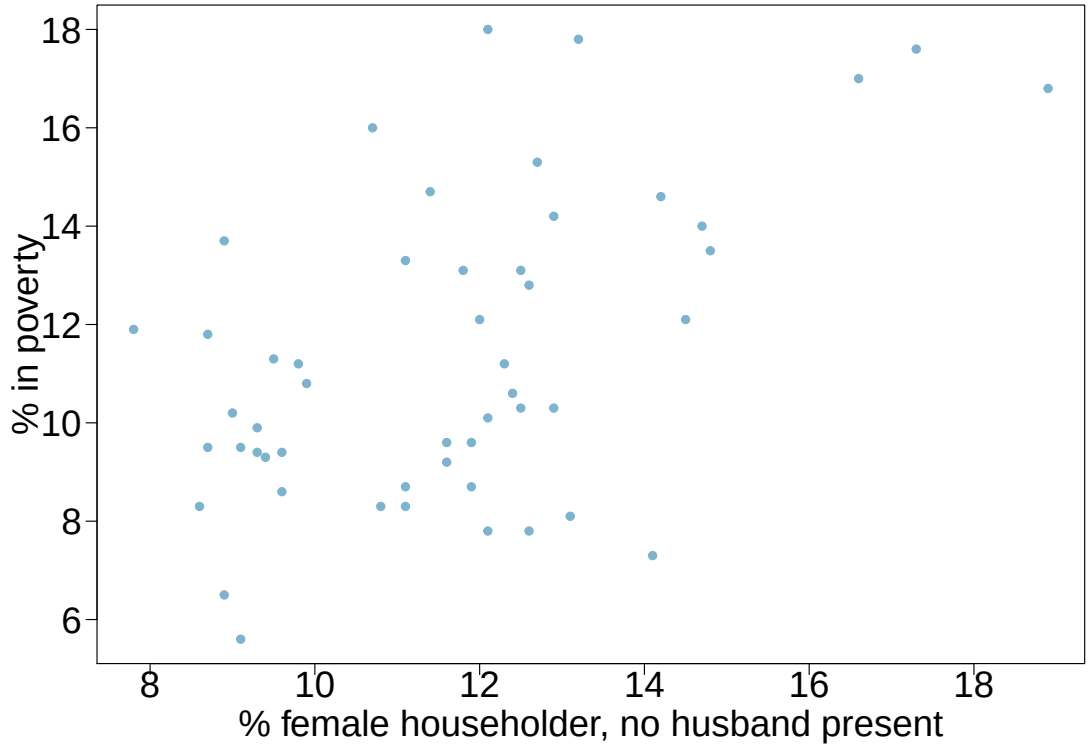
A) 0.6
B) -0.75
C) -0.1
D) 0.02
C) -1.5



Guessting the correlation

Which of the following is the best guess for the correlation between % in poverty and % female householder, no husband present?

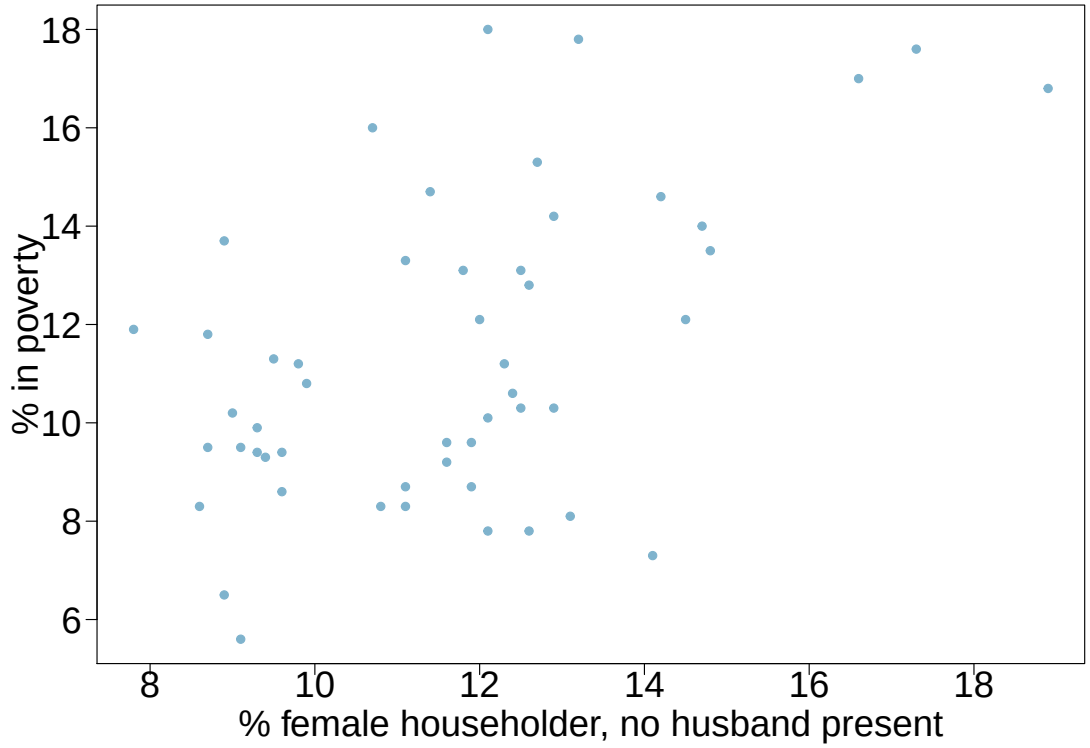
A) 0.1
B) -0.6
C) -0.4
D) 0.9
C) 0.5



Guessting the correlation

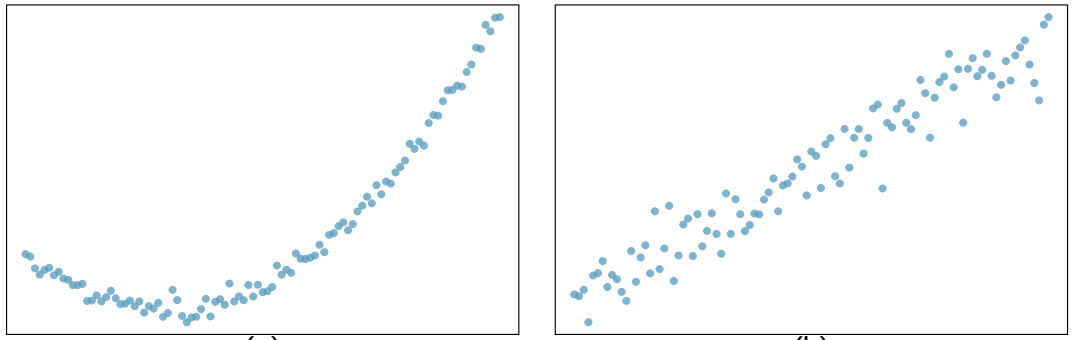
Which of the following is the best guess for the correlation between % in poverty and % female householder, no husband present?

A) 0.1
B) -0.6
C) -0.4
D) 0.9
C) 0.5



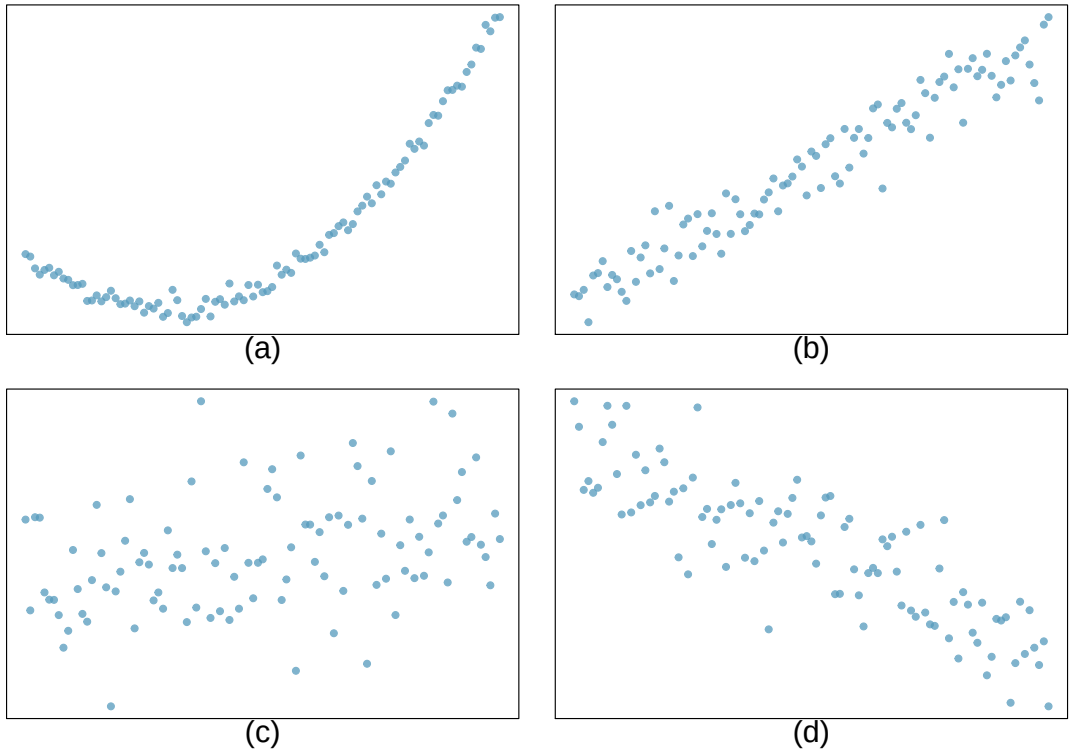
Assessing the correlation

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



Assessing the correlation

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



(b) $\rightarrow$correlation means linear association