Lecture 6: Linear Regression

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

3/15/2016

Overview

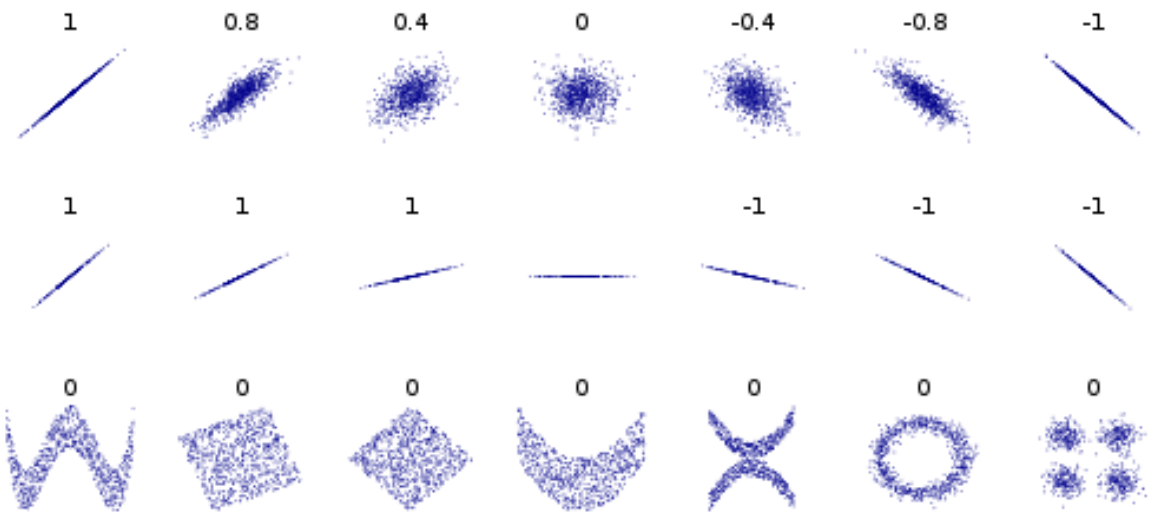
1. Fitting a line to data
2. Residuals
3. Interpretation
   - Conditions for the least squares line
   - R^2
4. Evaluating Error

Correlation

Correlation Coefficient

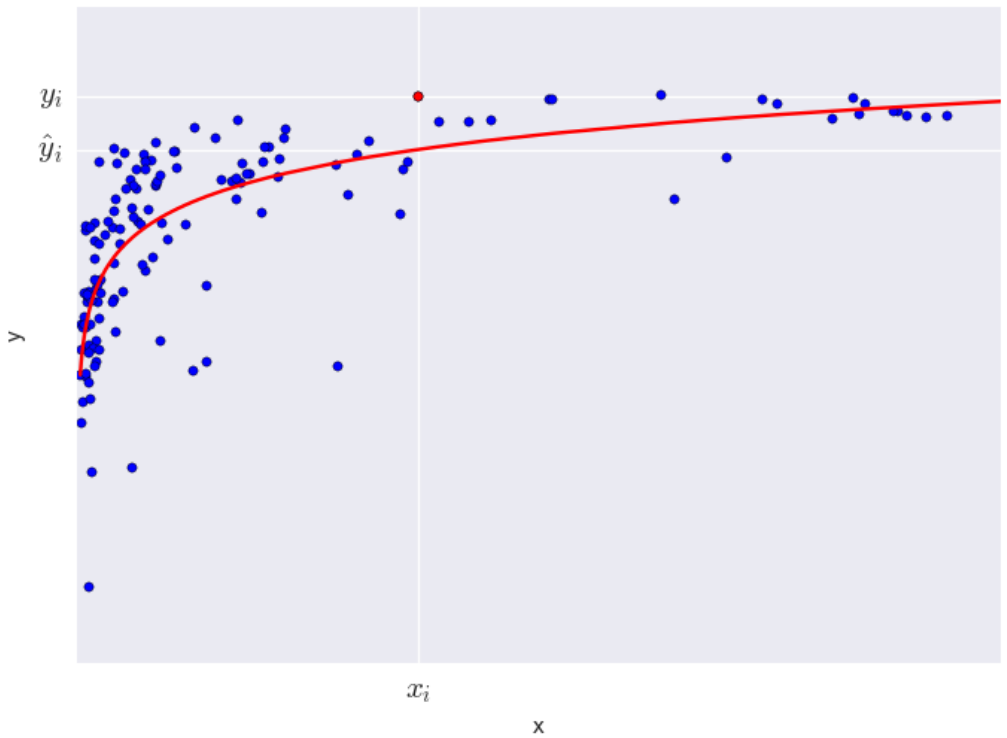
Also known as Pearson's [product-moment] coefficient measures the linear correlation between two [numerical] random variables X and Y.

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$



By DenisBoigelot, CC0

Regression Model



model: $f(x)$

$$y_i \approx f(x_i)$$

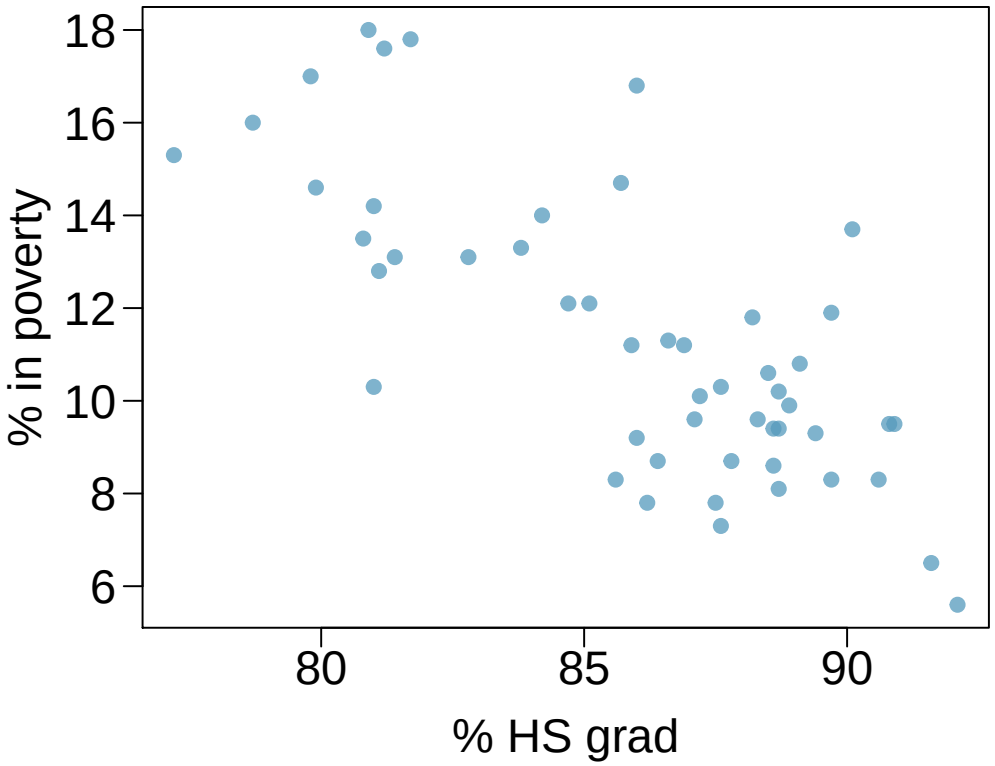
$$y_i = f(x_i) + \epsilon_i$$

$$E[\epsilon_i] = 0$$

$$\hat{y}_i = f(x_i)$$

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

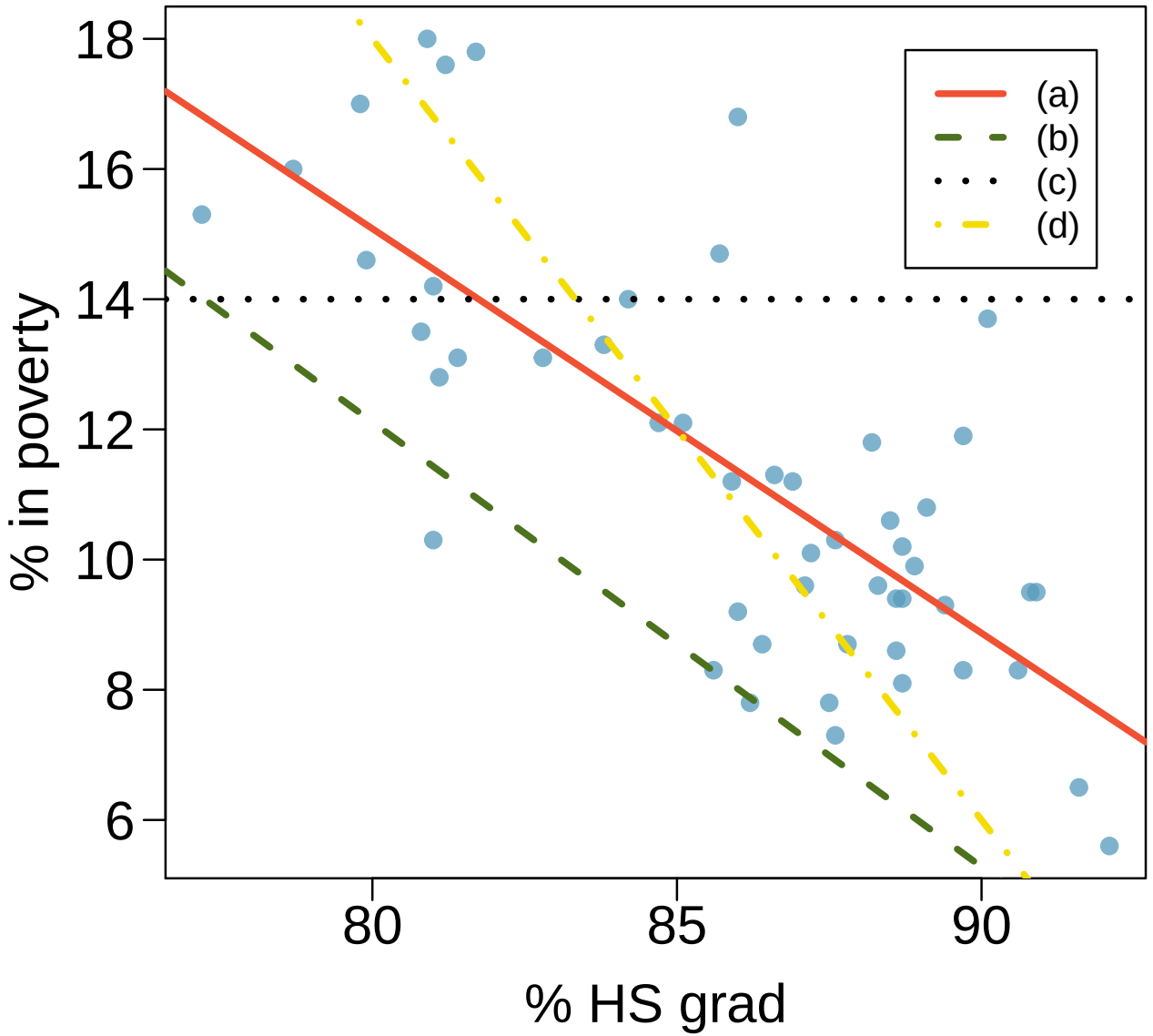
Relationship?

linear, negative, moderately strong

Eyeballing the line

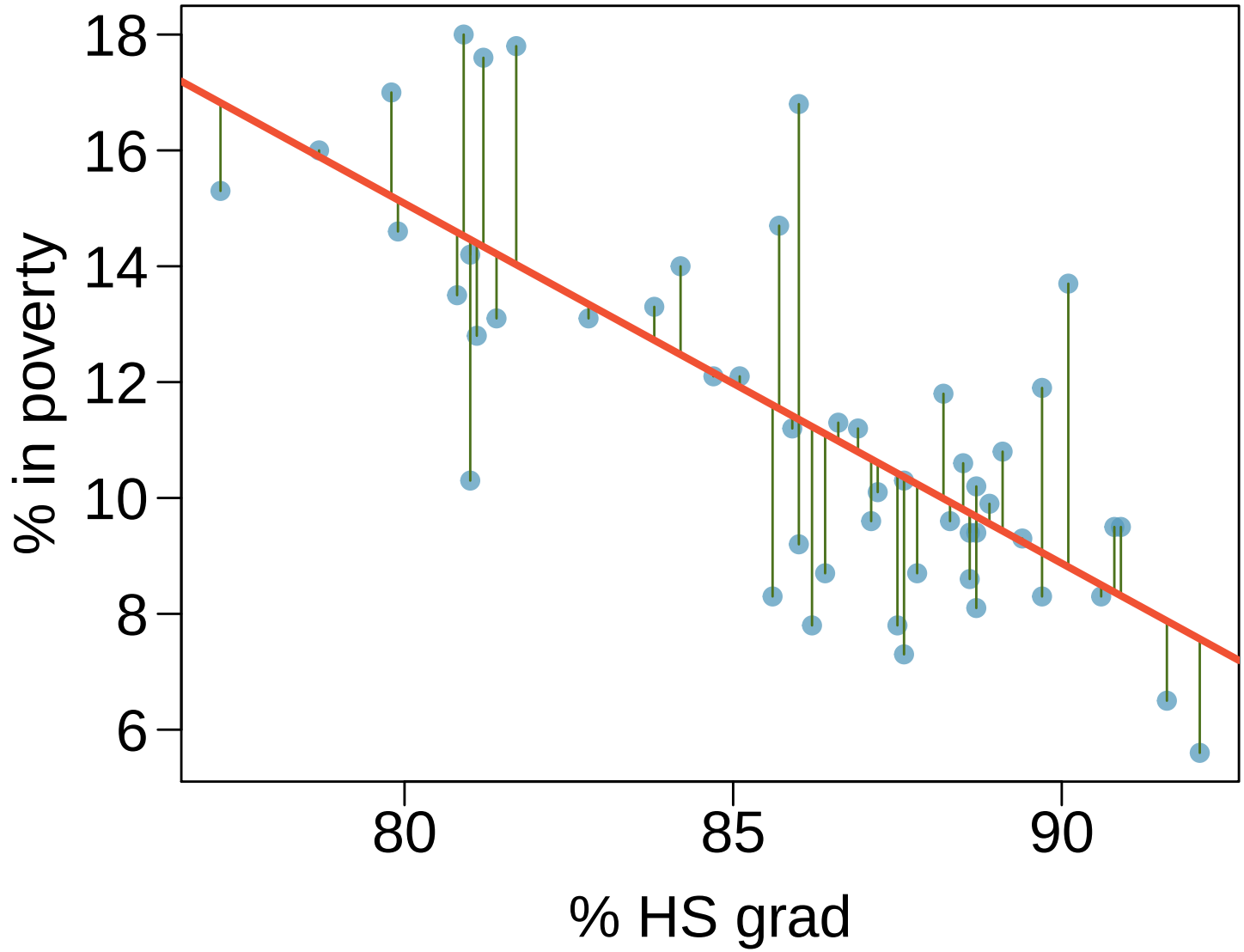
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



Residuals

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

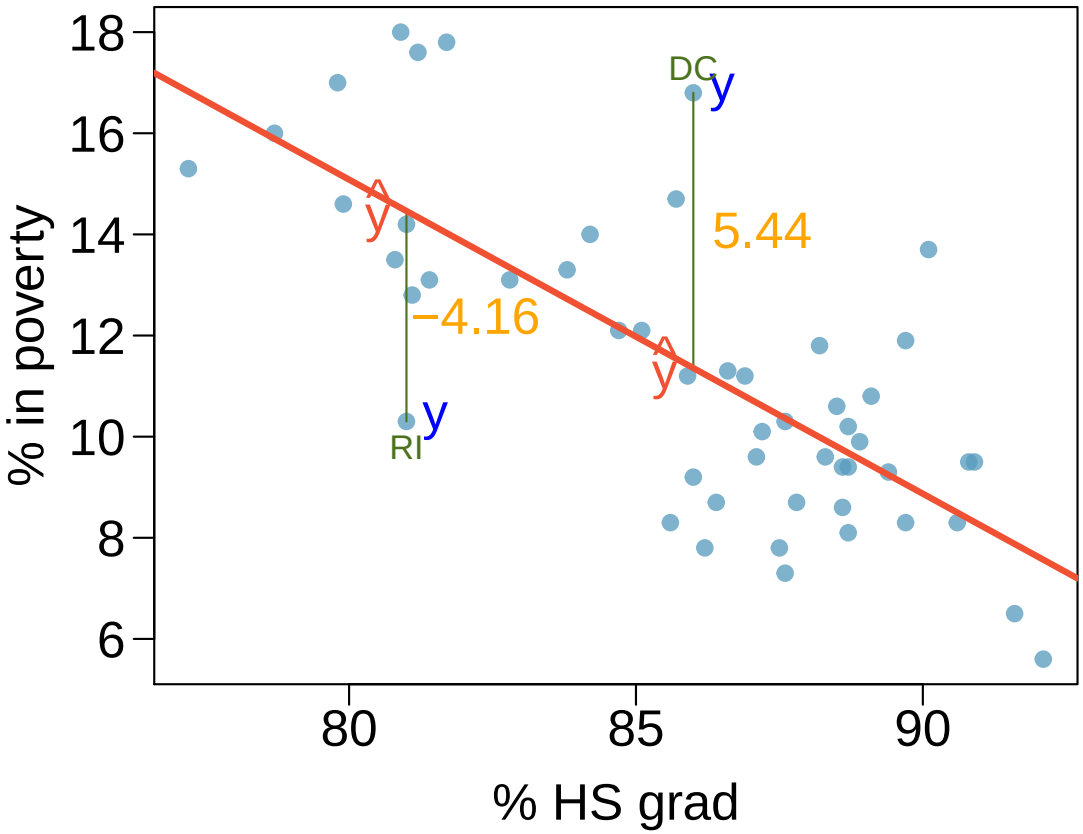


Residuals (cont.)

Residual

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

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



- % 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:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

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

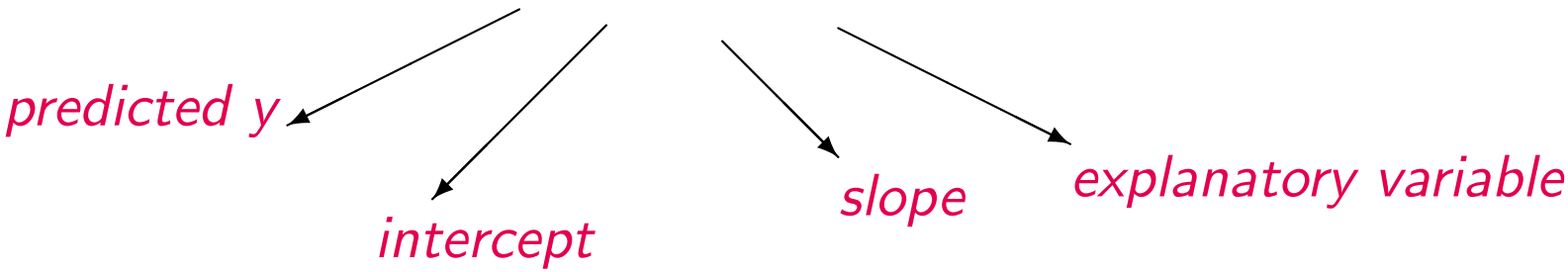
  2. Option 2: Minimize the sum of squared residuals – *least squares*

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

- 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

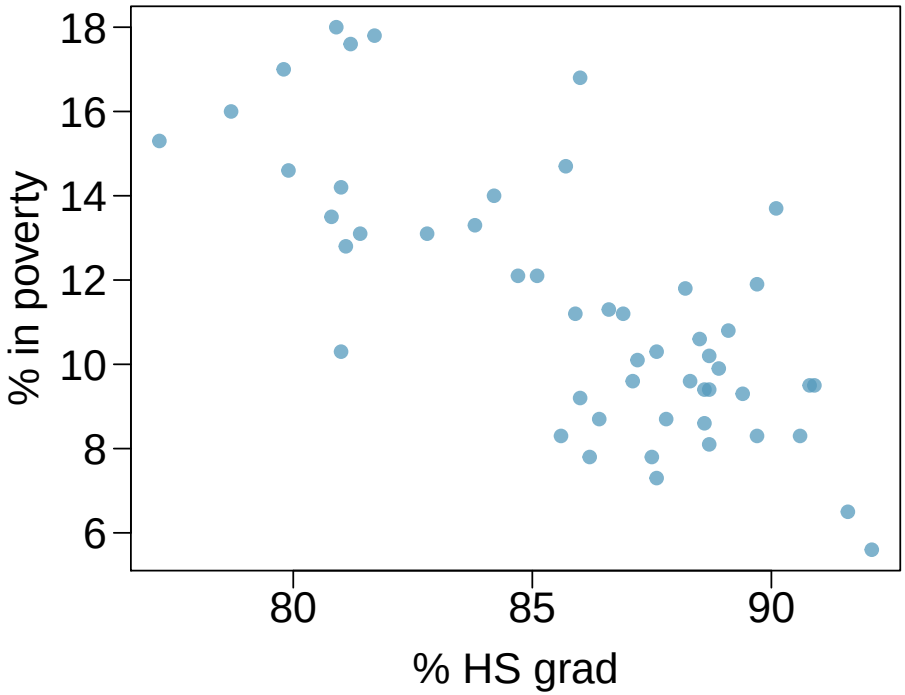
$$\hat{y} = \beta_0 + \beta_1 x$$



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

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

In context...

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

Interpretation

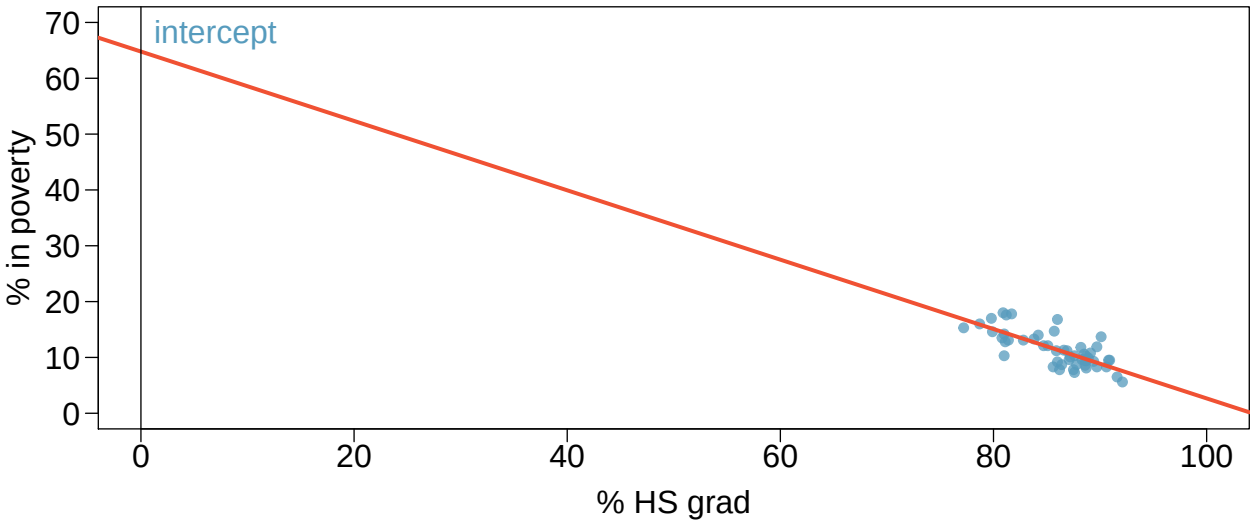
For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

Intercept

Intercept

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

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



$$b_0 = 11.35 - (-0.62) \times 86.0$$
$$= 64.68$$

Intercept

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.

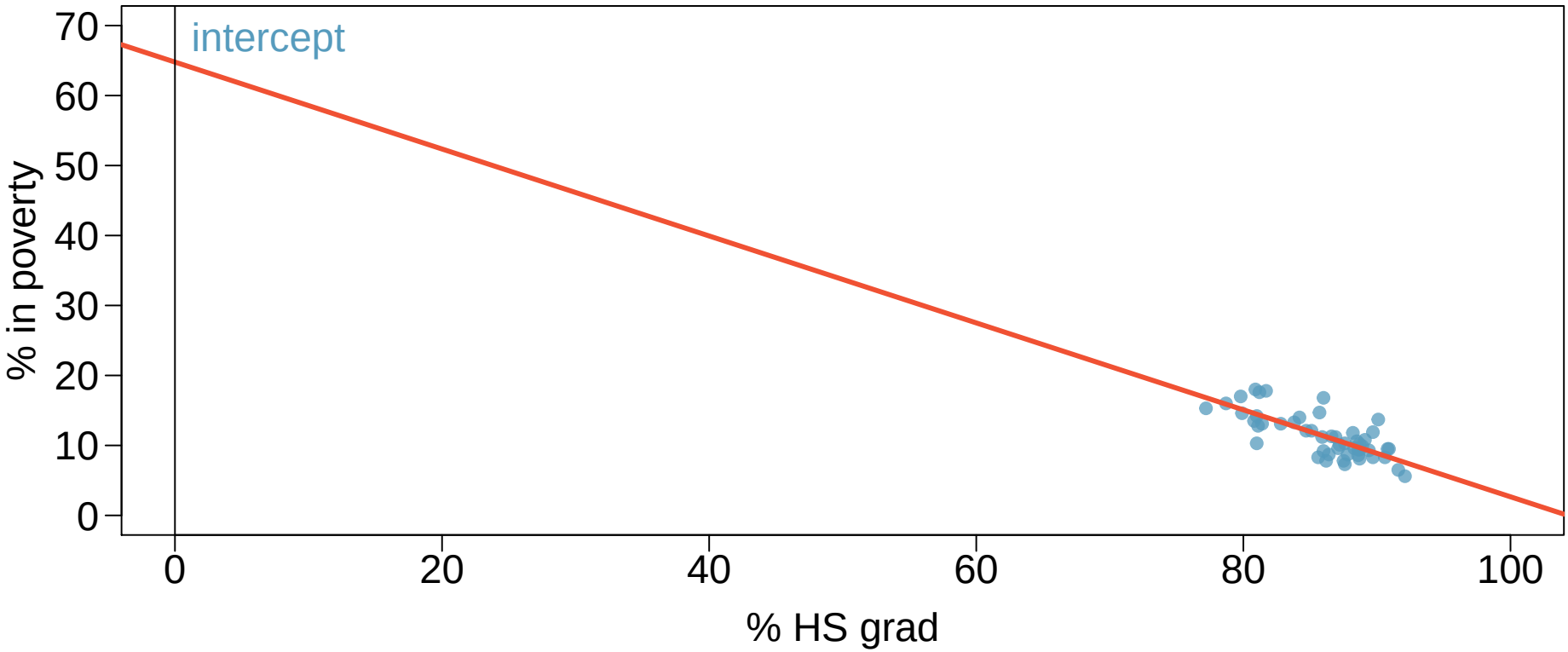
(d) States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

(e) *States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.*

(f) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

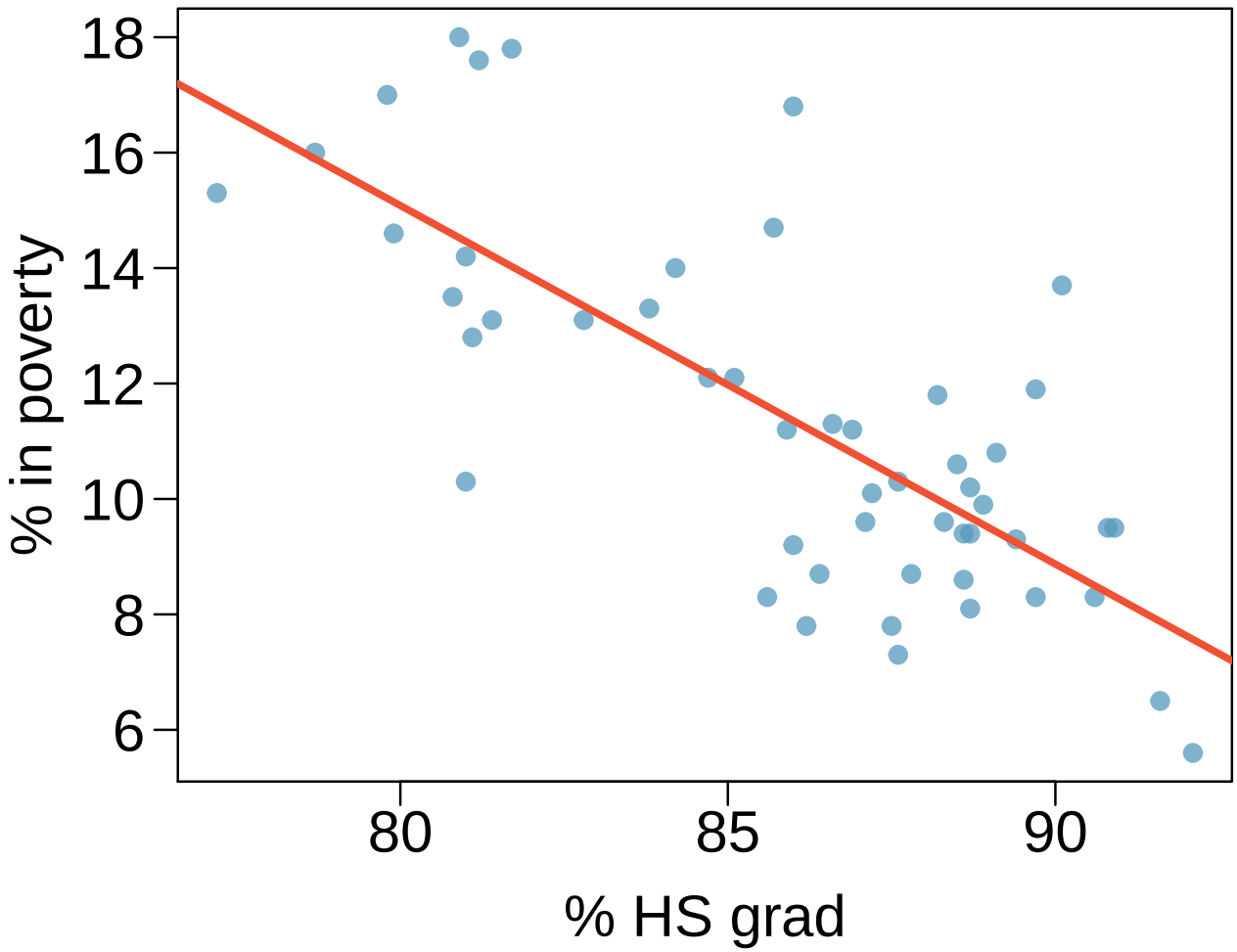
More on the intercept

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.



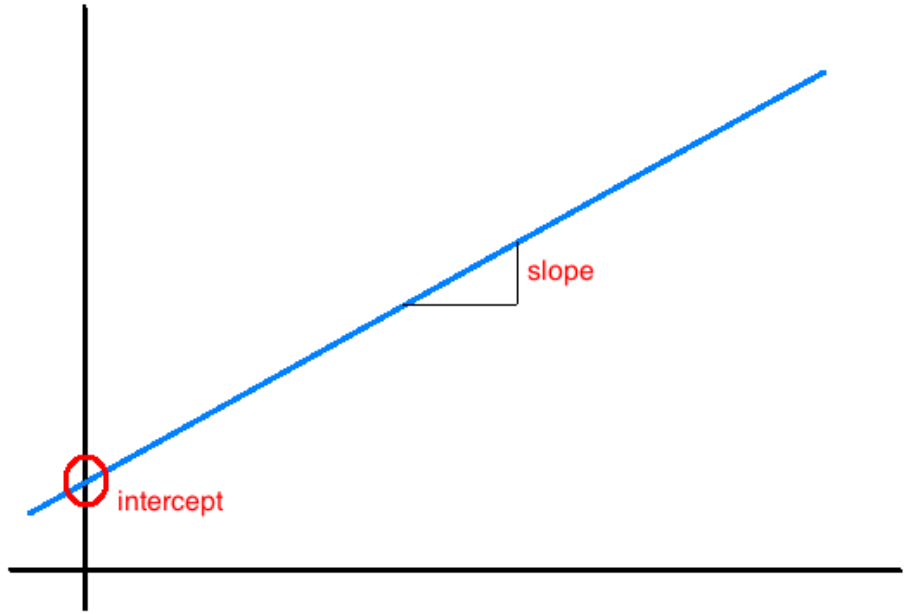
Regression line

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



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.

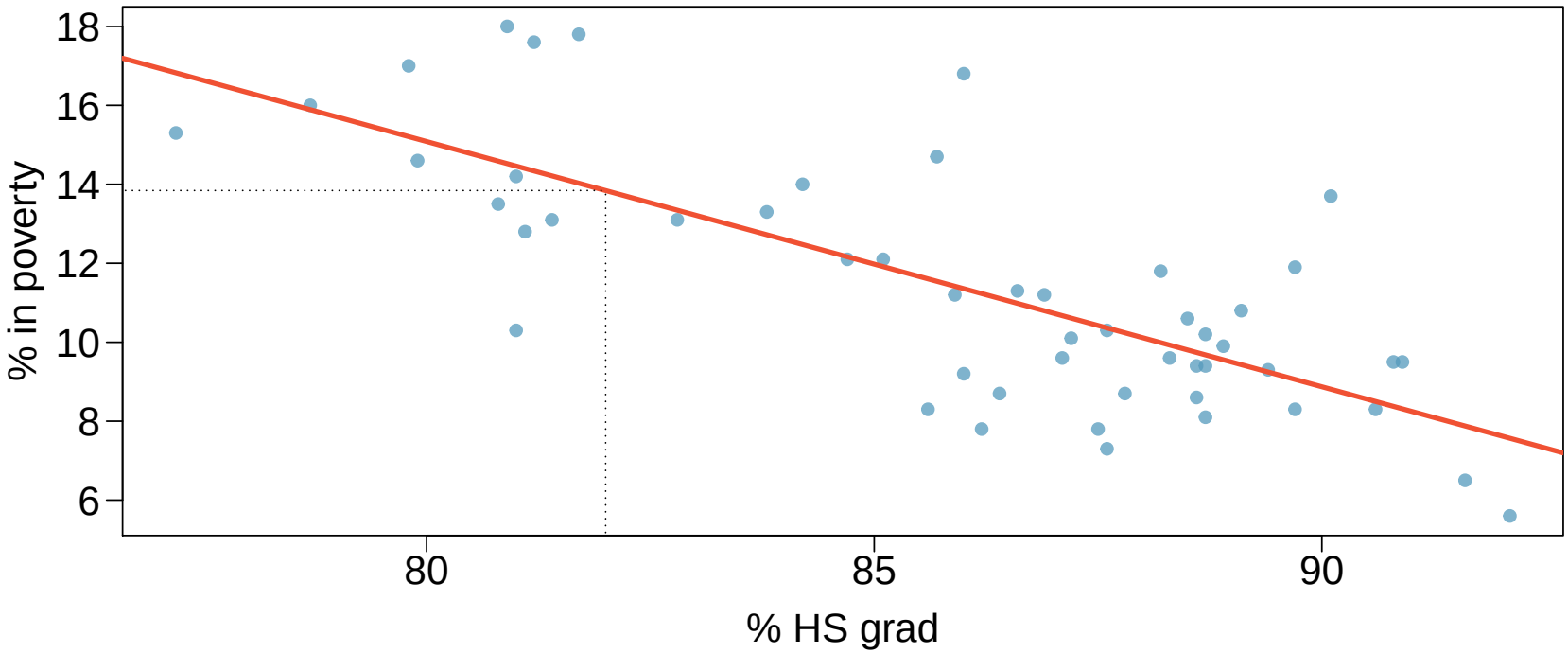


Exercise

LinearRegression.ipynb through Exercise 4.

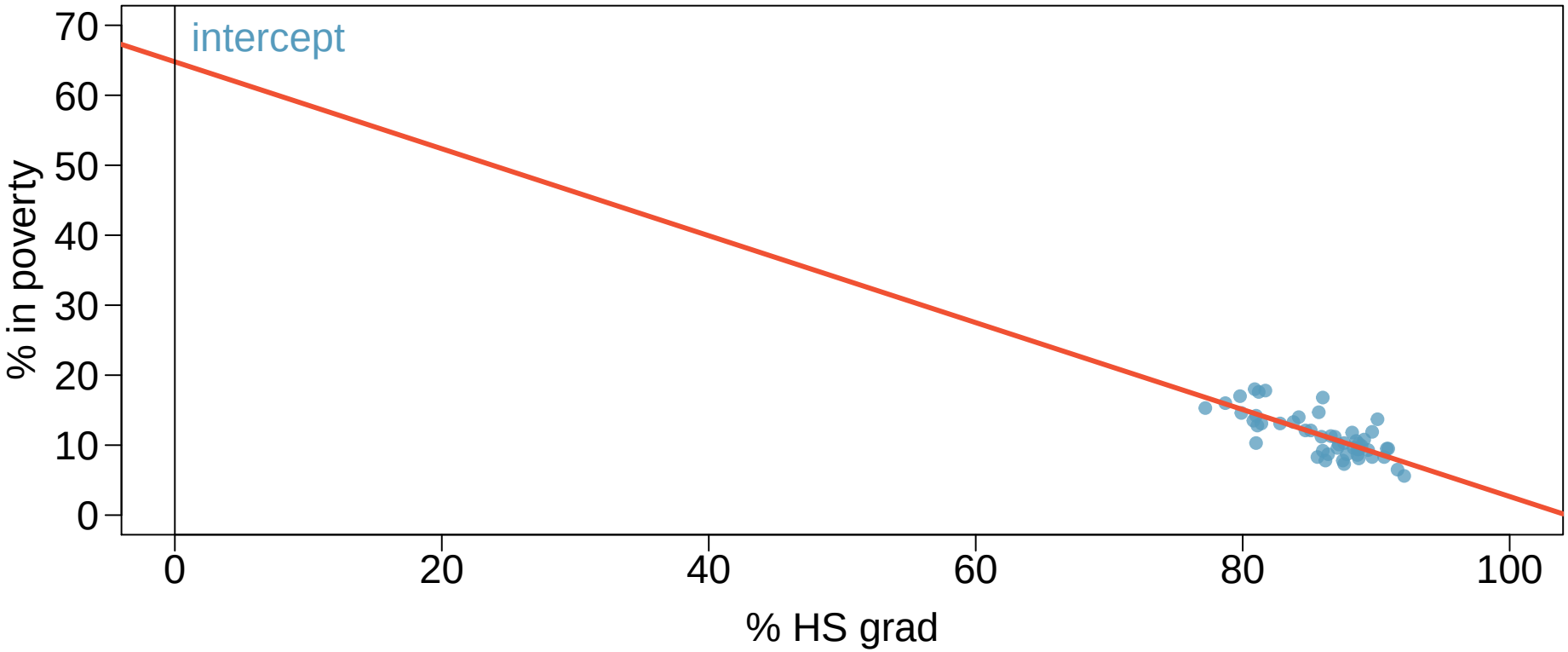
Prediction

- 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.

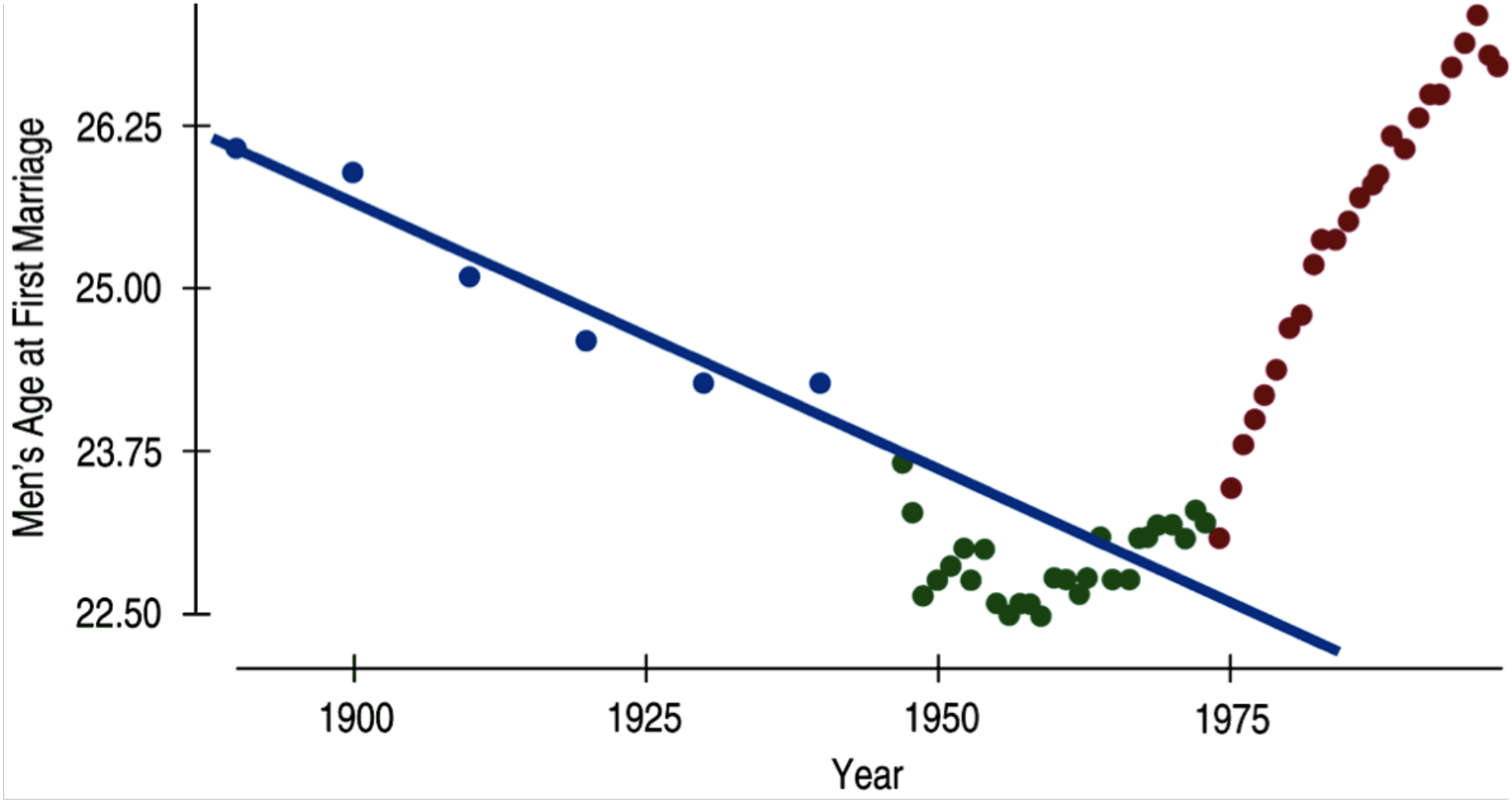


Extrapolation

- 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

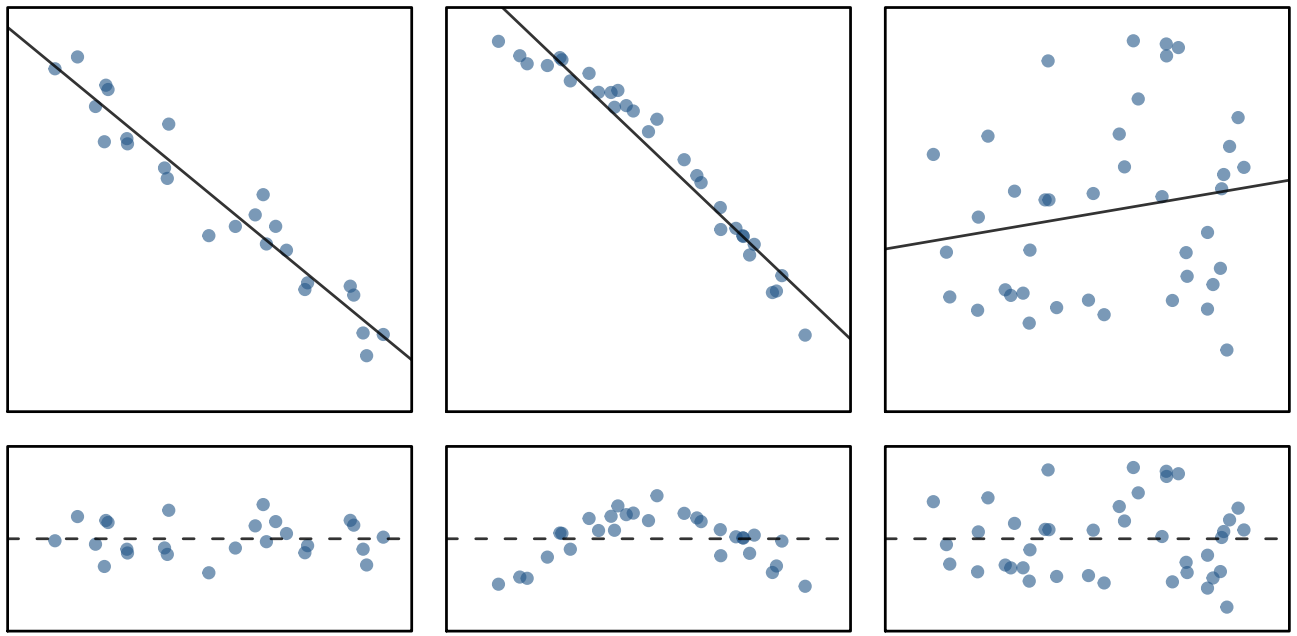


Conditions for the least squares line

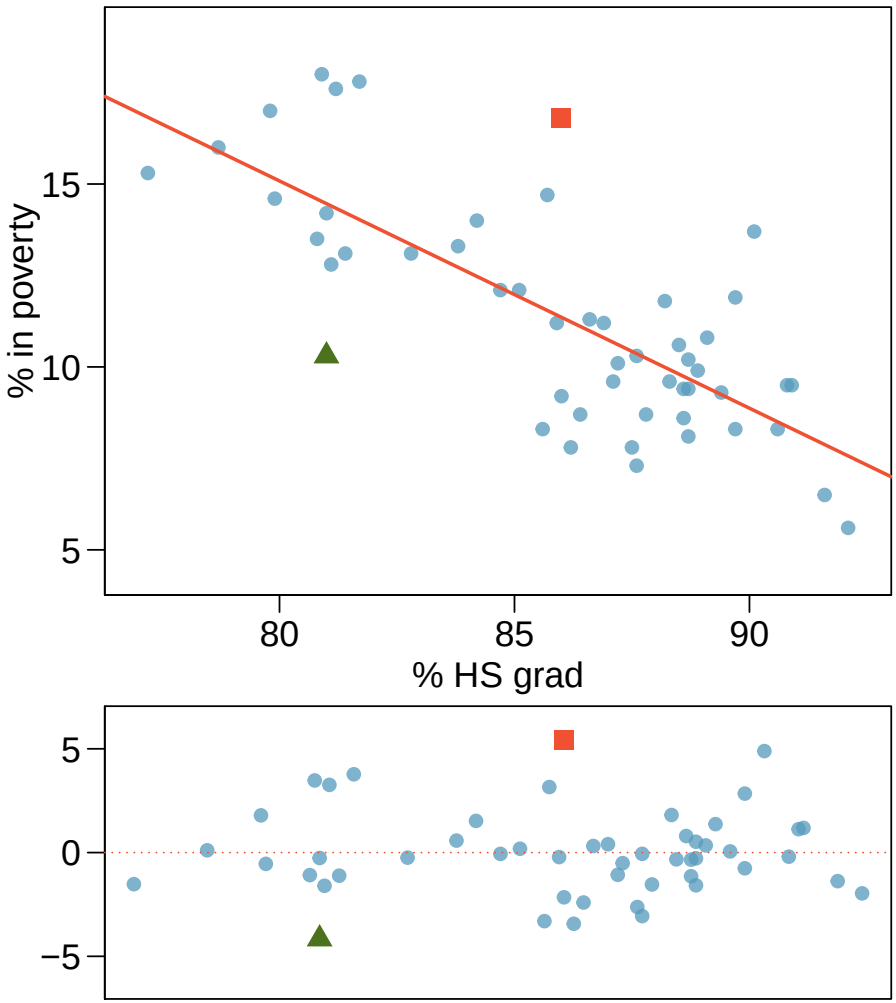
1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

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

$$\%\ HS\ grad = 81 \qquad \%\ in\ poverty = 10$$

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

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

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

$$\%\ HS\ grad = 86 \qquad \%\ in\ poverty = 16$$

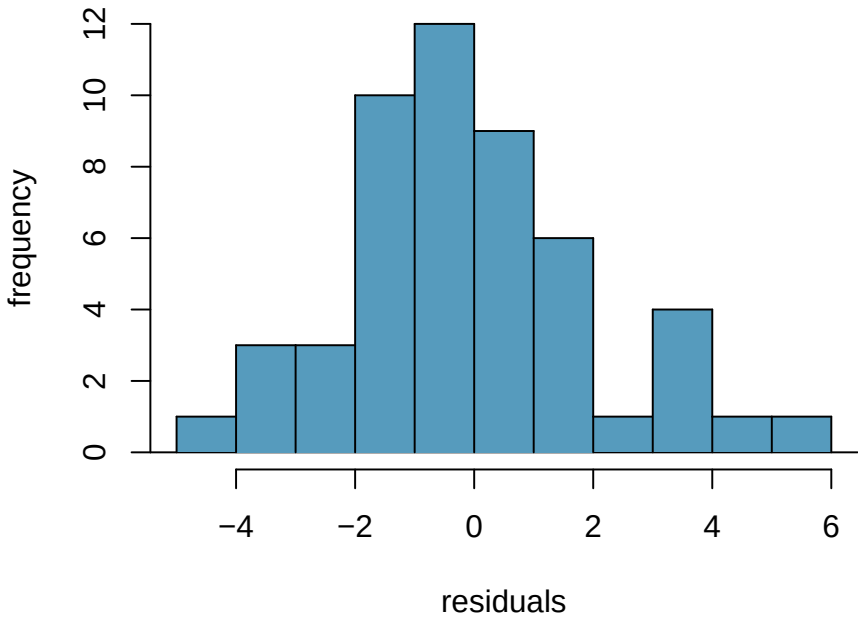
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 86 = 11.36$$

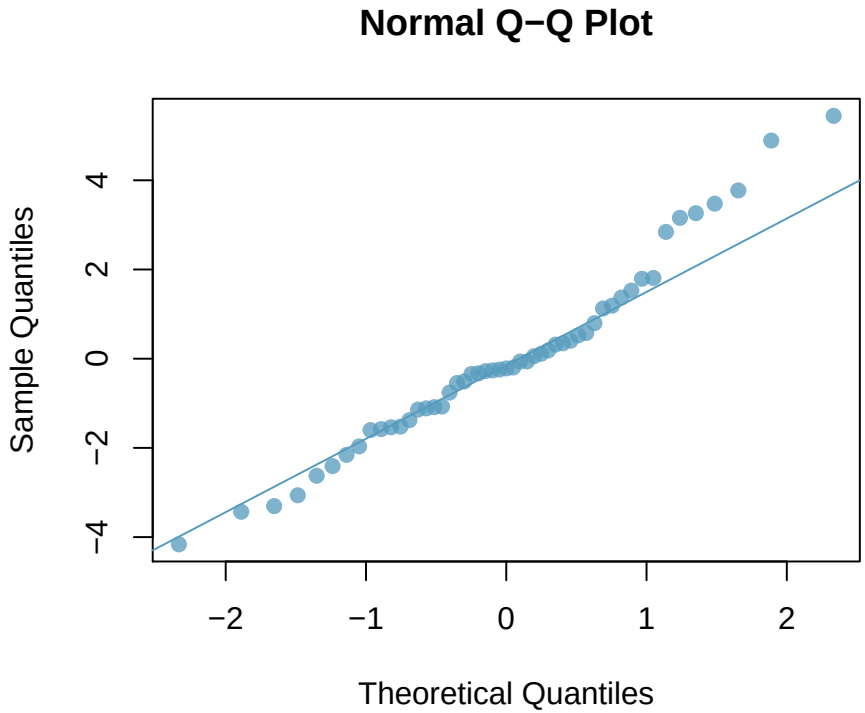
$$e = \%\ in\ poverty - \widehat{\%\ in\ pove}$$

$$= 16.8 - 11.36 = 5.44$$

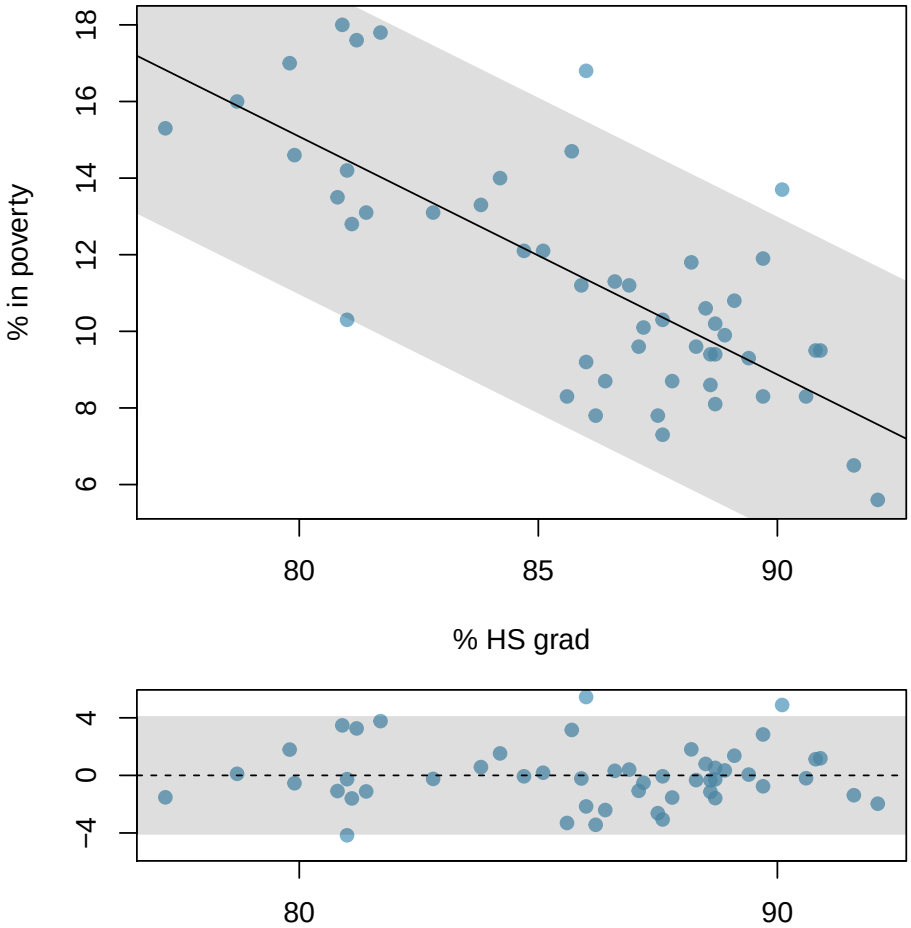
Conditions: (2) Nearly normal residuals

- 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

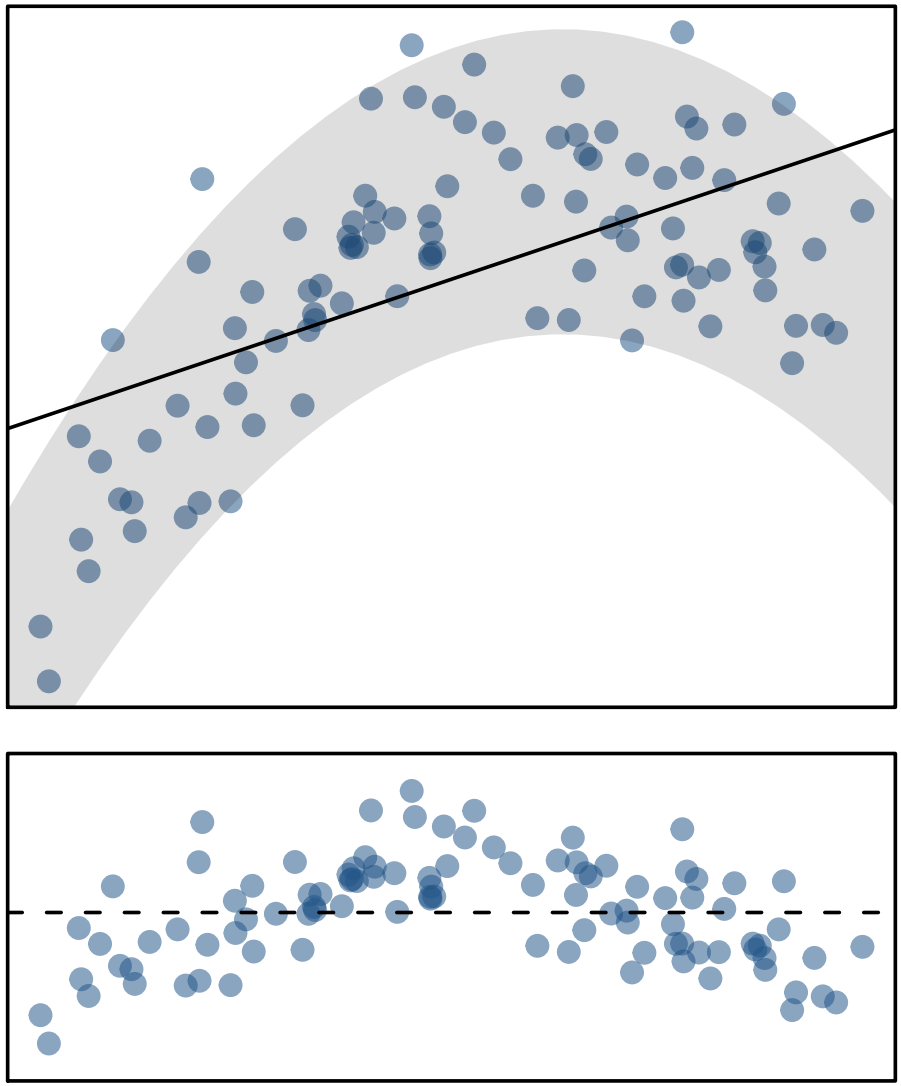


- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a histogram or normal probability plot of residuals.

Checking conditions

What condition is this linear model obviously violating?

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



Checking conditions

What condition is this linear model obviously violating?

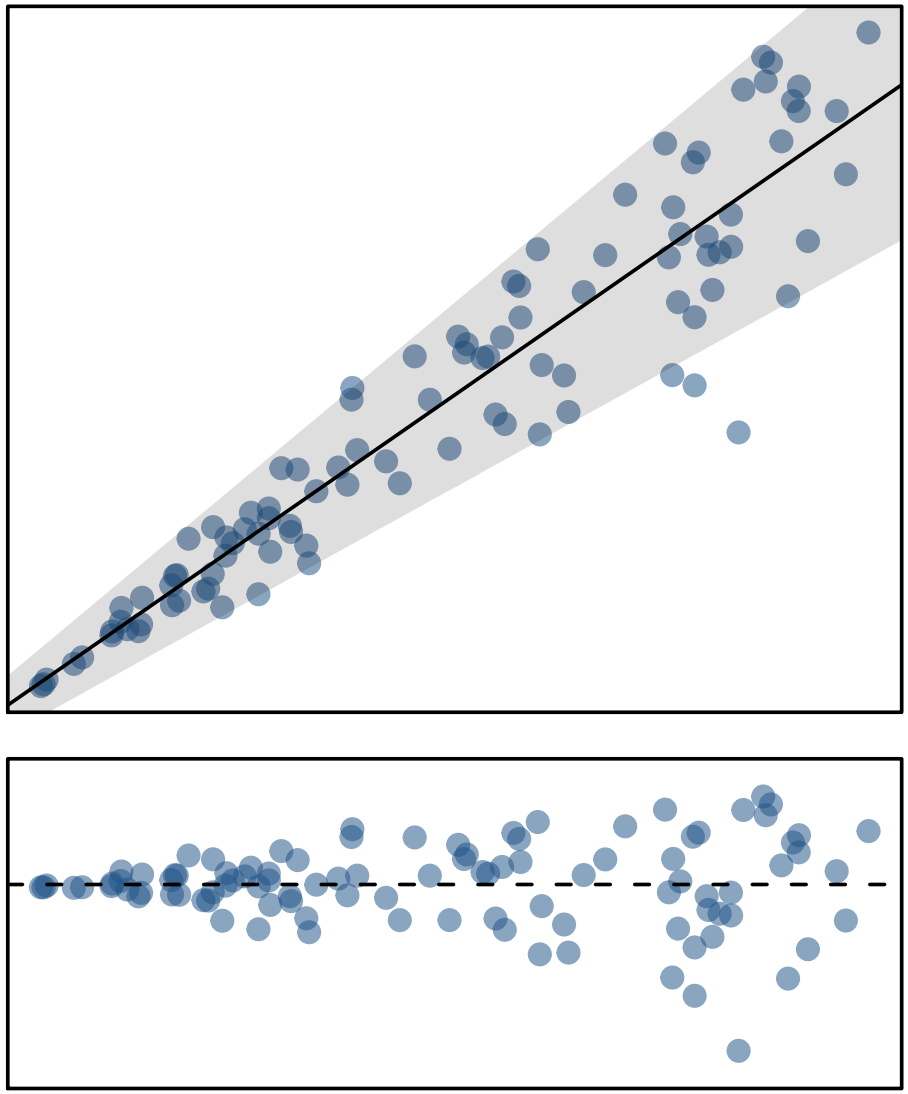
(a) Constant variability

(b) *Constant variability*

(c) Linear relationship

(d) Normal residuals

(e) No extreme outliers



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 $R = -0.62$, $R^2 = 0.38$?

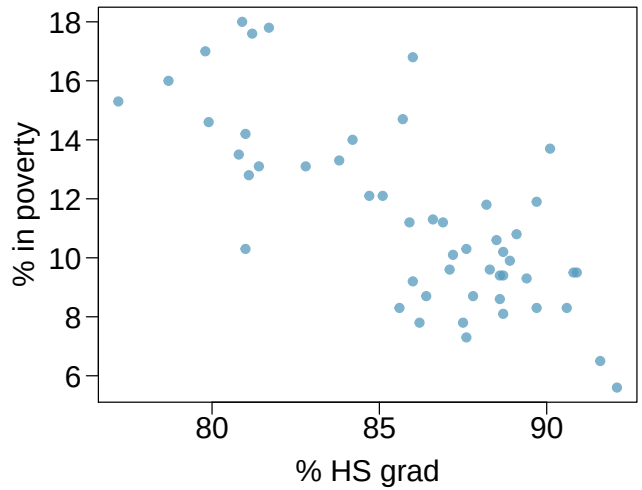
(a) 38% of the variability in the % of HG graduates among the 51 states is explained by the model.

(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 variability in the % of residents living in poverty among the 51 states is explained by the model.*

(d) 38% of the time % HS graduates predict % living in poverty correctly.

(e) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



Exercise

Continue with LinearRegression.ipynb

References


David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 7-8
Data Science from Scratch, Chapters 14-16
Art of Data Science, Chapter 7