Announcements

- HW3 due Monday
- Lab 3 write-up due Monday as well
- Next lab isn't until next Wednesday
- Read Ch 5.2 for Friday

Understanding Check

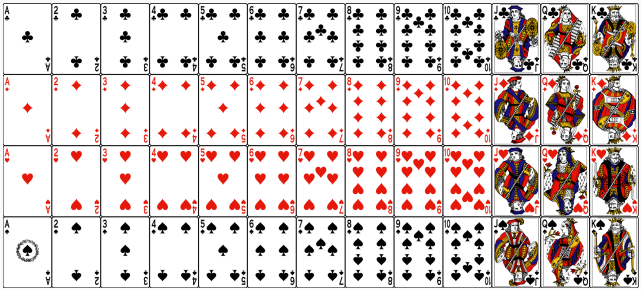
You draw two cards from a standard shuffled deck of cards, replacing the first after it is drawn. What is the probability that on the first you draw a King or club, and on the second you draw a number less than 5 or a red card?

A) 0.201

B) 0.289

C) 0.314

D) 0.385



Understanding Check

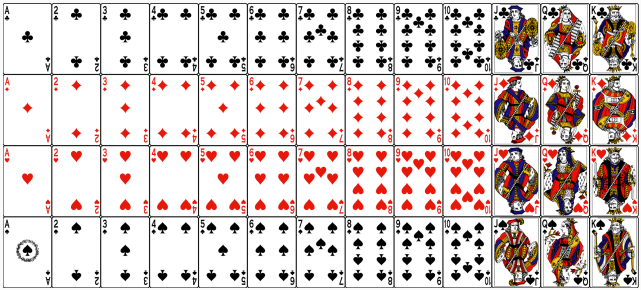
You draw two cards from a standard shuffled deck of cards, replacing the first after it is drawn. What is the probability that on the first you draw a King or club, and on the second you draw a number less than 5 or a red card?

A) 0.201

B) 0.289

C) 0.314

D) 0.385



Justification

Why linear regression and why now?

- We've already been talking about how data is related, linear fits just quantify that idea
- One of the more common methods that people have already been exposed to
- Let's us focus the second half of the semester on significance

Justification

Why linear regression and why now?

- We've already been talking about how data is related, linear fits just quantify that idea
- One of the more common methods that people have already been exposed to
- Let's us focus the second half of the semester on significance

Linear regression lets us quantify the relationship between two numerical variables, and can enable us to model or predict response variables from explanatory variables.

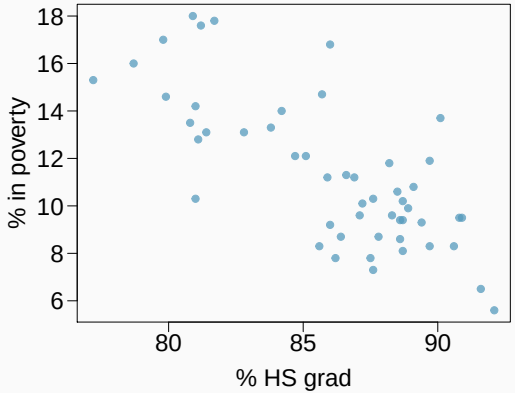
Poverty vs. HS graduation rate

The scatterplot to the right shows the relationship between HS graduation rates in all 50 US states and the % of residents who live below the poverty line.

Response variable?

Explanatory variable?

Relationship?



Poverty vs. HS graduation rate

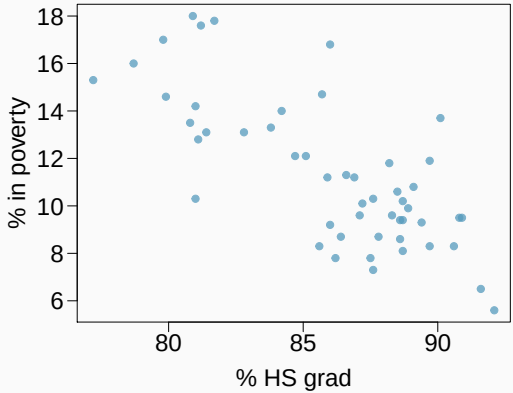
The scatterplot to the right shows the relationship between HS graduation rates in all 50 US states and the % of residents who live below the poverty line.

Response variable?

- % in poverty

Explanatory variable?

Relationship?



Poverty vs. HS graduation rate

The scatterplot to the right shows the relationship between HS graduation rates in all 50 US states and the % of residents who live below the poverty line.

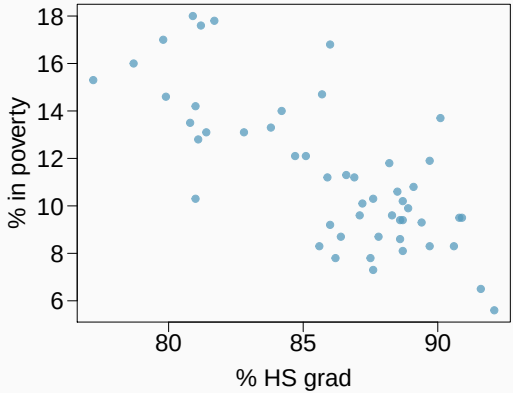
Response variable?

- % in poverty

Explanatory variable?

- % HS graduation rate

Relationship?



Poverty vs. HS graduation rate

The scatterplot to the right shows the relationship between HS graduation rates in all 50 US states and the % of residents who live below the poverty line.

Response variable?

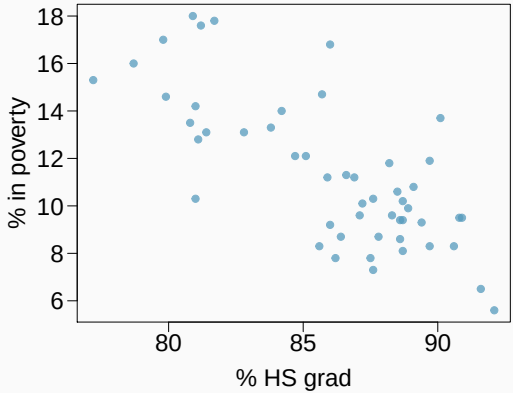
- % in poverty

Explanatory variable?

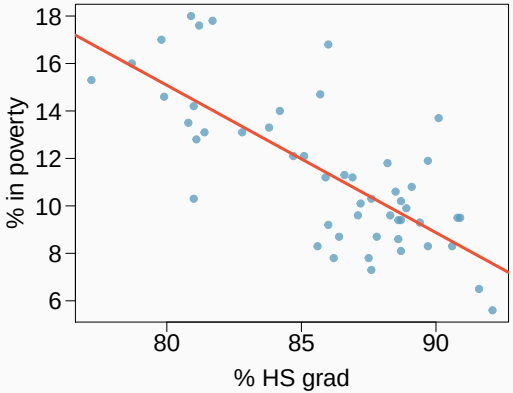
- % HS graduation rate

Relationship?

- linear, negative, moderately strong



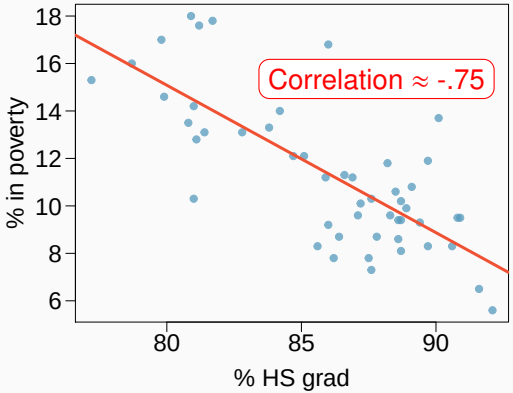
Quantifying your relationship with numbers



- ***Correlation*** describes the strengths 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
- Mathematically looks like:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

Quantifying your relationship with numbers



- ***Correlation*** describes the strengths 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
- Mathematically looks like:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y}$$

Guess the Correlation

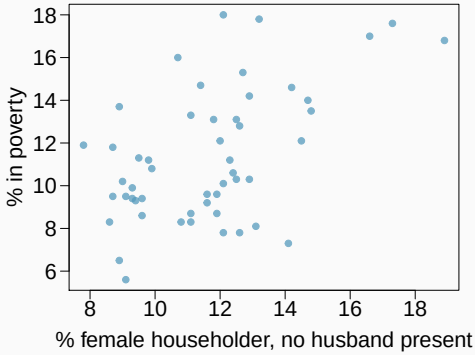
Which of the following is the best guess for the correlation between the % of female households with no husband present and the % in poverty?

A) 0.1

B) -0.4

C) 0.9

D) 0.5



Guess the Correlation

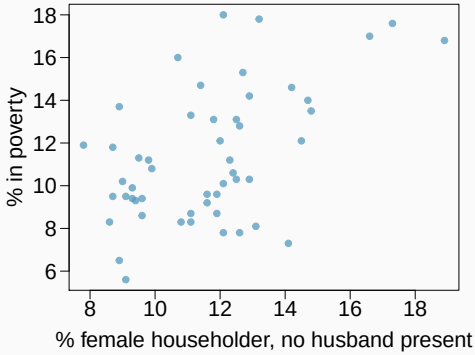
Which of the following is the best guess for the correlation between the % of female households with no husband present and the % in poverty?

A) 0.1

B) -0.4

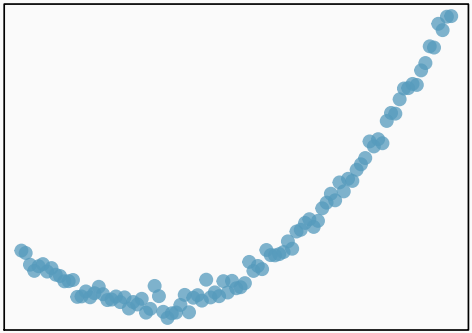
C) 0.9

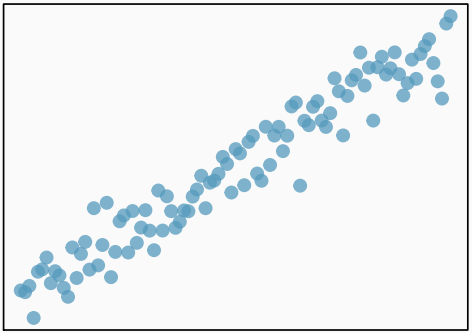
D) 0.5



Beware of Curves!

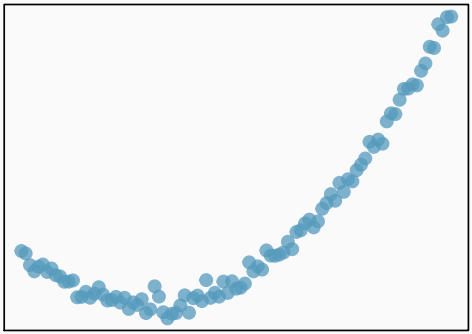
Strength of correlation is based off of linear relationships!

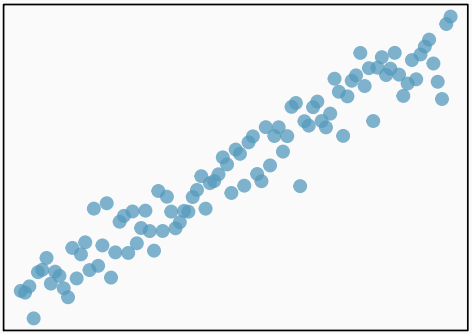




Beware of Curves!

Strength of correlation is based off of linear relationships!





Higher correlation

Having a fit

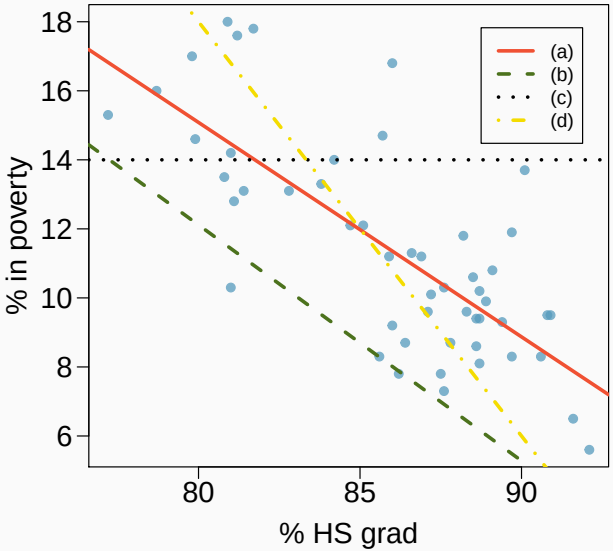
What is a linear fit?

- Attempts to model data with high degrees of correlation
- Never going to be perfect, but useful for making predictions
- Establishes relationships of the form:

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

  where β_0 and β_1 are the model parameters
- x is the explanatory variable, $\hat{y}$ the predicted response variable

Eyeballing it

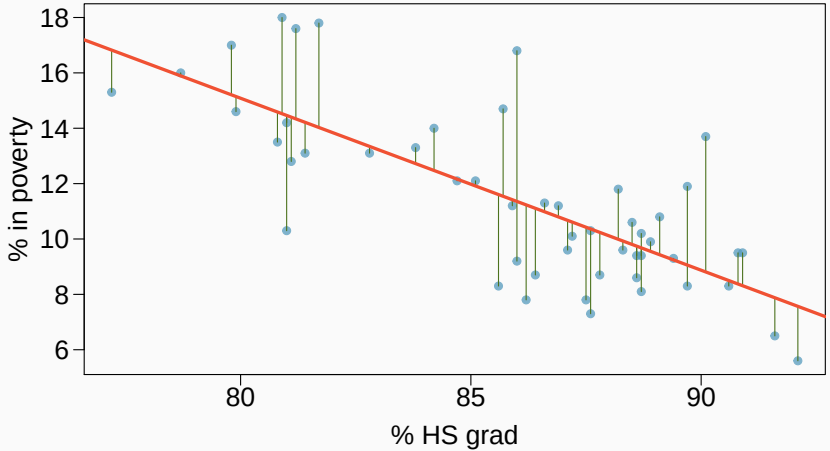


What makes one particular line describing a correlation better than another?

Residuals

Residuals are the leftovers from the model fit:

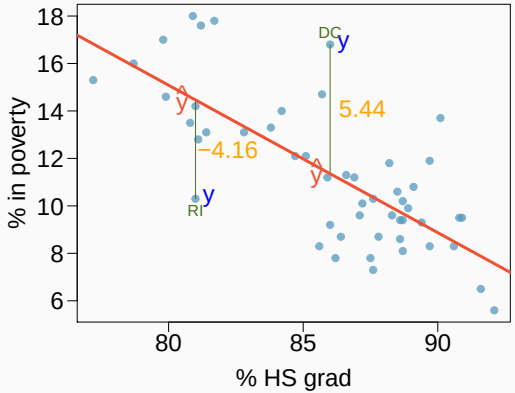
$$\text{Data} = \text{Fit} + \text{Residual}$$



Residual Math

Mathematically, the residual is the difference between the observed response variable (y_i), and the predicted response variable ($\hat{y}_i$).

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

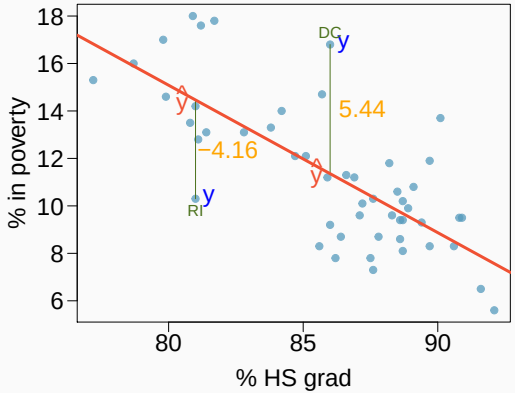


Residual Math

Mathematically, the residual is the difference between the observed response variable (y_i), and the predicted response variable ($\hat{y}_i$).

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

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

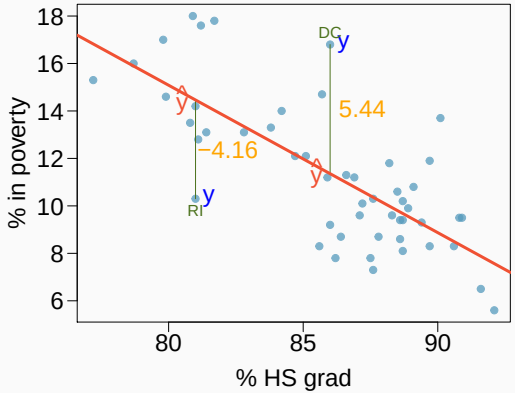


Residual Math

Mathematically, the residual is the difference between the observed response variable (y_i), and the predicted response variable ($\hat{y}_i$).

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

- % living in poverty in DC is 5.44% more than predicted
- % living in poverty in RI is 4.16% less than predicted



Understanding Check

The best fit line on the plot to the right is given by:

$$\hat{y} = 64.781 - 0.6212x$$

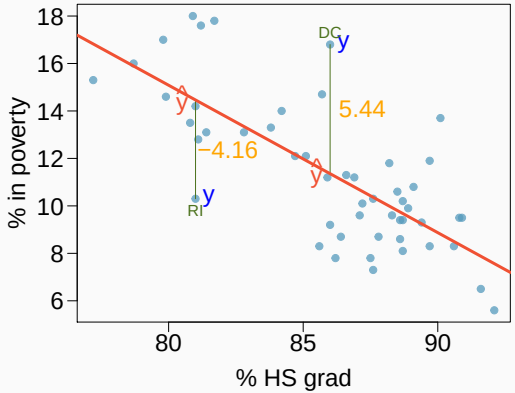
Oregon has an 86.9% HS graduate rate and a poverty rate of 11.2%. What is the residual for Oregon?

A) -0.5%

B) -1.4%

C) 0.4%

D) 29%



Understanding Check

The best fit line on the plot to the right is given by:

$$\hat{y} = 64.781 - 0.6212x$$

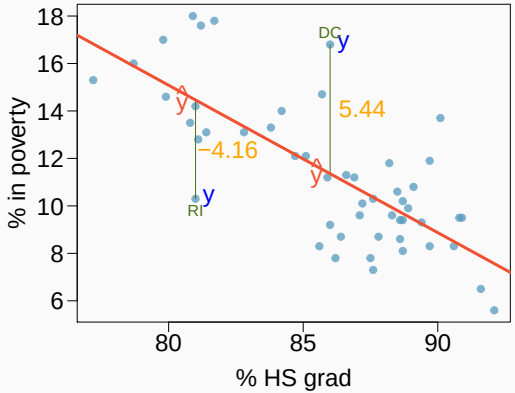
Oregon has an 86.9% HS graduate rate and a poverty rate of 11.2%. What is the residual for Oregon?

A) -0.5%

B) -1.4%

C) 0.4%

D) 29%



Fitting the best lines

- For a useful predictor, we want a line that has small residuals:

  Option 1: Minimize the sum of the absolute values of the residuals

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

  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 computer
  3. Frequently, a residual that is twice as large is more than twice as bad

Anatomy of a Line

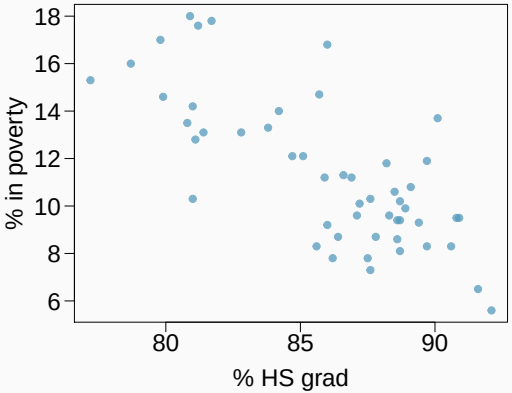
Recall that our best fit line takes the form of

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

- x is our explanatory variable
- $\hat{y}$ is our predicted response variable
- β_0 is the y-intercept of the line
  - β_0 itself corresponds to the population intercept
  - Our estimate of β_0 (from our sample) will be denoted b_0
- β_1 is the slope of the line
  - β_1 corresponds to the population slope
  - Our estimate of β_1 will be denoted b_1

Our Starting Point

Least square linear fits have certain properties that we can take advantage of. 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	$R = -0.75$

The Slope!

The slope of the regression can be calculated as

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

The Slope!

The slope of the regression can be calculated as

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

In this case:

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

The Slope!

The slope of the regression can be calculated as

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

In this case:

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

Interpretation:

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