Unit 4: Simple Regression

1. Intro to Linear Regression

(Chapter 5.1)

3/29/2021

Recap from Unit 3

In Unit 3, we learned to use distributions to answer questions like

- Is X different from what we would expect? (one-sample)
- Is X different from Y? (two-sample)
- Was there a change in X? (paired)

We did this for outcome variables that were categorical (e.g. atheist or not), or numerical (e.g. area). But our independent variables were always categorical.

In Unit 4, we'll talk about using distributions to understand the relationship between two numerical variables.

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	
Categorical Predictor (e.g. Men vs. women)		

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	
Categorical Predictor (e.g. Men vs. women)		

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	Numeric Outcome (e.g. Income)
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	Numeric Outcome (e.g. Income)
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	*Do men tend to earn more money than women?* *Hypothesis test for two means (3.4)*

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	Numeric Outcome (e.g. Income)
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	*Do men tend to earn more money than women?* *Hypothesis test for two means (3.4)*
Numeric Predictor (e.g. Age)		

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	Numeric Outcome (e.g. Income)
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	*Do men tend to earn more money than women?* *Hypothesis test for two means (3.4)*
Numeric Predictor (e.g. Age)	Are older people more likely to be promoted? *Logistic regression (5.3)*	

Recap from Unit 3

	Categorical Outcome (e.g. Promoted or not)	Numeric Outcome (e.g. Income)
Categorical Predictor (e.g. Men vs. women)	*Are men more likely to be promoted than women?* *Hypothesis test for two-proportions (3.2)*	*Do men tend to earn more money than women?* *Hypothesis test for two means (3.4)*
Numeric Predictor (e.g. Age)	Are older people more likely to be promoted? *Logistic regression (5.3)*	Do older people earn more money? *Linear regression (today)*

Key ideas

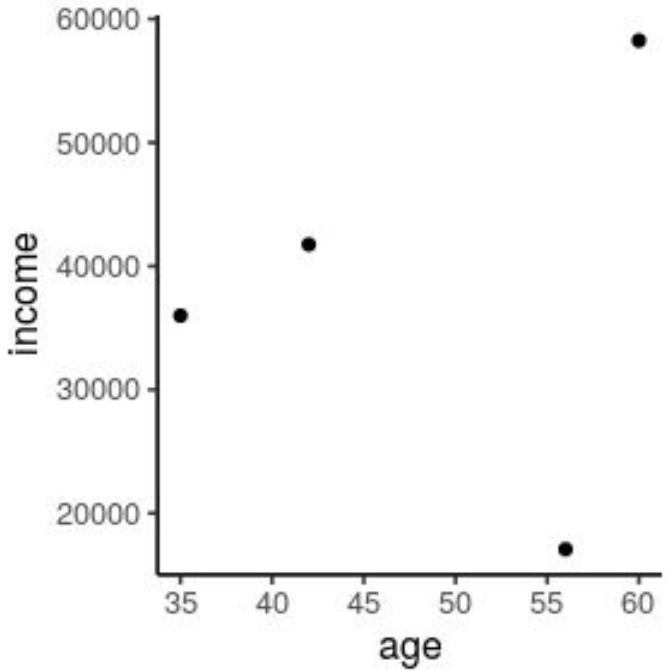
1. Correlation is a measure of the linear relationship between two factors.
2. We can use linear regression to estimate this correlations.
3. A regression line is the line that minimizes the residuals between each point and the line.

Scatterplots

A **scatterplot** shows the relationship between two numeric variables.

Each dot represents a single observation (e.g. a single person).

	age	income
Person 1	35	35,990
Person 2	42	41,750
Person 3	56	17,080
Person 4	60	58,255

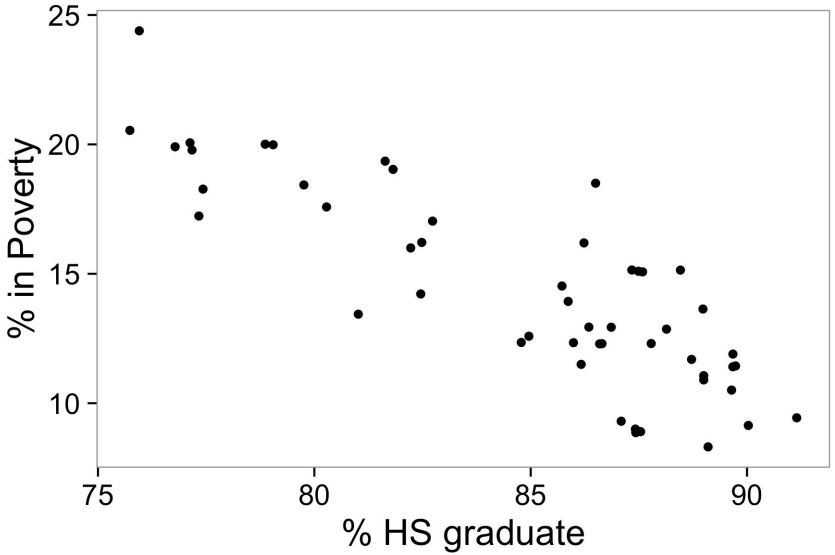


Poverty and high school graduation rate

This **scatterplot** shows the relationship between high school graduation rate and the percent of residents who live below the poverty line in all 50 US states + DC in 2012.

(income below $23,050 for a family of 4).

How would you describe this relationship?

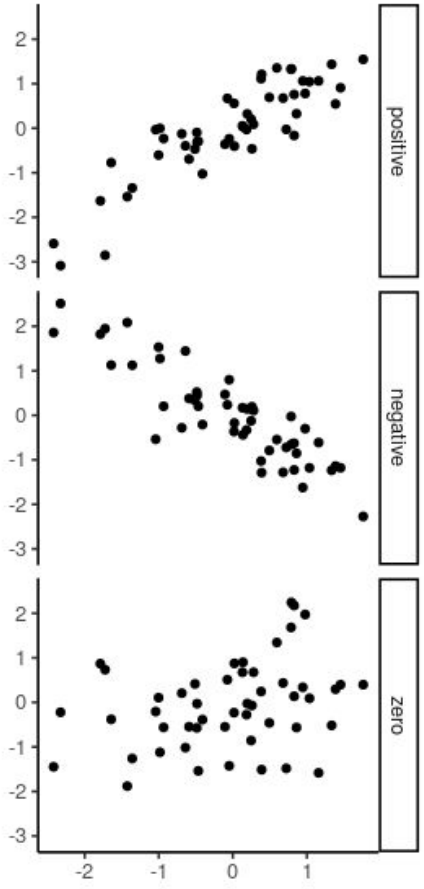


Quantifying the relationship between two numerical values

Correlation describes the strength of the **linear** association between two variables.

Correlation ranges from -1 (perfect negative) to +1 (perfect positive).

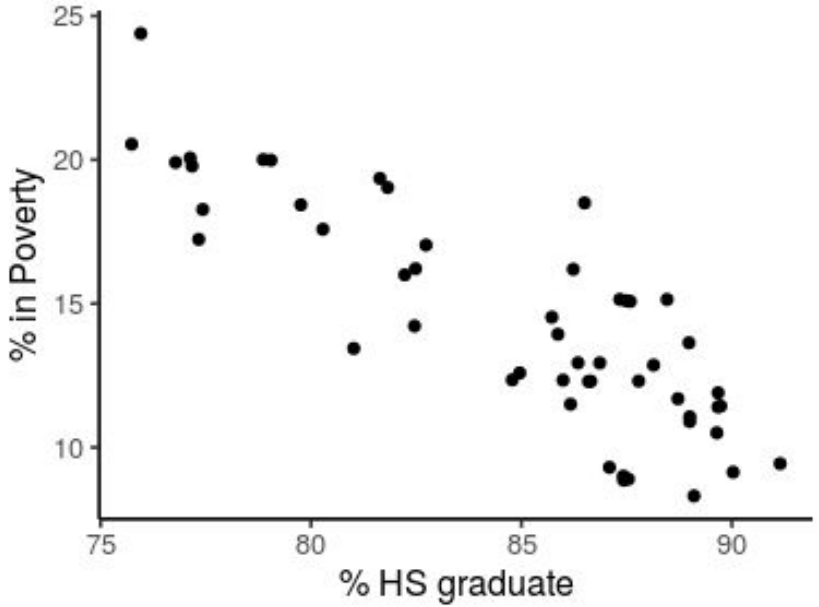
A value of 0 indicates no linear association.



Guess the correlation

Which of these is your best guess for the correlation between poverty and high school graduation?

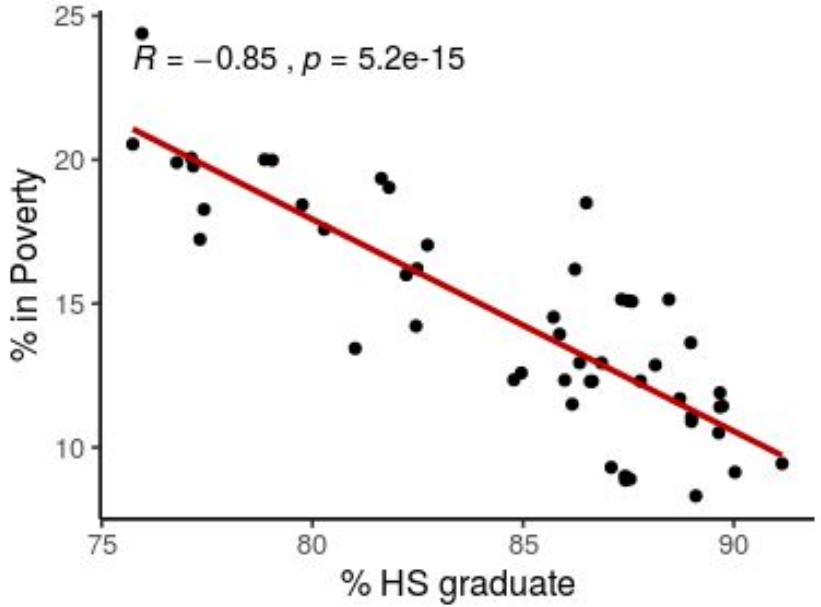
(a) .6
(b) -.85
(c) -.1
(d) .02
(e) -1.5



Guess the correlation

Which of these is your best guess for the correlation between poverty and high school graduation?

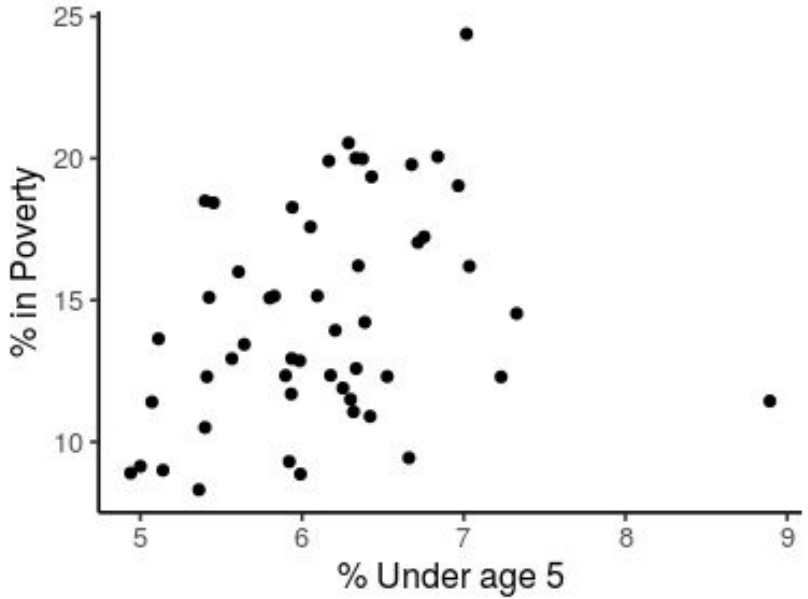
(a) .6
(b) -.85
(c) -.1
(d) .02
(e) -1.5



Guess the correlation

Which of these is your best guess for the correlation between poverty and and the proportion of the population under 5 years of age?

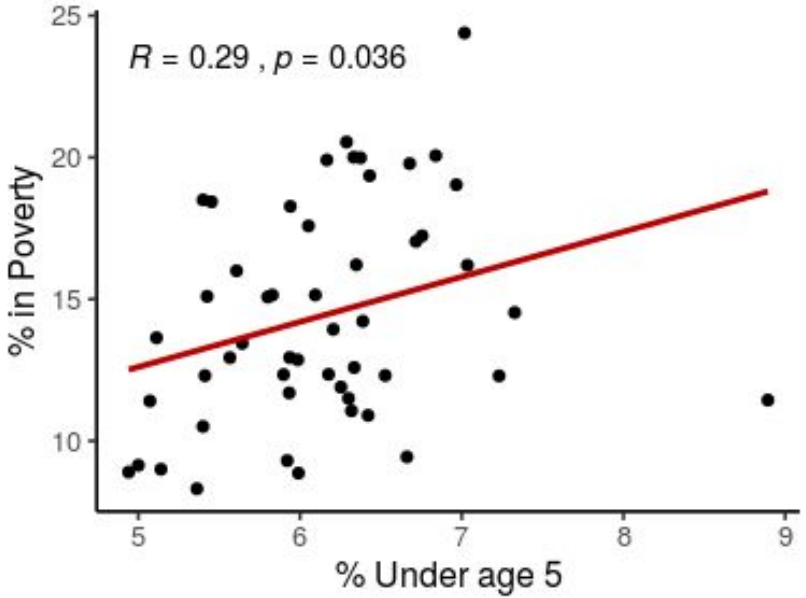
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.3



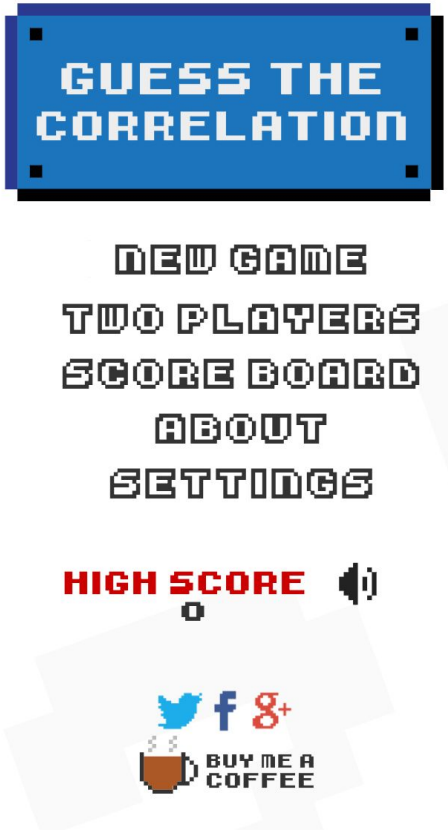
Guess the correlation

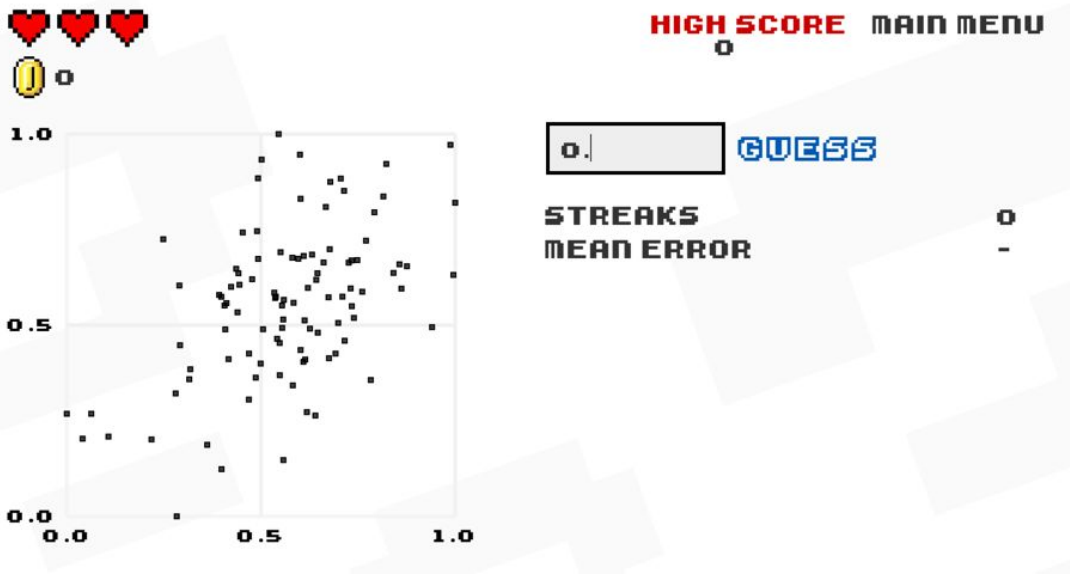
Which of these is your best guess for the correlation between poverty and and the proportion of the population under 5 years of age?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.3

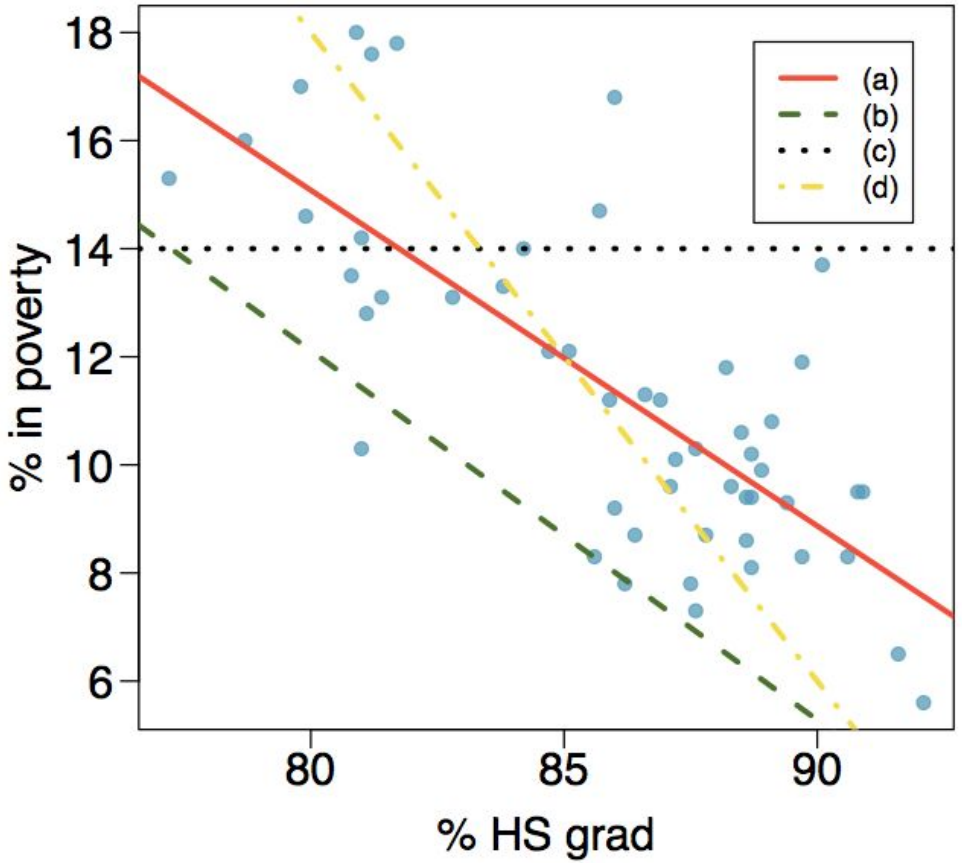


Play along at home: http://guessthecorrelation.com/





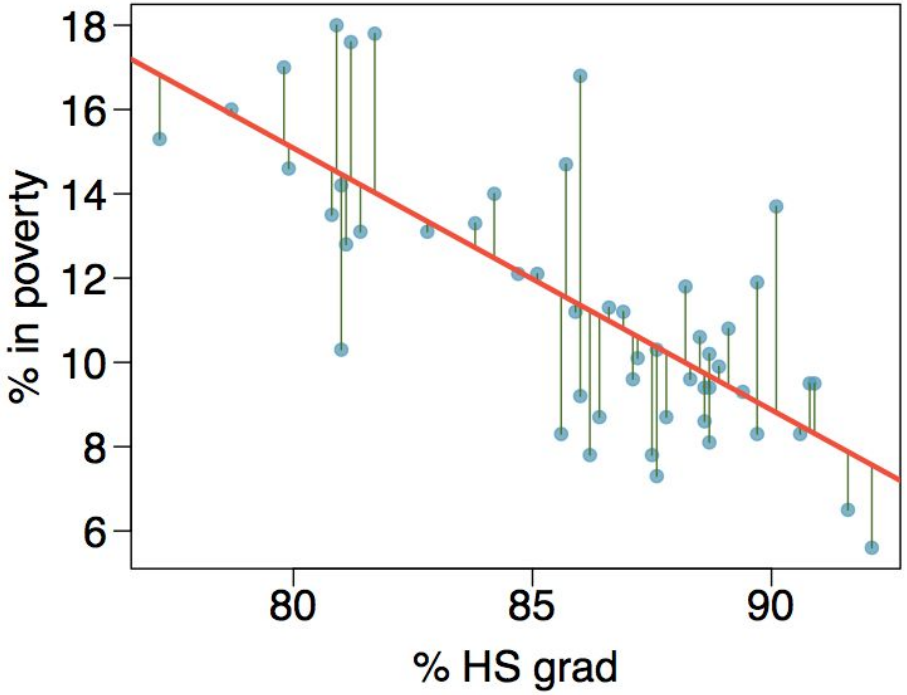
Which of these lines best represents the trend?



How do figure out that we want line (a)?

We want to find the line that minimizes the **residuals**: the distances between each point and the line.

A **regression** model is a model that says that your data is composed of two things:
(1) A best-fit line, and
(2) the residuals between each point and the line.



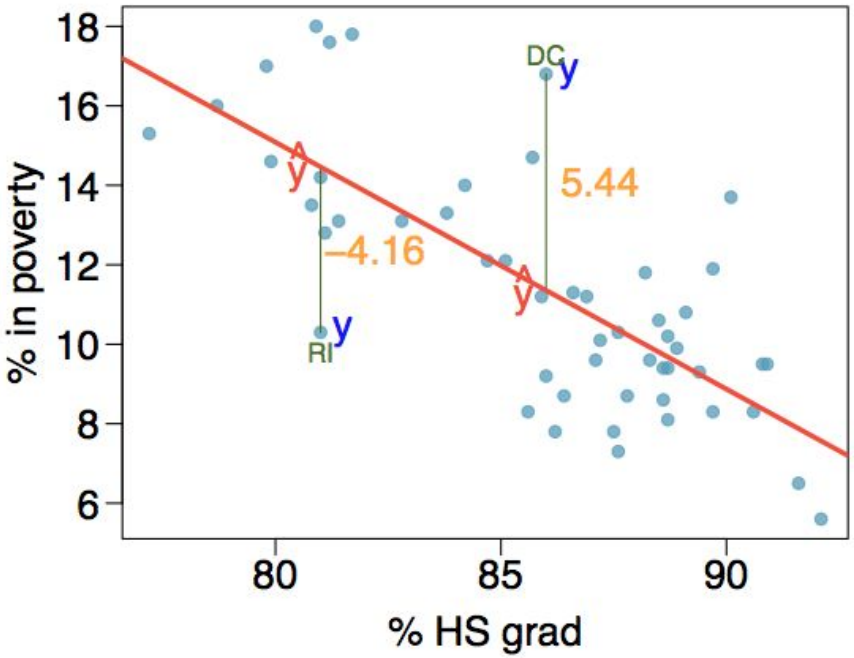
Residuals

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

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

For example, percent living in poverty in **DC** is 5.44% more than predicted based on HS grad % alone.

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



Finding the best line

We want to find the line that has the smallest residuals

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

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

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

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

Why least squares?

- Easier to compute by hand and using software
- Often, a residual twice as large as another is more than twice as bad

Key ideas

1. Correlation is a measure of the linear relationship between two factors.
2. We can use linear regression to estimate this correlations.
3. A regression line is the line that minimizes the residuals between each point and the line.