Unit 4: Regression and Prediction

1. Intro to Linear Regression

(Chapter 5.1)

11/14/2016

Quiz 7 - t-tests, paired data, and differences

Recap from Unit 3

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

- Is X different from what we would expect?
- Is X different from Y?
- Was there a change in X?

We did this for outcome variables that were categorical (e.g. hip roof 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.

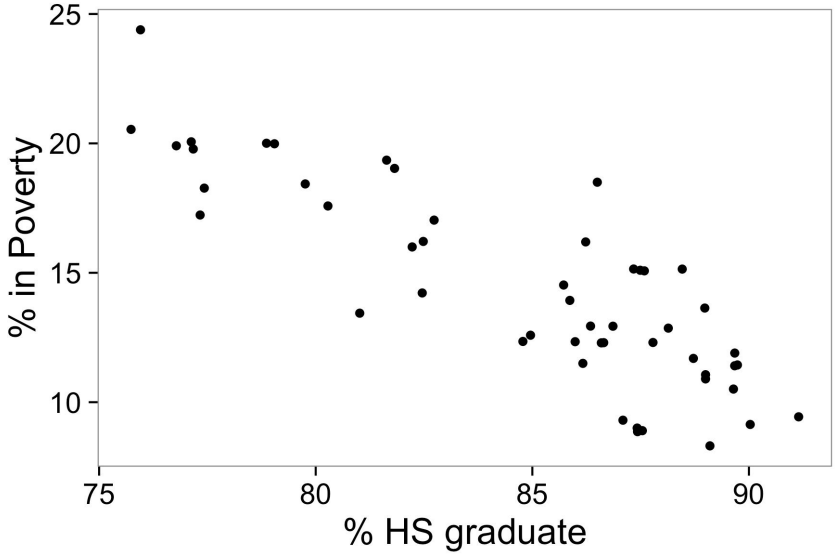
Key ideas

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

Poverty and highschool graduation rate

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

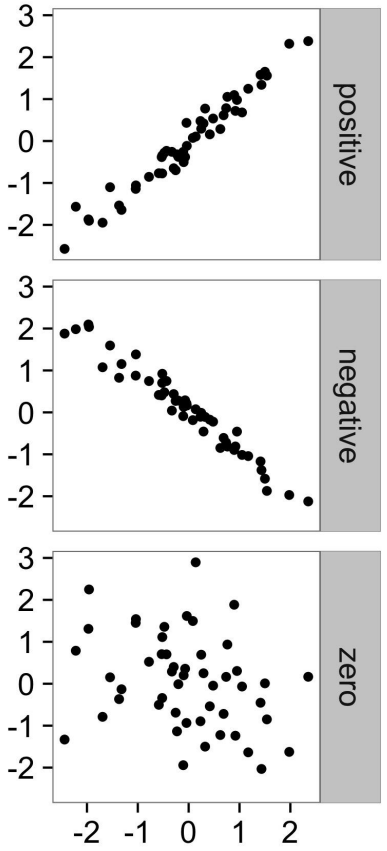
How would you describe this relationship?



Quantifying the relationship between two numerical values

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

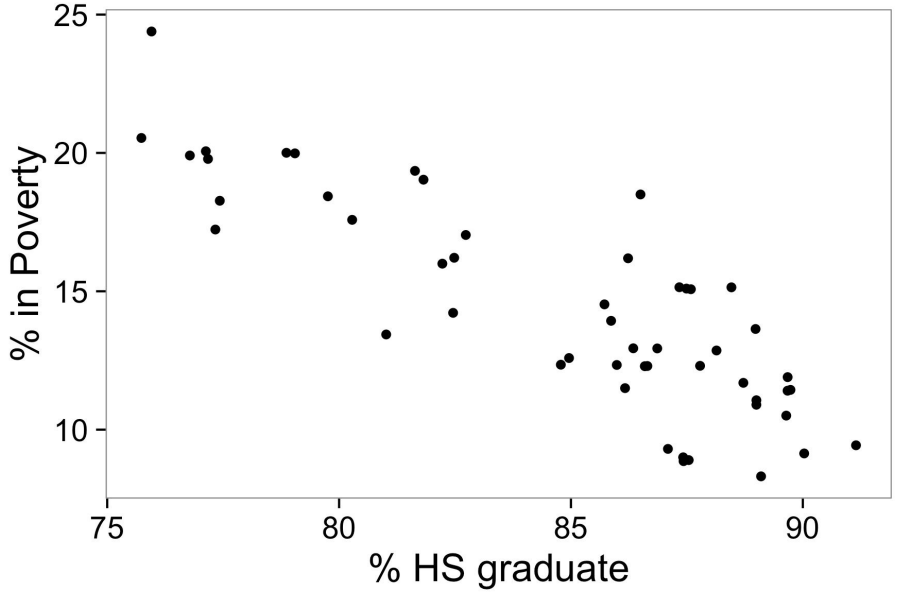
Correlation ranges from -1 (perfect negative) to +1 (perfect positive).
A value of 0 indicates no linear association.



Guess the correlation

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

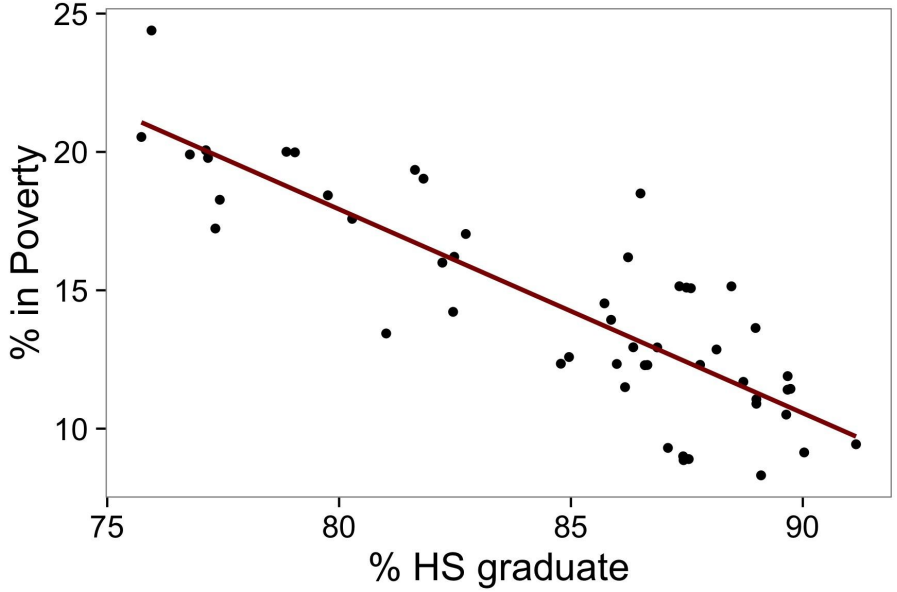
(a) .6

(b) -.85

(c) -.1

(d) .02

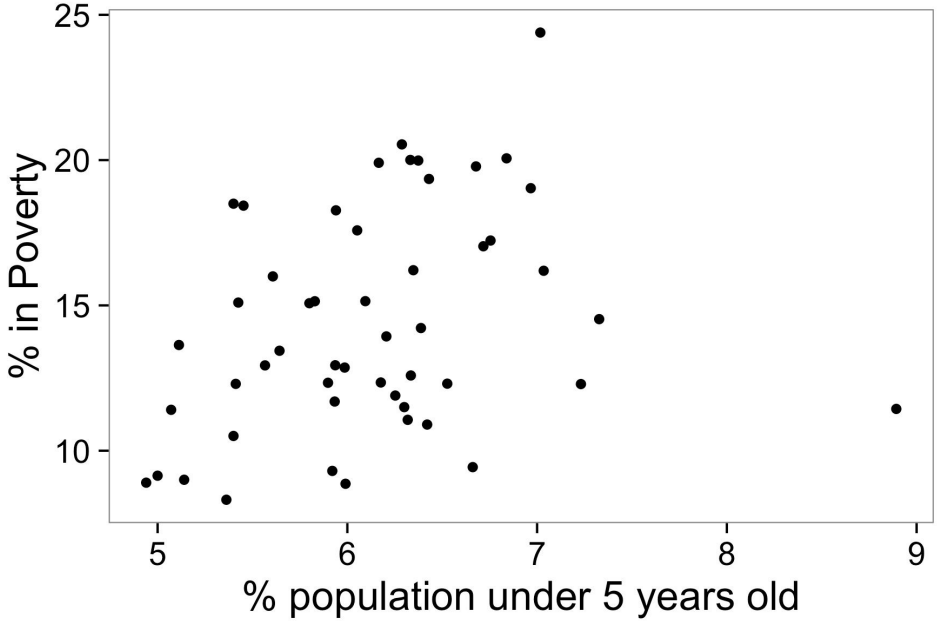
(e) -1.5



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?

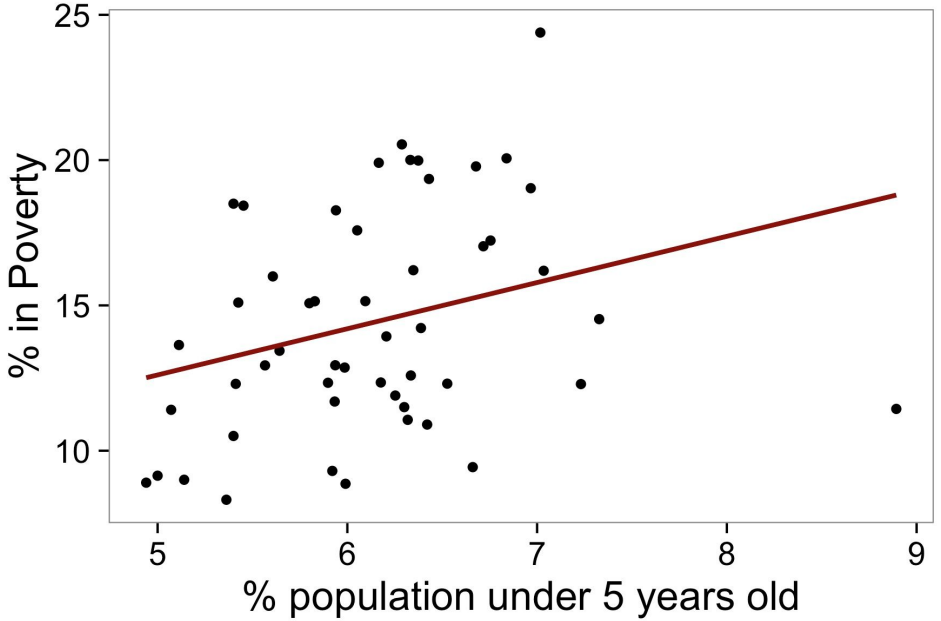
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.3



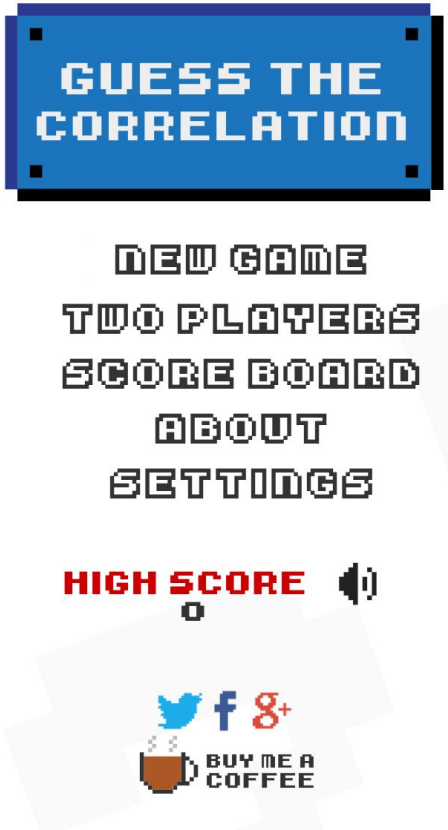
Guess the correlation

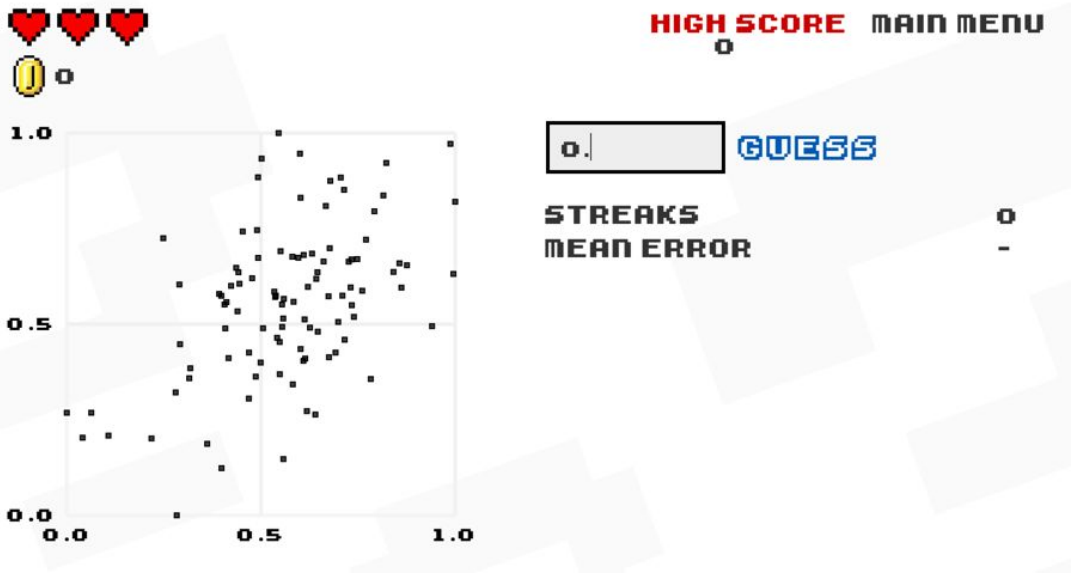
Which of these is your best guess for the correlation between poverty and and the proportion of the population under 5 years of age?

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

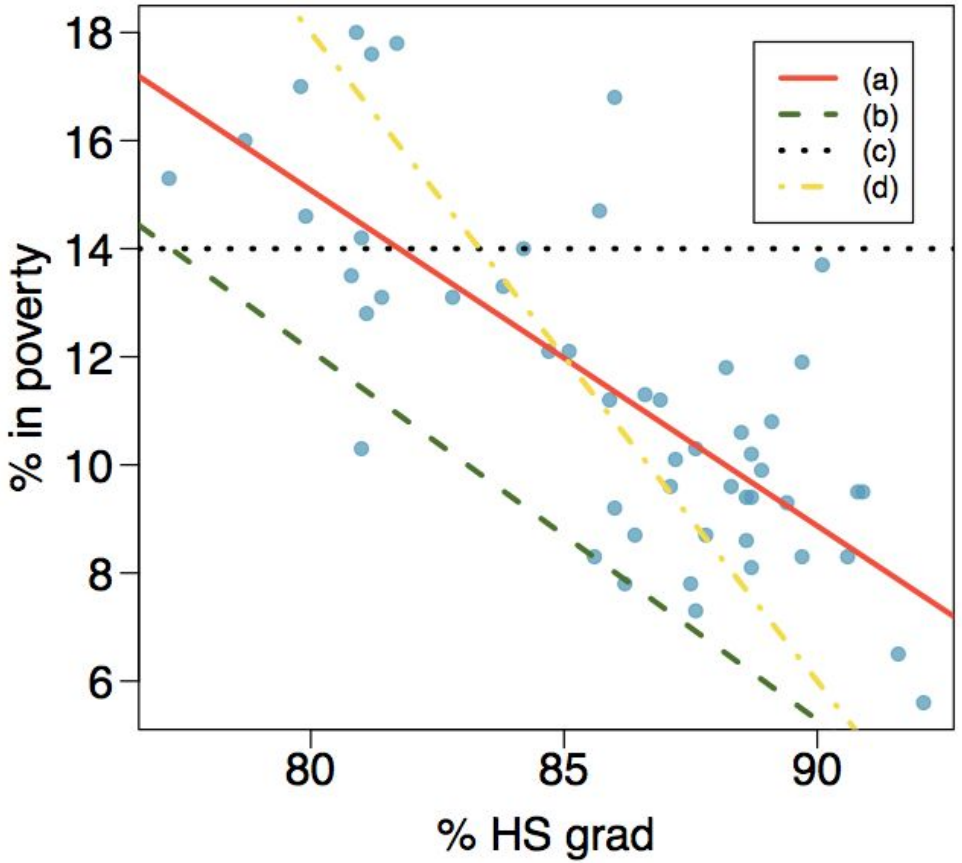


Play along at home: http://guessthecorrelation.com/





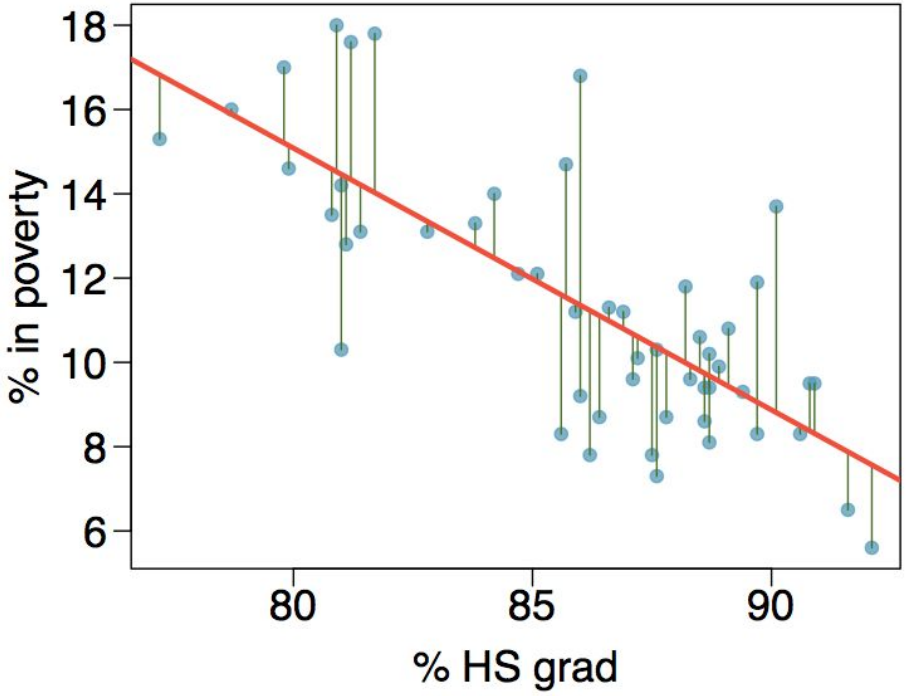
Which of these lines is the best representation of the trend?



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 that says that your data is composed of two things: A best-fit line + the residuals between each point and the line.



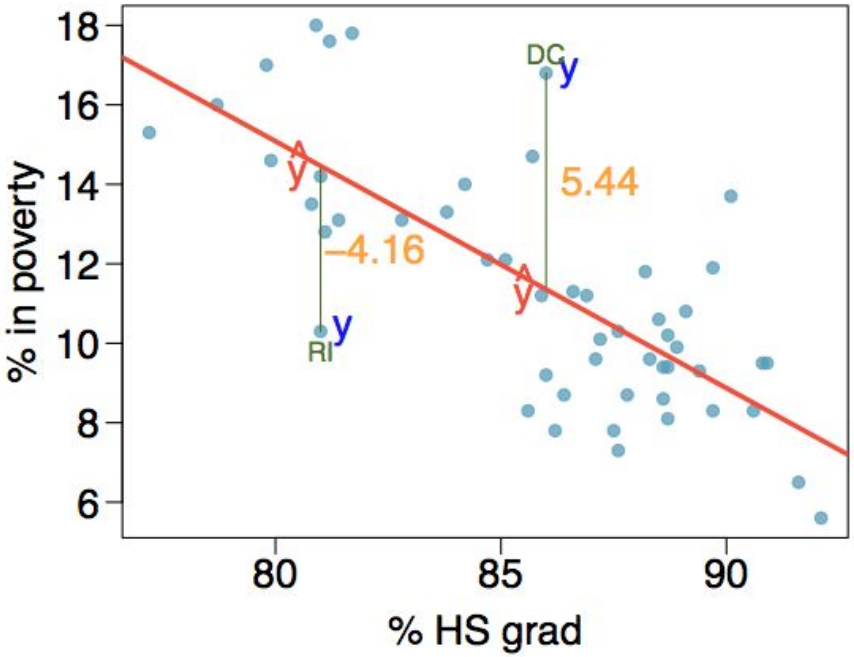
Residuals

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.

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



Key ideas

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