Sta 111 - Summer II 2017
Probability and Statistical Inference

18. Introduction to linear regression

Lu Wang

Duke University, Department of Statistical Science

July 31, 2017

Outline

1. Modeling numerical variables

2. Correlation describes the strength and direction of the linear association between two numerical variables

3. Least squares regression line minimizes squared residuals

4. Finding and interpreting the least squares line

5. Predict, but don't extrapolate

6. Homework 5

Modeling numerical variables

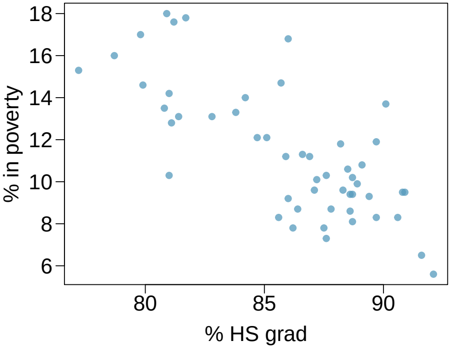
- So far we have worked with single numerical and categorical variables, and explored relationships between numerical and categorical, and two categorical variables.

- 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.

- In the next unit we'll learn to model numerical variables using many explanatory variables at once.

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between high-school (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?
Explanatory variable?
Relationship?

Quantifying the relationship - correlation

- *Correlation* describes the strength and direction of the *linear* association between two variables.
- Given a set of observations (x_1, y_1), (x_2, y_2),. . ., (x_n, y_n), the formula for computing the correlation is given by

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

  - This formula is rather complex, so we generally perform the calculations on a computer.
- Since the formula for calculating the correlation standardizes the variables, *changes in scale or units of measurement will not affect its value*.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.
- Correlation does not necessarily imply *causation*! - spurious correlations

Guessing the correlation

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

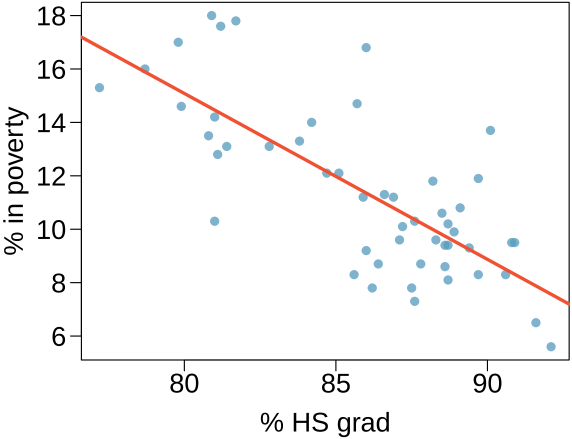
(a) 0.6

(b) -0.75

(c) -0.1

(d) 0.02

(e) -1.5



Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % female householder?

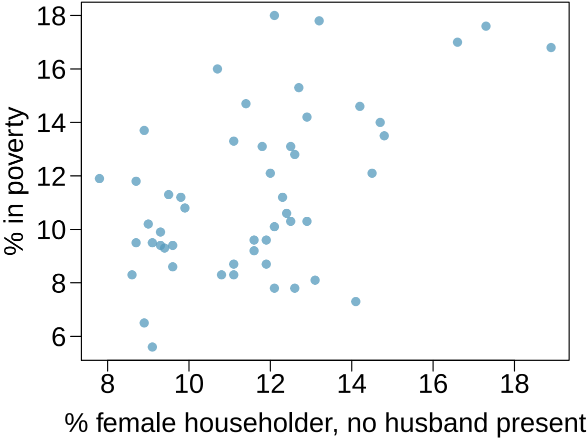
(a) 0.1

(b) -0.6

(c) -0.4

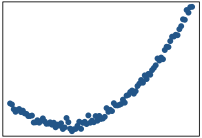
(d) 0.9

(e) 0.5

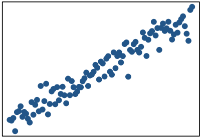


Assessing the correlation

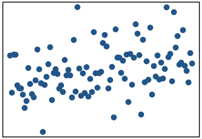
Which of the following has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



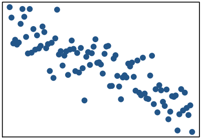
(a)



(b)



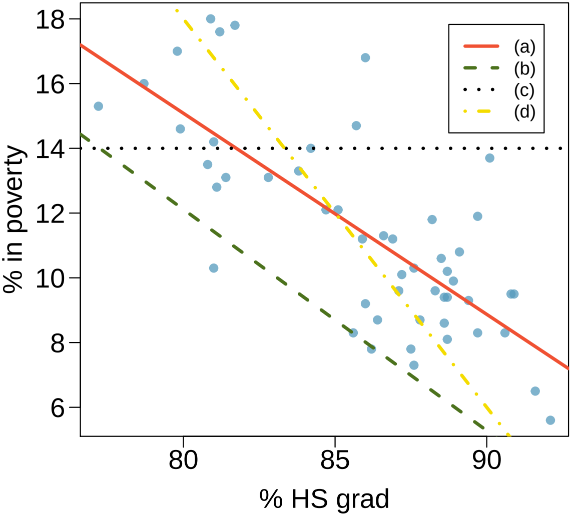
(c)



(d)

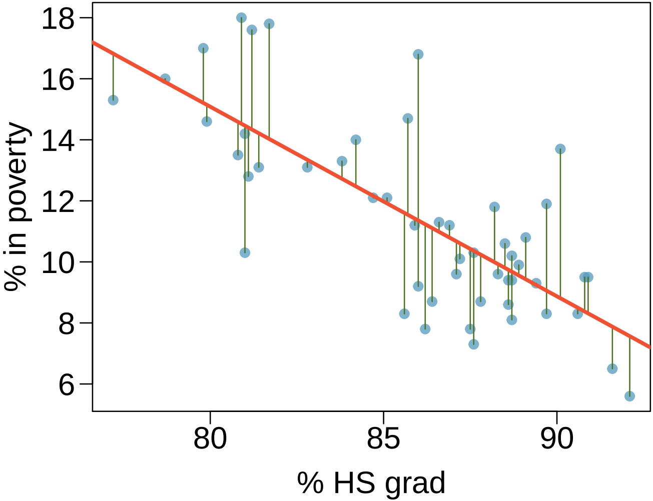
Eyeballing the line

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



Residuals

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

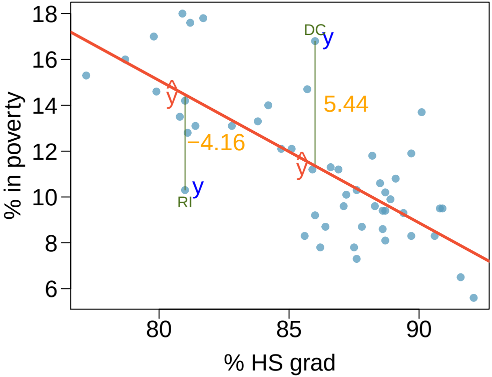


Residuals (cont.)

Residual

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

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

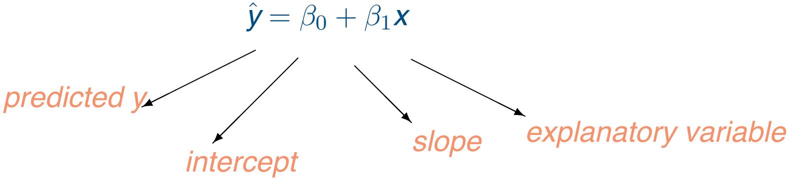


- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

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. The squared difference has nicer mathematical properties, e.g. continuously differentiable, while absolute values are difficult to work with in mathematics.
  2. Squaring the residuals gives more weight to large residuals, which may be helpful since in many applications, a residual twice as large as another is usually more than twice as bad

The least squares line



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

predicted y

intercept

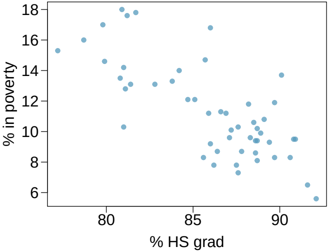
slope

explanatory variable

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	$corr = -0.75$

Slope

Slope

The slope of the regression can be calculated as

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

Interpretation: For each $\underline{\text{unit}}$ increase in $\underline{x}$, $\underline{y}$ is expected to be $\underline{\text{higher/lower}}$ on average by $\underline{\text{the slope}}$.

- 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

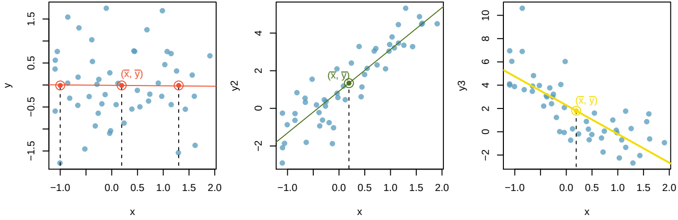
Intercept

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 $(\bar{x}, \bar{y})$.

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

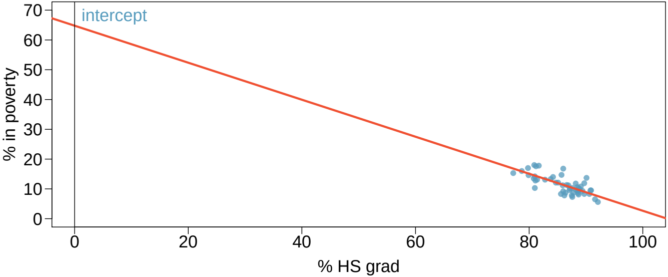
Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

Intuitively, whether there is a relationship between *x* and *y* or not, the best guess for $\hat{y}$ for any value of *x* is $\bar{y}$.



Intercept (cont.)

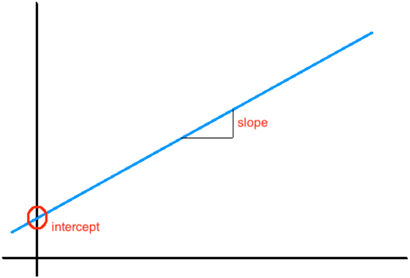
- $b_0 = 11.35 - (-0.62) \times 86.01 = 64.68$
- *Interpretation:* States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
- But since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.



Recap: Interpreting the slope and the intercept

- *Intercept:* When $\underline{x = 0}$, $\underline{y}$ is expected to equal the intercept.

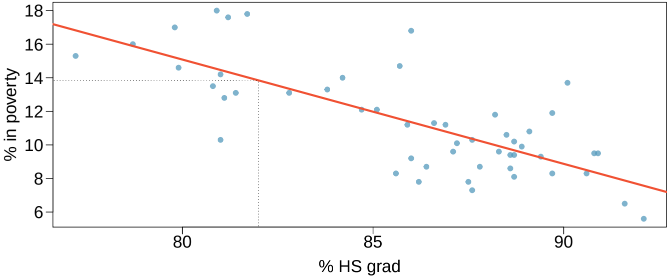
- *Slope:* For each unit increase in x, y is expected to increase / decrease on average by the slope.



Note: These statements are not causal, unless the study is a randomized controlled experiment.

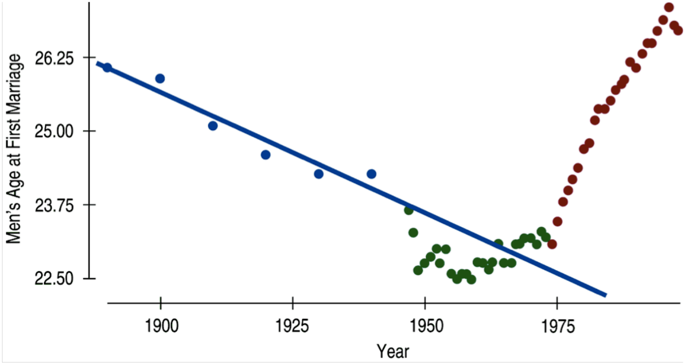
Prediction

- 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.



Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

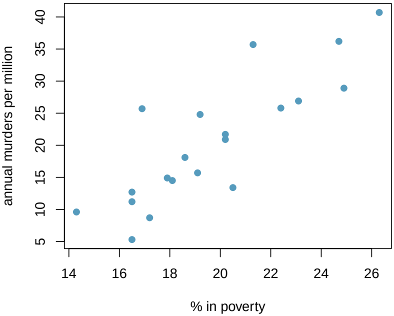
(a) 5%

(b) 15%

(c) 20%

(d) 26%

(e) 40%



Calculating predicted values

By hand: $\widehat{murder} = -29.91 + 2.56 \times poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata)
```

```
       1
21.28663
```

Homework 5

Graded questions:

- Ch 7: 7.14, 7.30, 7.34, 7.44

Practice questions:

- Relationship between two numerical variables: 7.1, 7.3, 7.7, 7.9, 7.11, 7.13, 7.15
- Linear regression with a single predictor: 7.17, 7.19, 7.25, 7.27, 7.29, 7.31, 7.33
- Inference for linear regression: 7.25, 7.37, 7.39, 7.41, 7.43