Unit 6: Introduction to linear regression

1. Introduction to regression

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

Prof. White

Slides posted at *https://www2.stat.duke.edu/courses/Summer18/sta104.001-1/*

Announcements

- Project Proposal is due tonight 11:55 pm.
- Problem Set 5 is due Saturday at 11:55 pm.
- Performance Assessment 5 is due Saturday at 11:55 pm.
- Lab 7 is due Monday at 12:45 pm
- PS 6 + PS 6 are due Tuesday (just so you're aware)

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.

Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

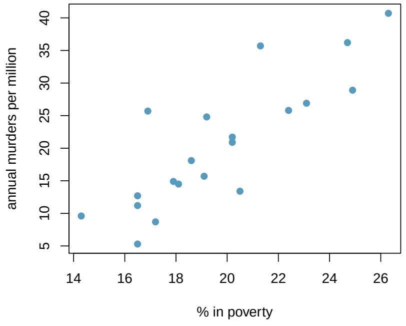
ⓐ -1.52

ⓑ -0.63

ⓒ -0.12

ⓓ 0.02

ⓔ 0.84

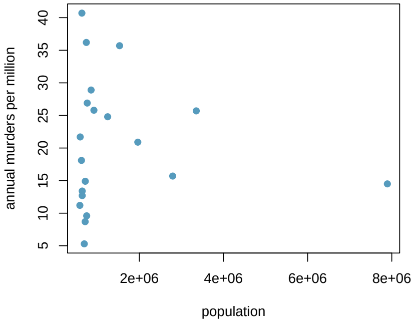


Guessing the correlation

Clicker question

Which of the following is the best guess for the correlation between annual murders per million and population size?

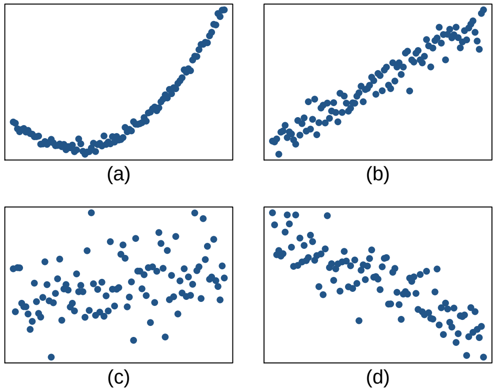
ⓐ -0.97
ⓑ -0.61
ⓒ -0.06
ⓓ 0.55
ⓔ 0.97



Assessing the correlation

Clicker question

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


(a) (b) (c) (d)

Play the game!

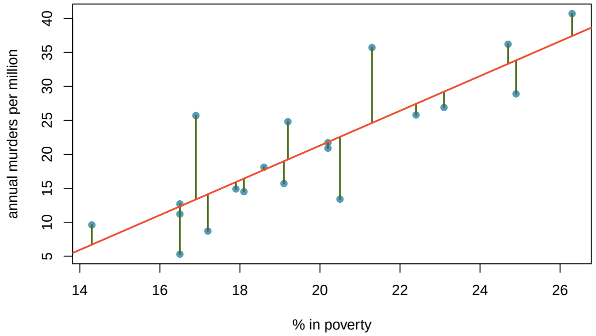
Send me a screen shot by midnight June 18 for extra credit on PS 6 (1 pt on the problem set).
http://guessthecorrelation.com/

Spurious correlations

Remember: correlation does not always imply causation!
http://www.tylervigen.com/

(2) Least squares line minimizes squared residuals

- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y: $e_i = y_i - \hat{y}_i$
- The least squares line minimizes squared residuals:
  - Population data: $\hat{y} = \beta_0 + \beta_1 x$
  - Sample data: $\hat{y} = b_0 + b_1 x$



(3) Interpreting the last squares line

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

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

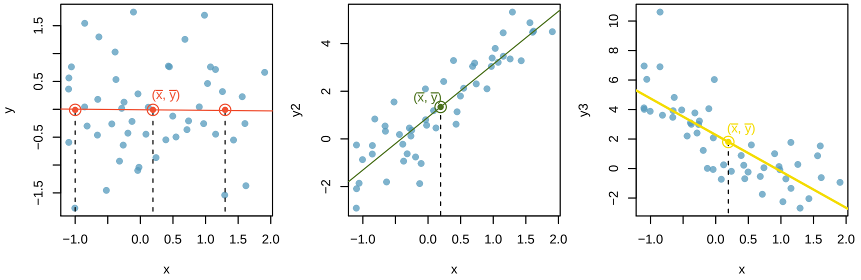
- *Intercept:* When $x = 0$, y is expected to equal the intercept.

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

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

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

- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.
- Even when there is a relationship between x and y ($b_1 \neq 0$), the best guess for $\hat{y}$ when $x = \bar{x}$ is still $\bar{y}$.



Application exercise: 6.1 Linear model

See course website for details

Clicker question

What is the interpretation of the slope?

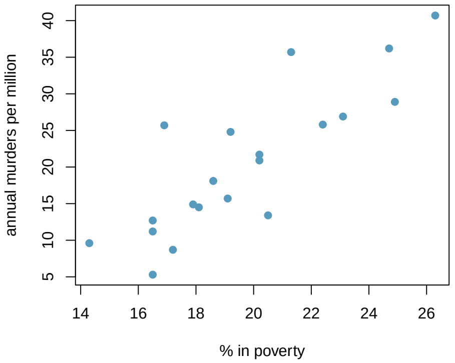
- (a) Each additional percentage in those living in poverty increases number of annual murders per million by 2.56.
- (b) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be higher by 2.56 on average.
- (c) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be lower by 29.91 on average.
- (d) For each percentage increase annual murders per million, the percentage of those living in poverty is expected to be higher by 2.56 on average.

Clicker question

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?

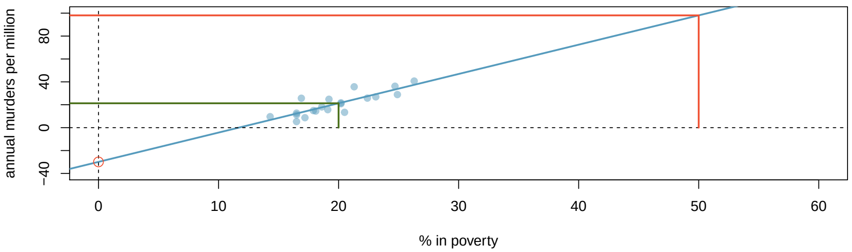
A district where % in poverty =

- (a) 5%
- (b) 15%
- (c) 20%
- (d) 26%
- (e) 40%



A note about the intercept

Sometimes the intercept might be an extrapolation: useful for adjusting the height of the line, but meaningless in the context of the data.



Calculating predicted values

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

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
```

Summary of main ideas

1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables
2. Least squares line minimizes squared residuals
3. Interpreting the least squares line
4. Predict, but don’t extrapolate