Statistical Inference

Introduction to Linear Regression

Behnam Bahrak

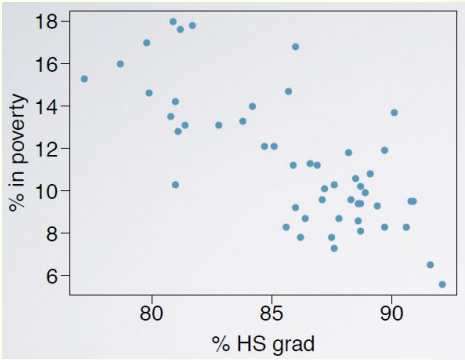
Spring 2020

Introduction to Linear Regression

- So far we worked with:
  - single numerical variable
  - single categorical variable
  - relationship between a numerical and a categorical variable
  - relationship between two categorical variables
- Linear Regression:
  - the relationship between two numerical variables.
  - Correlation: a measure of the strength of the linear relationship between two numerical variables

Poverty vs. HS Grad Rate

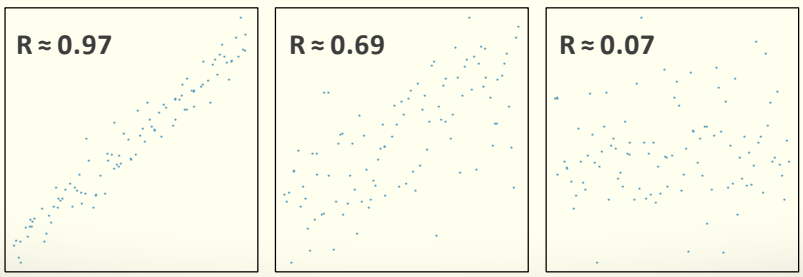
- Data: 50 states + DC
- Poverty line in US: income below $23,050 for a family of 4 in 2012
- Response?
  % in poverty
- Explanatory?
  % HS grad
- Relationship?
  linear, negative, moderately strong



Correlation

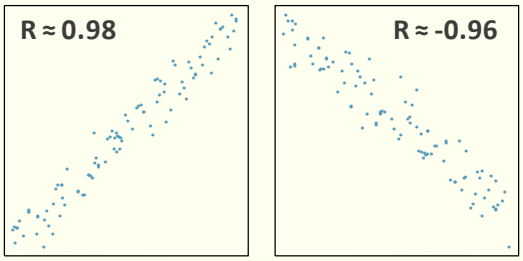
- Describes the strength of the linear association between two variables and is denoted as R
- Property 1. The magnitude (absolute value) of the correlation coefficient measures the strength of the **linear association** between two numerical variables

$$R = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$



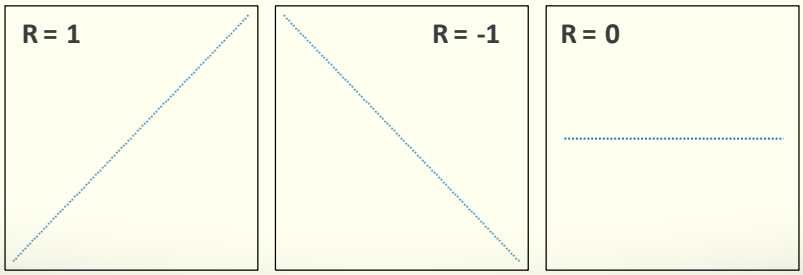
Properties of Correlation

- Property 2. The sign of the correlation coefficient indicates the direction of association



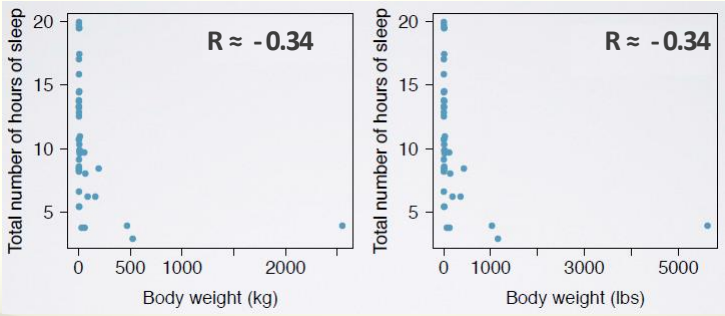
Properties of Correlation

- Property 3. The correlation coefficient is always between -1 (perfect negative linear association) and 1 (perfect positive linear association)
  - R = 0 indicates no **linear** relationship



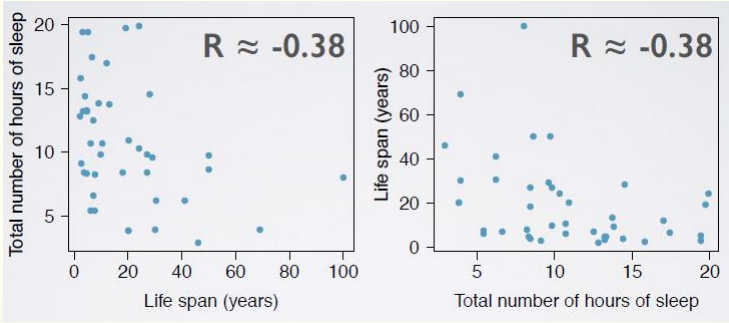
Properties of Correlation

- Property 4. The correlation coefficient is unitless, and is not affected by changes in the center or scale of either variable (such as unit conversions)



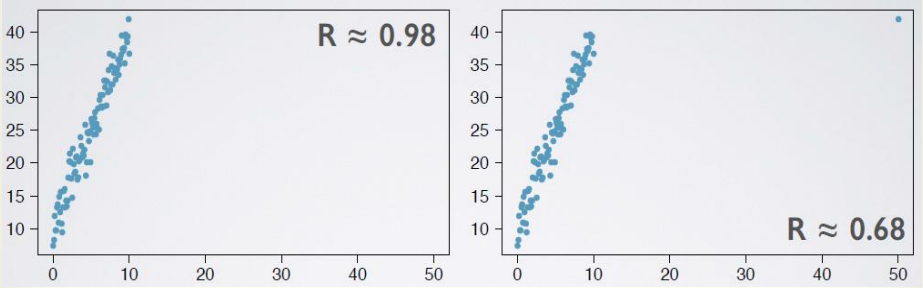
Properties of Correlation

- Property 5. The correlation of X with Y is the same as of Y with X.



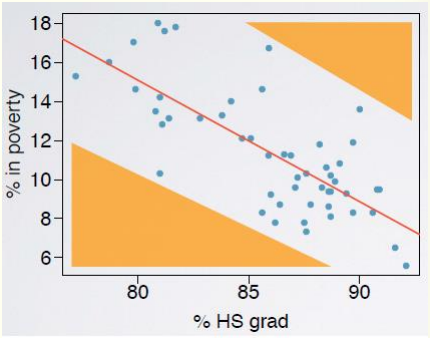
Properties of Correlation

- Property 6. The correlation coefficient is sensitive to outliers.



Question

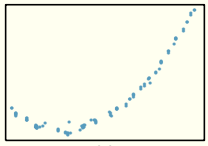
- Which of the following is the best guess for the correlation between % in poverty and % HS grad?
  - (a) 0.6
  - (b) -0.75
  - (c) -0.1
  - (d) 0.02
  - (e) -1.5

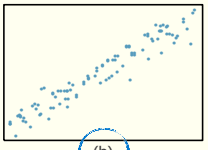


Question

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

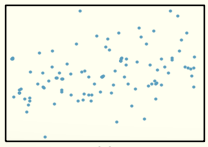
very strong, but not linear

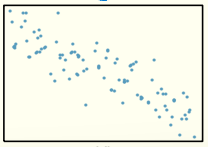

(a)


(b)

strongest linear

even weaker


(c)

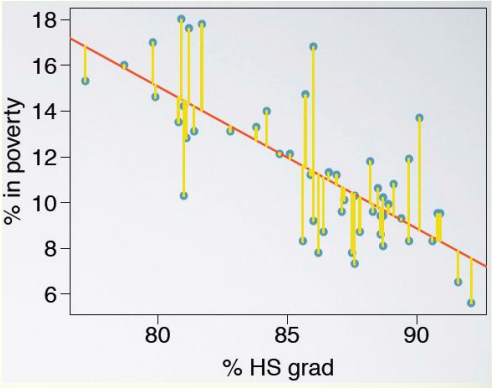

(d)

weaker

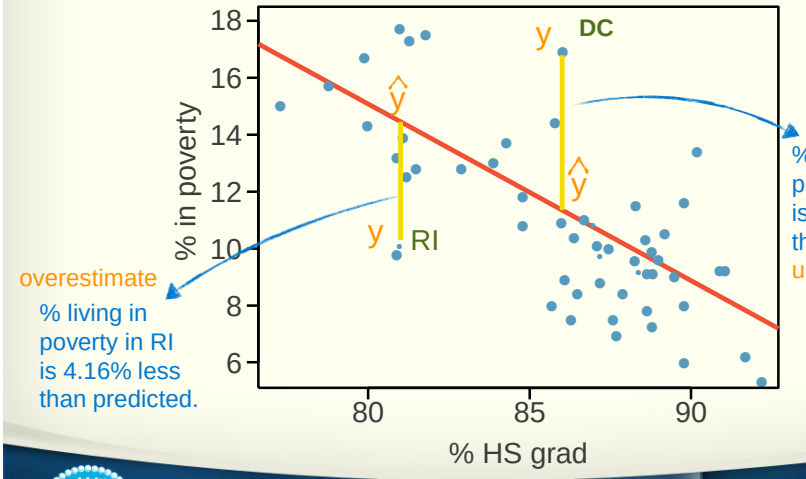
Residuals

- Leftovers from the model fit
- Data = Fit + Residual
- Difference between the observed and predicted y

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



Residuals



overestimate
% living in poverty in RI is 4.16% less than predicted.

% living in poverty in DC is 5.44% more than predicted.
underestimate

A measure for the best line

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

- Most commonly used
- Easier to compute by hand and using software
- In many applications, a residual twice as large as another is more than twice as bad

Least Square Line

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

predicted response

intercept

slope

explanatory

Notation

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

Estimating the regression parameters: slope

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

s_x : SD of x

s_y : SD of y

$R = cor(x, y)$

Example

- The standard deviation of % living in poverty is 3.1%, and the standard deviation of % HS graduates is 3.73%. Given that the correlation between these variable is -0.75, what is the slope of the regression line for predicting % living poverty from % HS graduates?

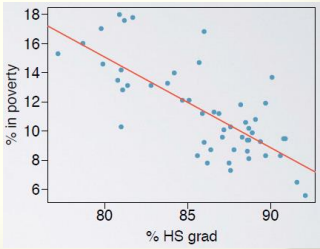
$S_y = 3.1\%$

$S_x = 3.73\%$

$R = -0.75$

$$b_1 = \frac{S_y}{S_x} \times R = \frac{3.1}{3.73} \times (-0.75) = -0.62$$

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

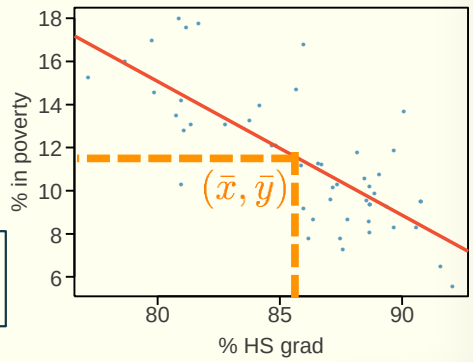


Estimating the regression parameters: intercept

- The least squares line always goes through $(\bar{x}, \bar{y})$

$$\bar{y} \quad \hat{y} = b_0 + b_1 x \quad \bar{x}$$

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



Example

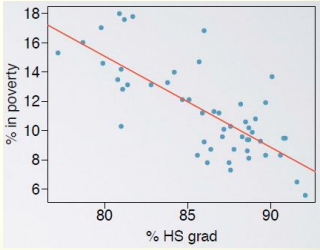
- Given that the average % living in poverty is 11.35%, and the average % HS graduates is 86.01%, what is the intercept of the regression line for predicting % living poverty from % HS graduates?

$\bar{y} = 11.35\%$

$\bar{x} = 86.01\%$

$b_0 = \bar{y} - b_1\bar{x} = 11.35 - (-0.62)(86.01) = 64.68$

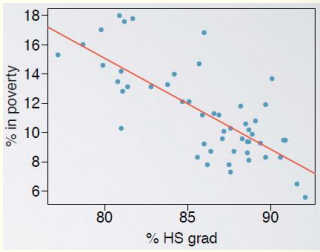
States with no HS graduates are expected on average to have 64.68% of their residents living below the poverty line.



Example

$\widehat{\%\text{ in poverty}} = 64.68 - 0.62\ \%\text{ HS grad}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.78	6.80	9.52	0.00
hsgrad	-0.62	0.08	-7.86	0.00



Recap

- intercept: When $x = 0, y$ is expected to equal the intercept.
  - may be meaningless in context of the data, and only serve to adjust the height of the line
- slope: For each unit increase in x, y is expected to be higher/lower on average by the slope.