DSC 10, Spring 2018
Lecture 23

Least Squares

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Regression Line
Slope & Intercept

Regression Line Equation

In standard units, the regression line has this equation:

Regression Line

$$y_{(\text{su})} = r \times x_{(\text{su})}$$

Correlation

In original units, the regression line has this equation:

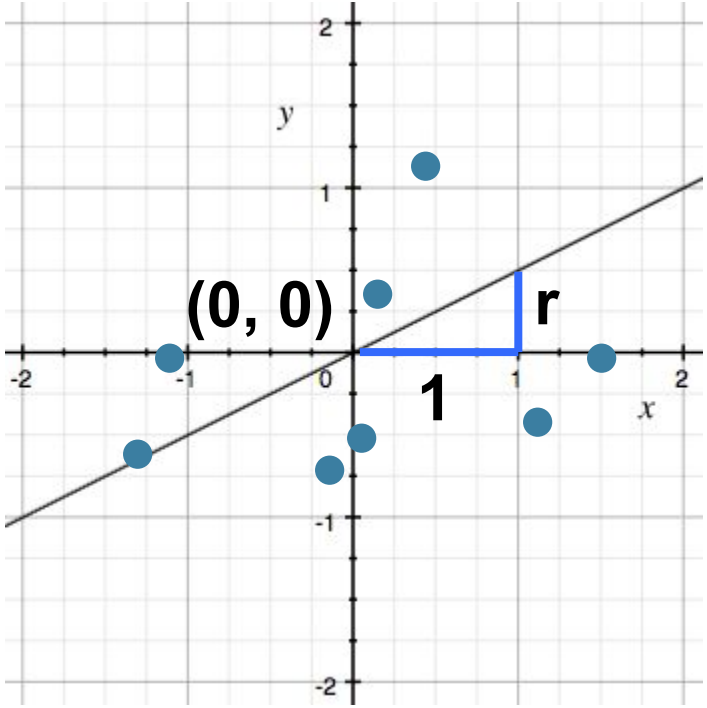
$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

y in standard units

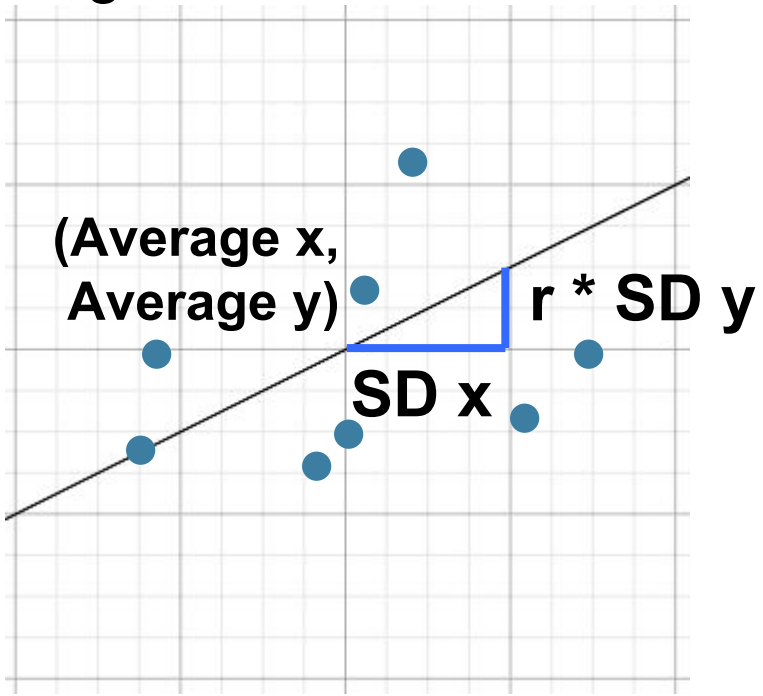
x in standard units

Regression Line

Standard Units



Original Units



Slope and Intercept

estimate of y = slope * x + intercept

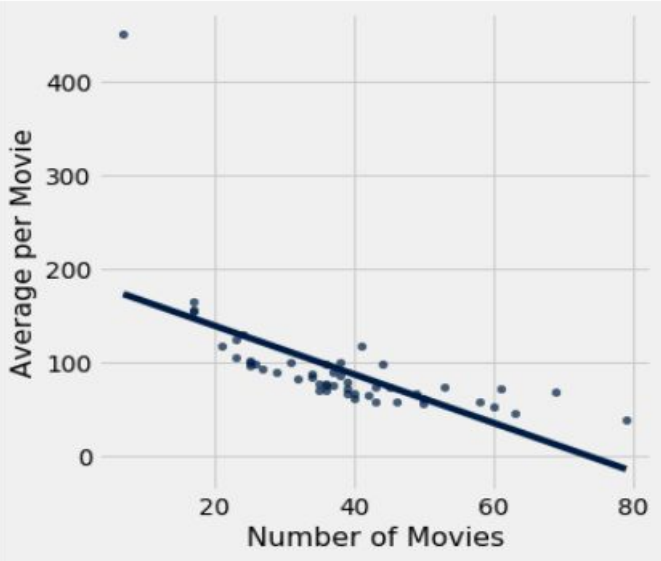
$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

Discussion Question 1

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$



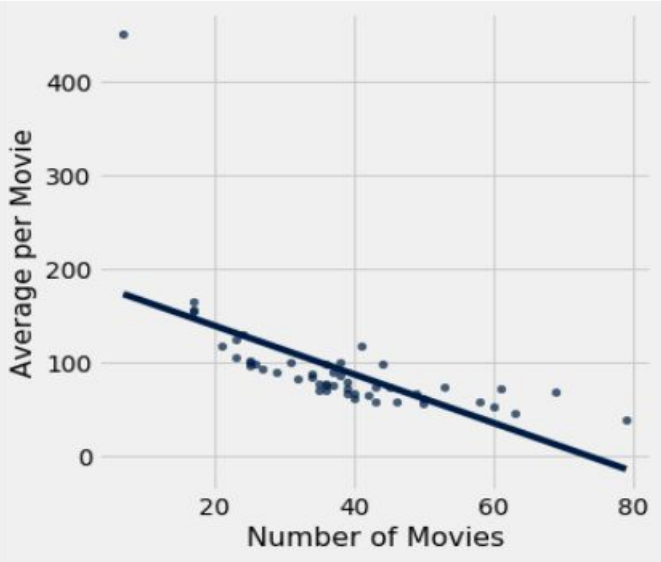
What are the units of the slope of the regression line?

A. million dollars per movie
B. movie per million dollars
C. million dollars per movie per movie
D. movie per million dollars per movie
E. none of the above

Discussion Question 2

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x} = -2.5$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$



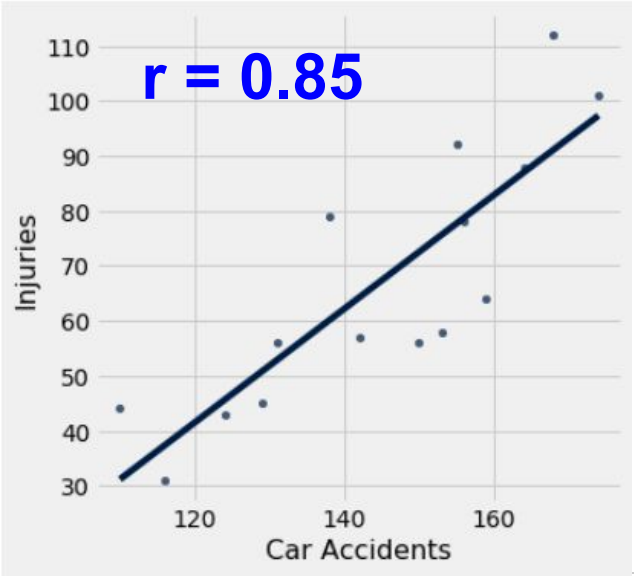
Actor A appeared in *m* movies, which made an average of 87 million each. If Actor B appeared in *m+2* movies, estimate his average earnings per movie.

A. 82 million dollars
B. 84 million dollars
C. 84.5 million dollars
D. 89 million dollars
E. 89.5 million dollars

Discussion Question 3

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$



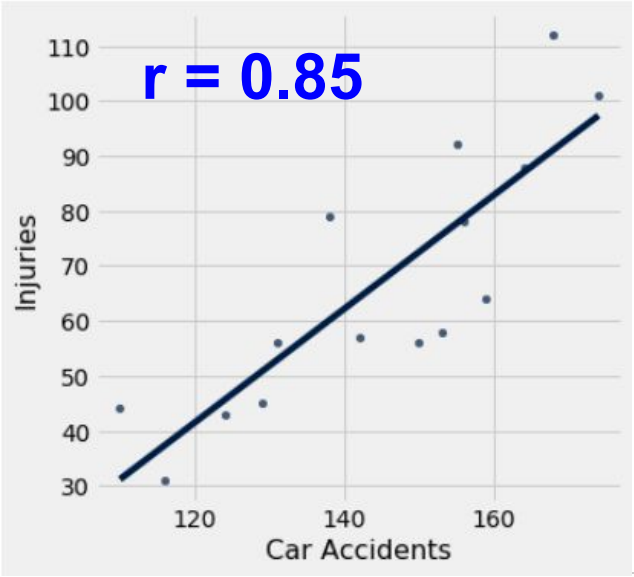
If r = 0.85, and we know that 150 car accidents occurred this month, estimate the number of car accident-related injuries this month.

A. 150 * 0.85
B. 150 / 0.85
C. 150 * sqrt(1-0.85)
D. 150 ** 0.85
E. None of the above

Discussion Question 3

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$



If r = 0.85, and we know that 150 car accidents occurred this month, estimate the number of car accident-related injuries this month.

	mean	sd
Car Accidents	144	19
Injuries	70	23

Least Squares

(Demo)

Error in Estimation

- **error = actual value - estimate**

  = actual value - predicted value

- Typically, some errors are positive and some negative
  - What does a positive error mean? negative?

Error in Estimation

- **error = actual value - estimate**

  = actual value - predicted value

- Typically, some errors are positive and some negative
  - What does a positive error mean? negative?
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square erro**r (rmse)

(Demo)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

Numerical Optimization

- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- Idea: Given a function that returns a real number,
  - Search among all possible inputs to the function
  - Find the input to the function that results in the function returning the smallest possible real number
  - Approximate because we cannot search *all* possible inputs

Numerical Optimization of MSE

If the function **`mse(a, b)`** returns the mse of estimation using the line "estimate = a*x* + b",

- then **`minimize(mse)`** returns array $[a_0, b_0]$
- a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

(Demo)

Discussion question

```
def my_func(c):
    if c < -2:
        return 4
    elif c > 2:
        return 4
    else:
        return abs(c)+2
```

Pick the option that best completes the sentence:

"The expression **minimize(my_func)** evaluates to…"

A: -3
B: 0
C: 1
D: 2
E: 4

Residuals

Residuals

- Error in regression estimate
- One residual corresponding to each point (x, y)
- residual = observed y - regression estimate of y
  = observed y - height of regression line at x
  = vertical *difference* between point and line

Residual Plot

For good regressions, the residual plot

- Should look like a blob
- About half above and half below the horizontal line at 0
- Similar vertical spread throughout
- No pattern

For bad regressions…?

Dugong

(Demo)

Spotting Problems

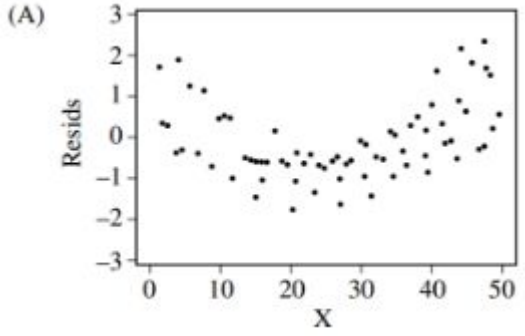
Residual plots can be used to detect:

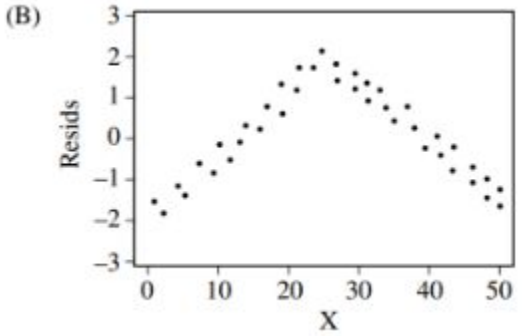
- Non-linearity
  - Shape of scatter plot is curved, not a straight line
- Heteroscedasticity
  - Uneven spread

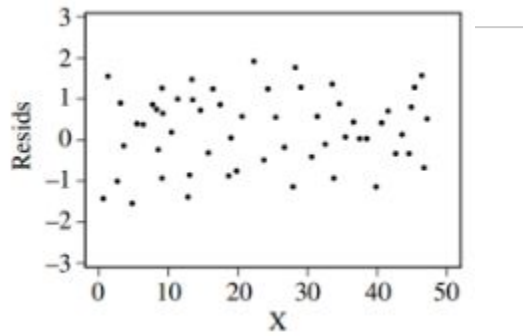
Residual Plots are Flat Overall

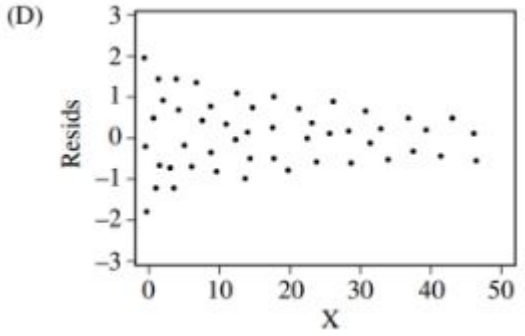
- For any regression, no matter what the shape of the original scatterplot:
- The residual plot will not have any linear trend, neither upwards nor downwards.
- The correlation between the residuals and the predictor variable is 0.

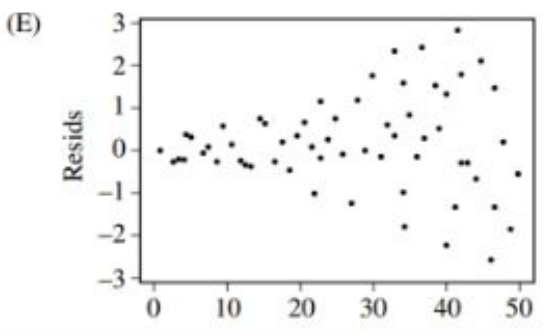
Discussion Question











Which of the plots provides the strongest evidence that the regression line is an appropriate model?