Simple Linear Regression

Introduction

Prof. Maria Tackett

STA 210

[Click for PDF of slides]

Topics

- Use simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.

Topics

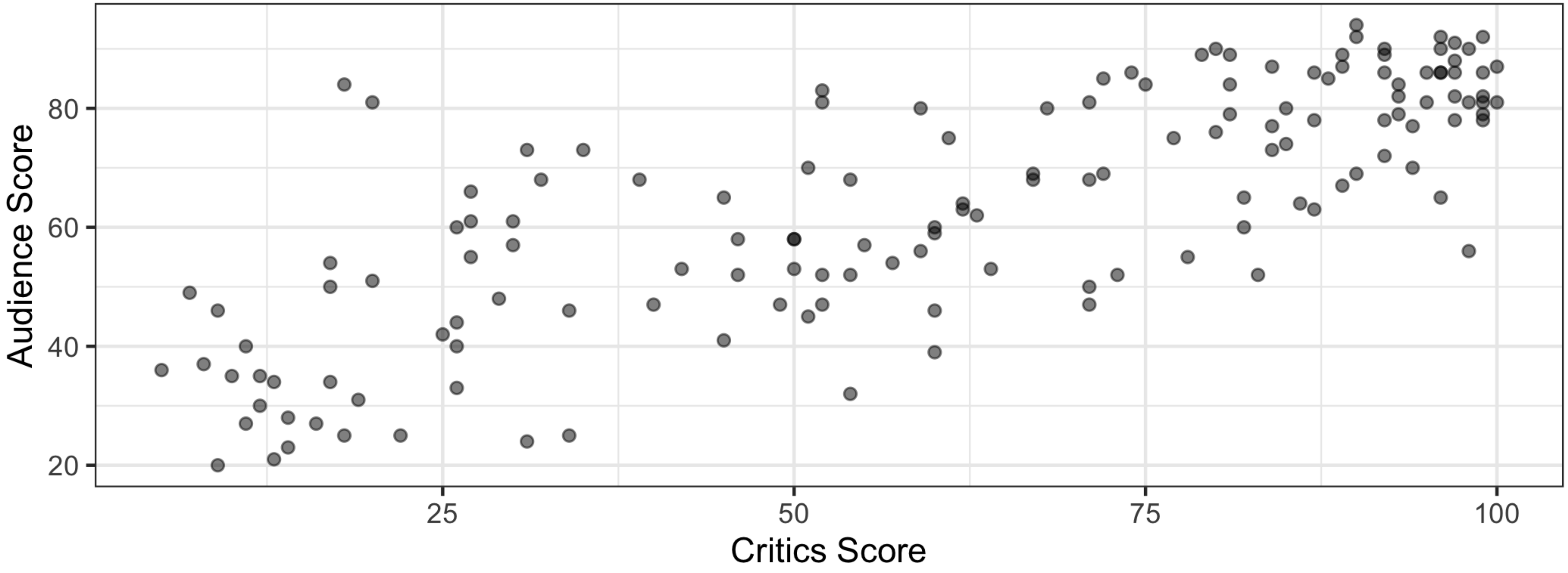
- Use simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.
- Estimate the slope and intercept of the regression line using the least squares method.

Topics

- Use simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.
- Estimate the slope and intercept of the regression line using the least squares method.
- Interpret the slope and intercept of the regression line.

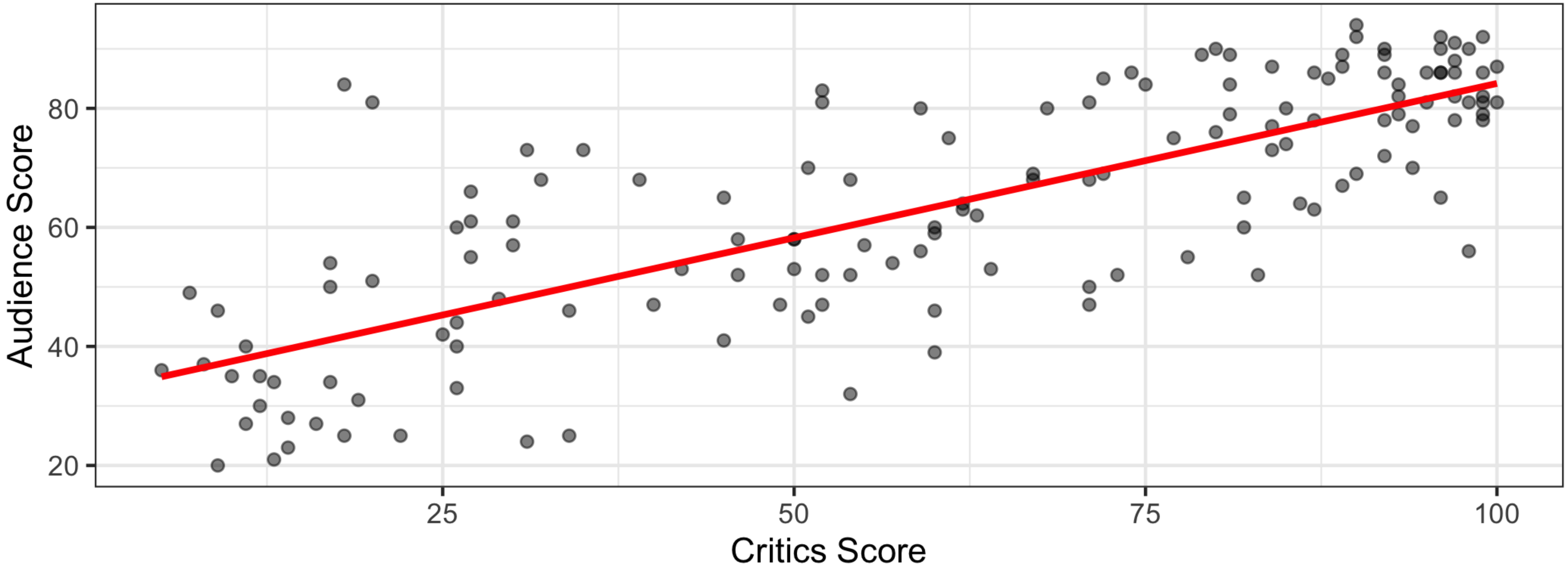
Movie ratings data

The data set contains the "Tomatometer" score (**`critics`**) and audience score (**`audience`**) for 146 movies rated on rottentomatoes.com.



Movie ratings data

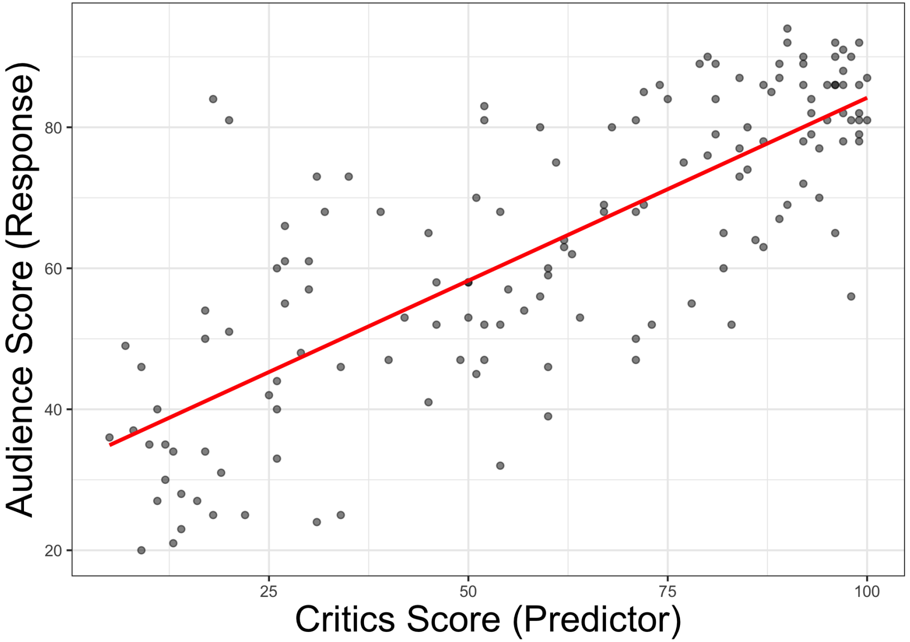
We want to fit a line to describe the relationship between the critics score and audience score.



Terminology

The **response, *Y***, is the variable describing the outcome of interest.

The **predictor, *X***, is the variable we use to help understand the variability in the response.



Regression model

A regression model is a function that describes the relationship between the response, Y, and the predictor, X.

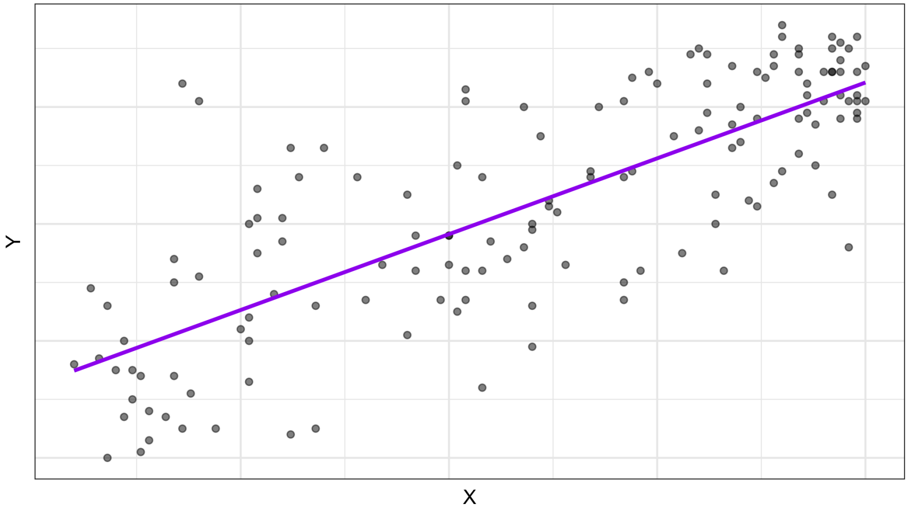
$$
\begin{aligned}
Y &= \textbf{Model} + \text{Error} \\
&= \mathbf{f(X)} + \epsilon \\
&= \boldsymbol{\mu_{Y|X}} + \epsilon
\end{aligned}
$$

STA 210

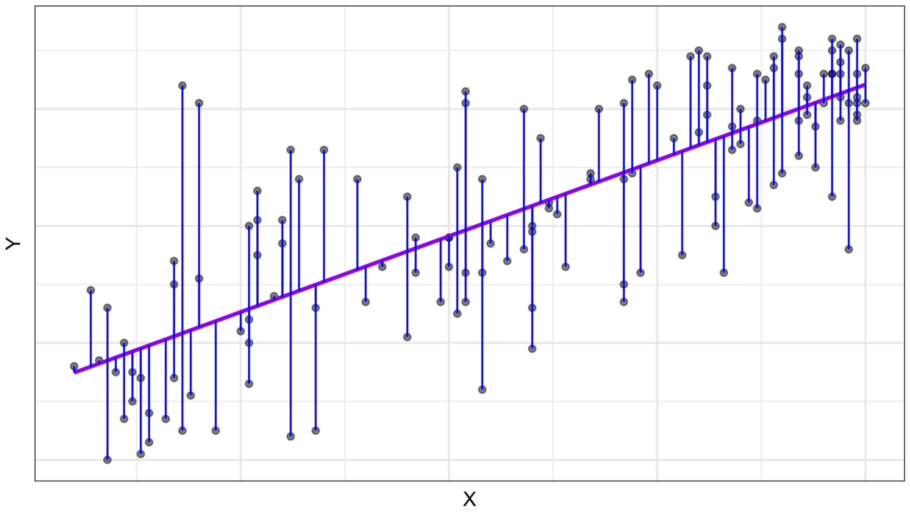
$$Y = \textbf{Model} + \text{Error}$$

$$= \mathbf{f(X)} + \epsilon$$

$$= \boldsymbol{\mu_{Y|X}} + \epsilon$$



$$
\begin{aligned}
Y &= \textbf{Model} + \textbf{Error} \\
&= \mathbf{f(X)} + \epsilon \\
&= \mu_{Y|X} + \epsilon
\end{aligned}
$$



STA 210

Simple linear regression

When we have a quantitative response, Y, and a single quantitative predictor, X, we can use a **simple linear regression** model to describe the relationship between Y and X.

$$Y = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{X} + \epsilon$$

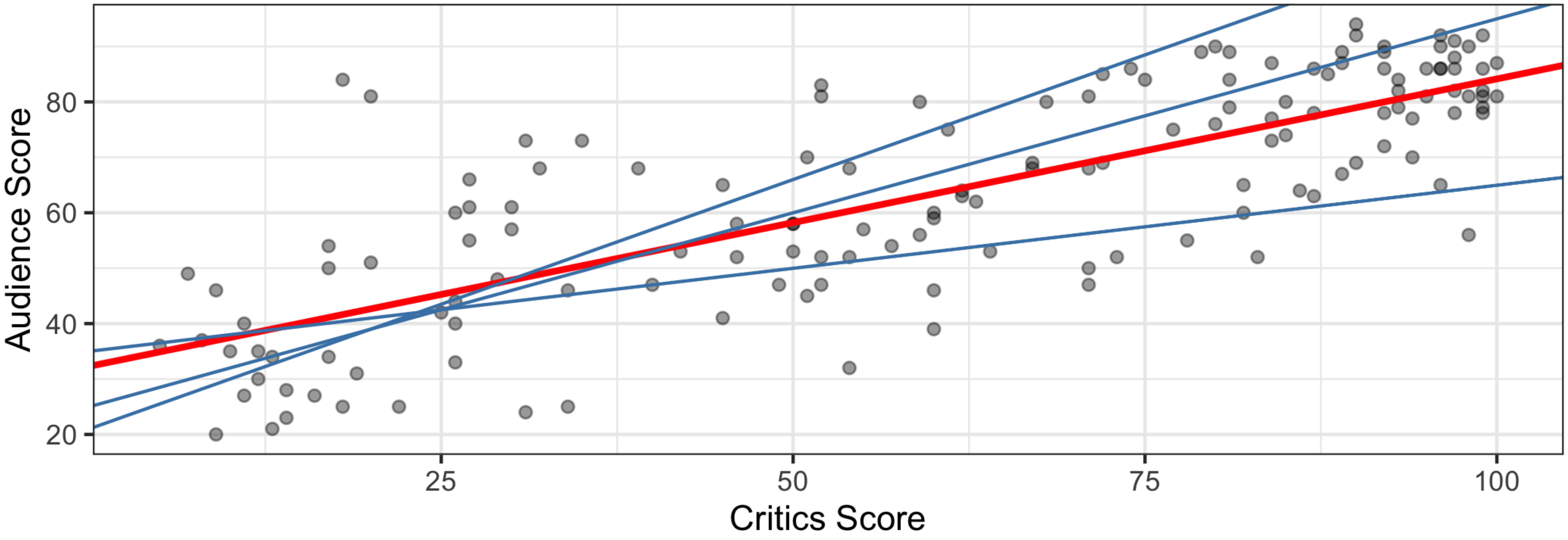
$\boldsymbol{\beta}_1$: Slope $\qquad$ $\boldsymbol{\beta}_0$: Intercept

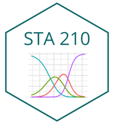
STA 210

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

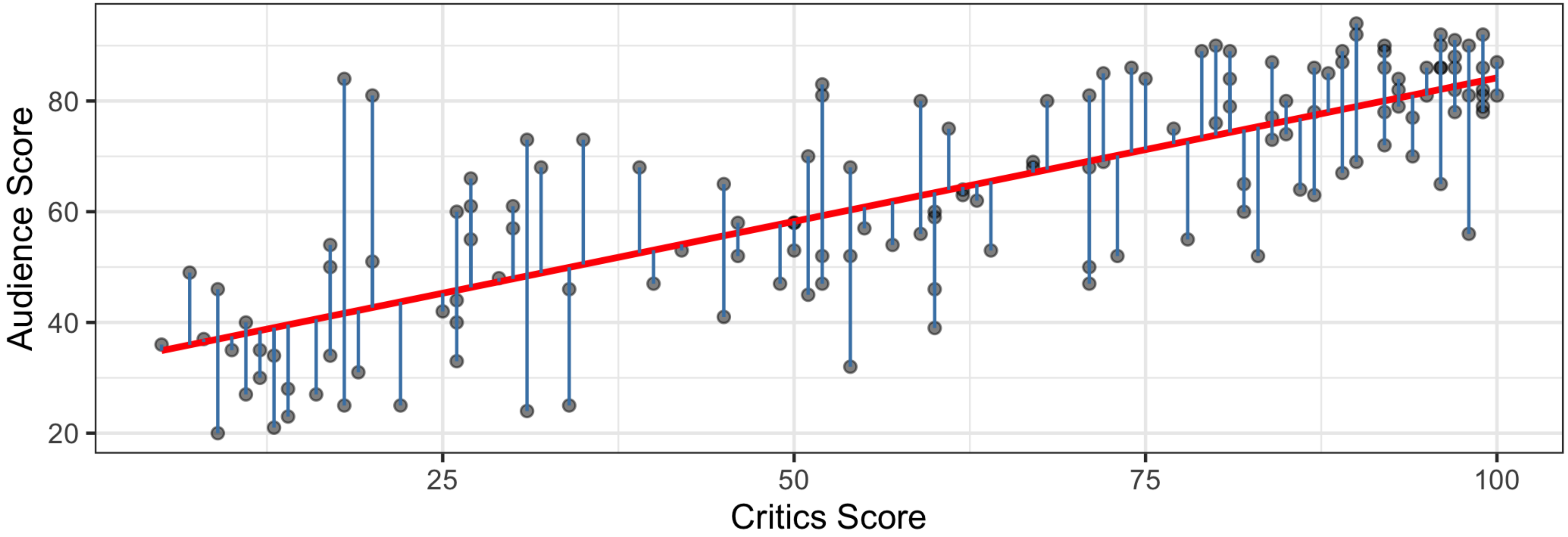
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

How do we choose values for $\hat{\beta}_1$ and $\hat{\beta}_0$?





Residuals



$$\text{residual} = \text{observed} - \text{predicted} = y - \hat{y}$$

STA 210

Least squares line

Least squares line

- The residual for the i^{th} observation is

$$e_i = \text{observed} - \text{predicted} = y_i - \hat{y}_i$$

Least squares line

- The residual for the i^{th} observation is

$$e_i = \text{observed} - \text{predicted} = y_i - \hat{y}_i$$

- The **sum of squared residuals** is

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

Least squares line

- The residual for the i^{th} observation is

$$e_i = \text{observed} - \text{predicted} = y_i - \hat{y}_i$$

- The **sum of squared residuals** is

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- The **least squares line** is the one that minimizes the sum of squared residuals

STA 210

Estimating the slope

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

STA 210

Estimating the slope

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

$$s_X = 30.169$$

$$s_Y = 20.024$$

$$r = 0.781$$

STA 210

Estimating the slope

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

$$s_X = 30.169$$

$$s_Y = 20.024$$

$$r = 0.781$$

$$\hat{\beta}_1 = 0.781 \times \frac{20.024}{30.169}$$

$$= \mathbf{0.518}$$

Estimating the intercept

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Estimating the intercept

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\bar{x} = 60.850$$

$$\bar{y} = 63.877$$

$$\hat{\beta}_1 = 0.518$$

STA 210

Estimating the intercept

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\bar{x} = 60.850$$

$$\bar{y} = 63.877$$

$$\hat{\beta}_1 = 0.518$$

$$\hat{\beta}_0 = 63.877 - 0.518 \times 60.850$$

$$= \mathbf{32.296}$$

STA 210

Interpreting slope & intercept

$$\widehat{\text{audience}} = 32.296 + 0.518 \times \text{critics}$$

STA 210

Interpreting slope & intercept

$$\widehat{\text{audience}} = 32.296 + 0.518 \times \text{critics}$$

Slope: For every one point increase in the critics score, we expect the audience score to increase by 0.518 points, on average.

STA 210

Interpreting slope & intercept

$$\widehat{\text{audience}} = 32.296 + 0.518 \times \text{critics}$$

Slope: For every one point increase in the critics score, we expect the audience score to increase by 0.518 points, on average.

Intercept: If the critics score is 0 points, we expect the audience score to be 32.296 points.

Does it make sense to interpret the intercept?

Does it make sense to interpret the intercept?

✅ **Interpret the intercept if**

- the predictor can feasibly take values equal to or near zero.
- there are values near zero in the data.

STA 210

Does it make sense to interpret the intercept?

✅ **Interpret the intercept if**

- the predictor can feasibly take values equal to or near zero.
- there are values near zero in the data.

🛑 Otherwise, don't interpret the intercept!

STA 210

Recap

STA 210

Recap

- Used simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.

Recap

- Used simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.
- Used the least squares method to estimate the slope and intercept.

Recap

- Used simple linear regression to describe the relationship between a quantitative predictor and quantitative response variable.
- Used the least squares method to estimate the slope and intercept.
- We interpreted the slope and intercept.
  - **Slope**: For every one unit increase in x, we expect y to change by $\hat{\beta}_1$ units, on average.
  - **Intercept**: If x is 0, then we expect y to be $\hat{\beta}_0$ units.

STA 210