Simple Linear Regression

Prof. Maria Tackett

01.22.20

STA 210

[Click for PDF of slides]

Announcements

- Lab 01 due **TODAY at 11:59p**
- HW 01 due **Wed, Jan 29 at 11:59p**
- Lab groups start tomorrow!
- Daily engagement survey starts today at the end of class (check your email)
- Reading for today & Monday
- Check email for info about Jan 29 class

Check in

- Any questions from last class?
- Any questions about the lab?
- Any questions about course logistics?

In-class exercise:

- Answer the questions: http://bit.ly/sta210-sp20-indep
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

03:00

STA 210

Today's Agenda

- Simple Linear Regression
  - Estimating & interpreting coefficients
  - Assessing model fit: R^2
  - Residuals and model assumptions
  - Prediction

STA 210

Packages and Data

```r
library(tidyverse)
library(broom)
library(modelr)
library(knitr)
library(fivethirtyeight) #fandango dataset
library(cowplot) #plot_grid() function
```

```r
movie_scores <- fandango %>%
  rename(critics = rottentomatoes,
         audience = rottentomatoes_user)
```

Motivating Regression Analysis

rottentomatoes.com

Can the ratings from movie critics be used to predict what movies the audience will like?

DORA AND THE LOST CITY OF GOLD

Critics Consensus

Led by a winning performance from Isabela Moner, *Dora and the Lost City of Gold* is a family-friendly adventure that retains its source material's youthful spirit.

CERTIFIED FRESH Rotten Tomatoes **83%** — **TOMATOMETER** Total Count: 129

88% — **AUDIENCE SCORE** Verified Ratings: 5,605

NEW **MORE INFO**

ALADDIN

Critics Consensus

Aladdin retells its classic source material's story with sufficient spectacle and skill, even if it never approaches the dazzling splendor of the animated original.

57% — **TOMATOMETER** Total Count: 347

94% — **AUDIENCE SCORE** Verified Ratings: 58,961

NEW **MORE INFO**

Critic vs. Audience Ratings

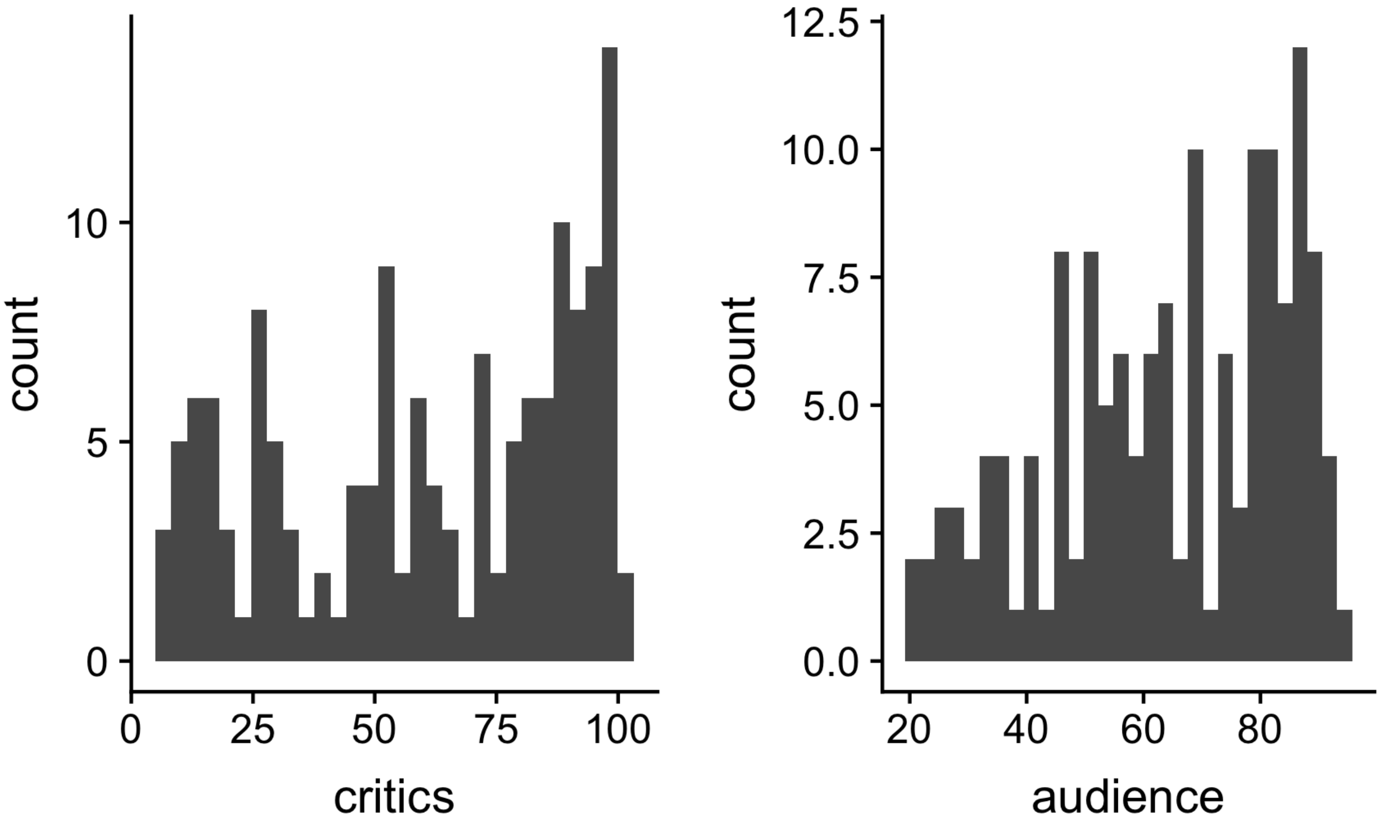
- To answer this question, we will analyze the critic and audience scores from rottentomatoes.com.
    - The data was first used in the article [Be Suspicious of Online Movie Ratings, Especially Fandango's].
- Variables:
    - **`critics`**: Tomatometer score for the film (0 - 100)
    - **`audience`**: Audience score for the film (0 - 100)

`glimpse` of the data

```
glimpse(movie_scores)
```

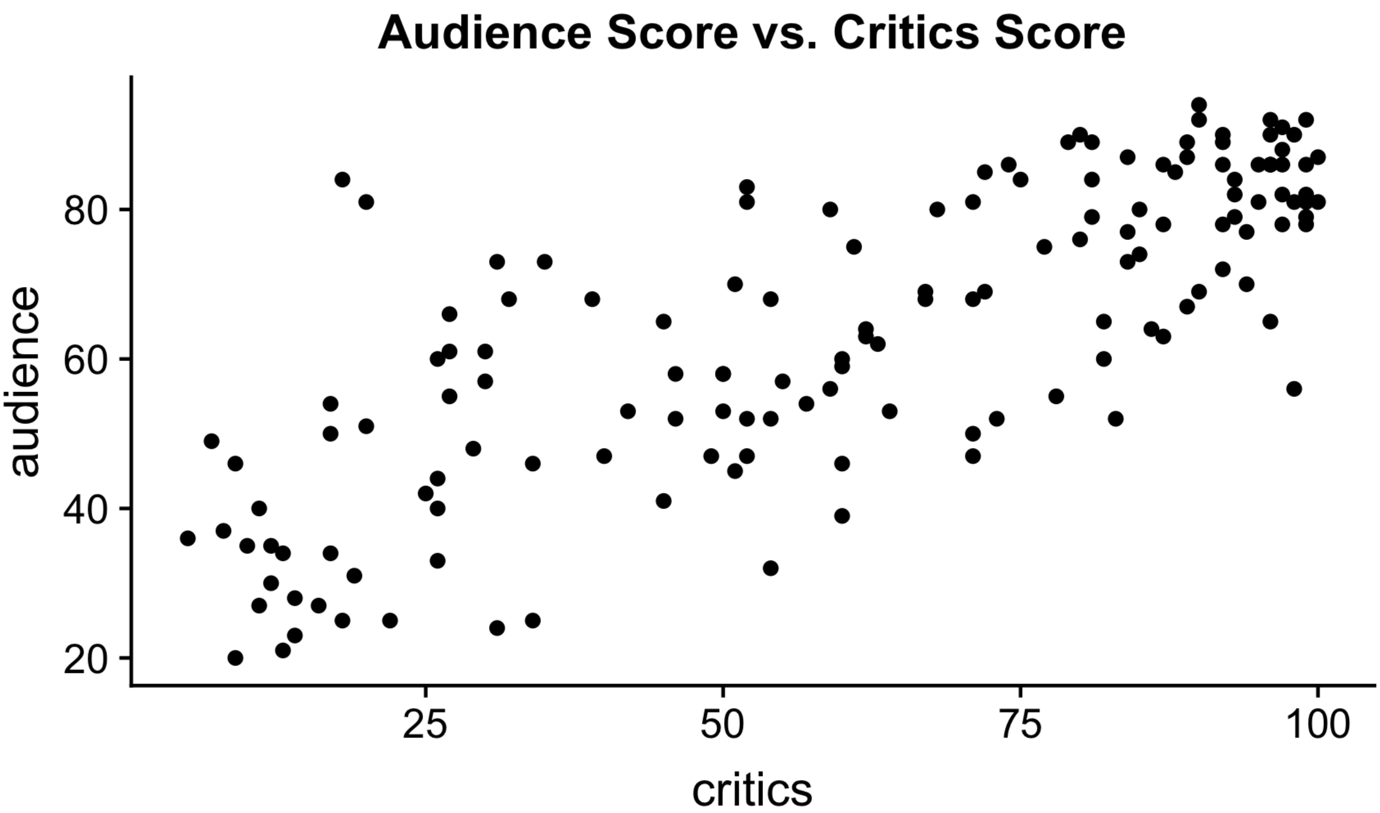
```
## Observations: 146
## Variables: 23
## $ film                         <chr> "Avengers: Age of Ultron", "Cinderell…
## $ year                         <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2…
## $ critics                      <int> 74, 85, 80, 18, 14, 63, 42, 86, 99, 8…
## $ audience                     <int> 86, 80, 90, 84, 28, 62, 53, 64, 82, 8…
## $ metacritic                   <int> 66, 67, 64, 22, 29, 50, 53, 81, 81, 8…
## $ metacritic_user              <dbl> 7.1, 7.5, 8.1, 4.7, 3.4, 6.8, 7.6, 6.…
## $ imdb                         <dbl> 7.8, 7.1, 7.8, 5.4, 5.1, 7.2, 6.9, 6.…
## $ fandango_stars               <dbl> 5.0, 5.0, 5.0, 5.0, 3.5, 4.5, 4.0, 4.…
## $ fandango_ratingvalue         <dbl> 4.5, 4.5, 4.5, 4.5, 3.0, 4.0, 3.5, 3.…
## $ rt_norm                      <dbl> 3.70, 4.25, 4.00, 0.90, 0.70, 3.15, 2…
## $ rt_user_norm                 <dbl> 4.30, 4.00, 4.50, 4.20, 1.40, 3.10, 2…
## $ metacritic_norm              <dbl> 3.30, 3.35, 3.20, 1.10, 1.45, 2.50, 2…
## $ metacritic_user_nom          <dbl> 3.55, 3.75, 4.05, 2.35, 1.70, 3.40, 3…
## $ imdb_norm                    <dbl> 3.90, 3.55, 3.90, 2.70, 2.55, 3.60, 3…
## $ rt_norm_round                <dbl> 3.5, 4.5, 4.0, 1.0, 0.5, 3.0, 2.0, 4.…
## $ rt_user_norm_round           <dbl> 4.5, 4.0, 4.5, 4.0, 1.5, 3.0, 2.5, 3.…
## $ metacritic_norm_round        <dbl> 3.5, 3.5, 3.0, 1.0, 1.5, 2.5, 2.5, 4.…
## $ metacritic_user_norm_round   <dbl> 3.5, 4.0, 4.0, 2.5, 1.5, 3.5, 4.0, 3.…
## $ imdb_norm_round              <dbl> 4.0, 3.5, 4.0, 2.5, 2.5, 3.5, 3.5, 3.…
## $ metacritic_user_vote_count   <int> 1330, 249, 627, 31, 88, 34, 17, 124, …
## $ imdb_user_vote_count         <int> 271107, 65709, 103660, 3136, 19560, 3…
## $ fandango_votes               <int> 14846, 12640, 12055, 1793, 1021, 397,…
## $ fandango_difference          <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.…
```

```r
p1 <- ggplot(data = movie_scores,mapping = aes(x = critics)) +
  geom_histogram()
p2 <- ggplot(data = movie_scores,mapping = aes(x = audience)) +
  geom_histogram()
plot_grid(p1, p2, ncol = 2)
```

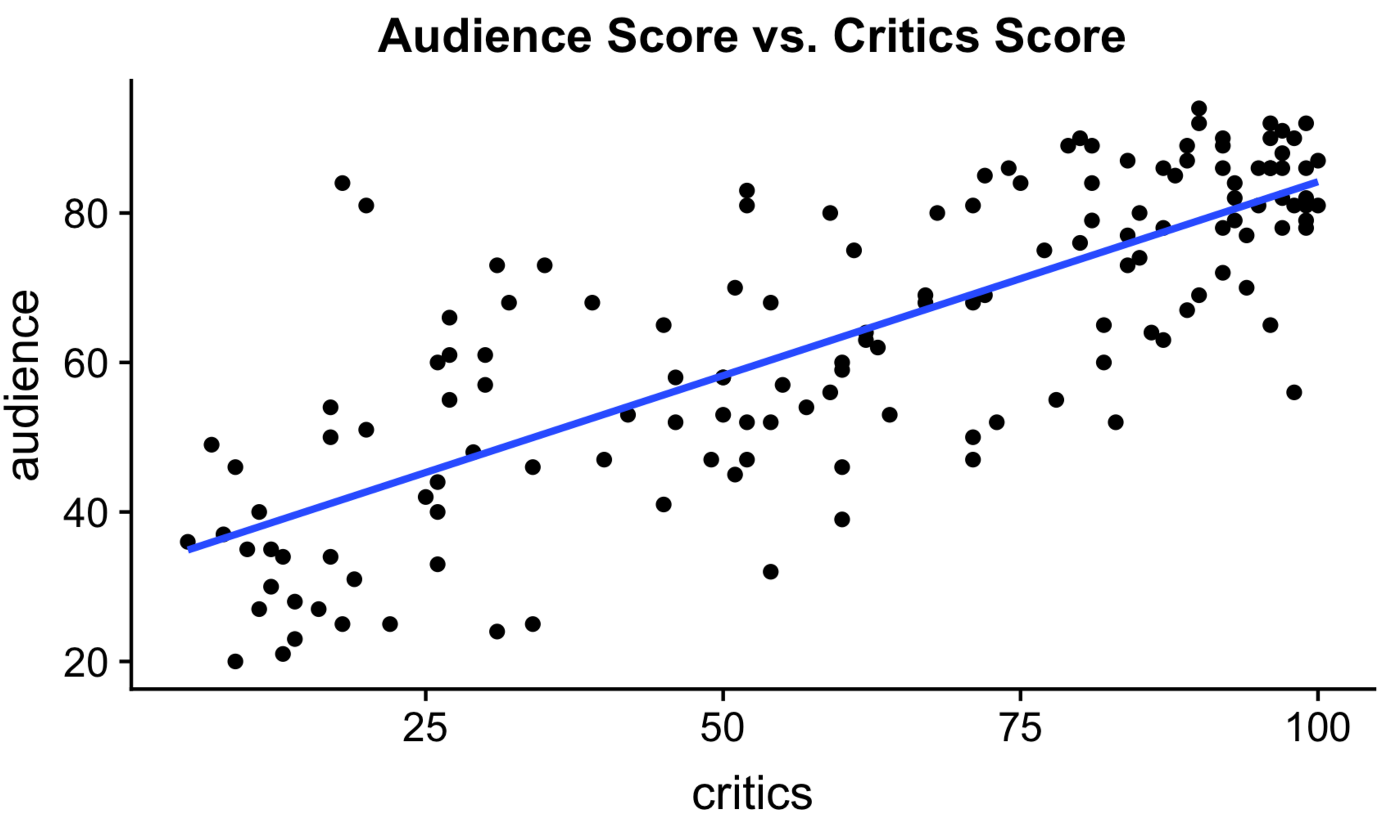


STA 210

```r
ggplot(data = movie_scores, mapping = aes(x = critics, y = audience)) +
  geom_point() +
  labs(title = "Audience Score vs. Critics Score")
```



```r
ggplot(data = movie_scores, mapping = aes(x = critics, y = audience)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Audience Score vs. Critics Score")
```



Terminology

- **`audience`** is the response variable (Y)
  - variable whose variation we want to understand and/or variable we wish to predict
  - also known as *dependent*, *outcome*, *target* variable

- **`critics`** is the predictor variable (X)
  - variable used to account for variation in the response
  - also known as *independent*

STA 210

Model

$$\text{audience} = f(critics) + \epsilon$$

We want to estimate f. How do we do it?

General form of model

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

- Y: quantitative response variable
- $\mathbf{X} = (X_1, X_2, \ldots, X_p)$: predictor variables
- f: fixed but unknown function
  - systematic information $\mathbf{X}$ provides about Y
- ϵ: random error term with mean 0 that is independent of $\mathbf{X}$

STA 210

How to estimate f?

In general, we will use the following steps to estimate f

- Choose the functional form of f, i.e. choose the appropriate model given the data
  - Ex: f is a linear model

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

- Use the data to fit the model, i.e. estimate the model parameters
  - Ex: Use a method to estimate the model parameters $\beta_0, \beta_1, \ldots, \beta_p$

Why estimate f?

Suppose we have the model

$$\text{audience} = \beta_0 + \beta_1 \times \text{critics} + \epsilon$$

- What is one question you can answer using this model?
- Submit your response at: http://bit.ly/sta210-sp20-q
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

03:00

Why estimate f?

There are two types of questions we may wish to answer using our model:

- **Prediction:** What is the expected Y given particular values of $X_1, X_2, \ldots, X_p$?
  - Ex: What is the expected audience score for a movie that receives a critic score of 70%?
- **Inference:** What is the relationship between $\mathbf{X}$ and Y. How does Y change as a function of $\mathbf{X}$?
  - Ex: How much can we expect the audience score to change for each additional point in the critic score?

Course Outline

- **Unit 1: Quantitative Response Variables**
  - Simple Linear Regression
  - Multiple Linear Regression

- **Unit 2: Categorical Response Variable**
  - Logistic Regression
  - Multinomial Logistic Regression

- **Unit 3: Looking Ahead**
  - Log-linear models
  - Weighted least squares
  - Missing data
  - Special topics

STA 210

Simple Linear Regression

Least-Squares Regression

- There is some true relationship between X and Y that exists in the population

$$Y = f(X) + \epsilon$$

- If f is approximated by a linear function, then we can write the relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

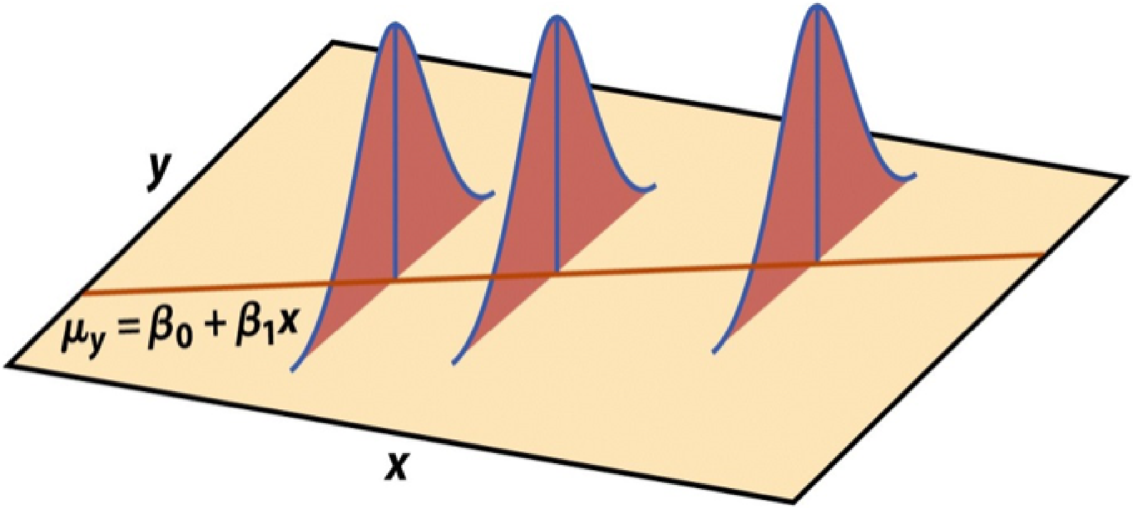
- We'll estimate the slope and intercept of this linear function using least-squares regression

- We'll use statistical inference to determine if the relationship we observe in the data is statistically significant or if it's due to random chance (we'll talk about this more next class)

STA 210

Regression Model

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$



- σ: the standard deviation of Y as a function of X
- *Assumption*: σ is equal for all values of X

Regression Model

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

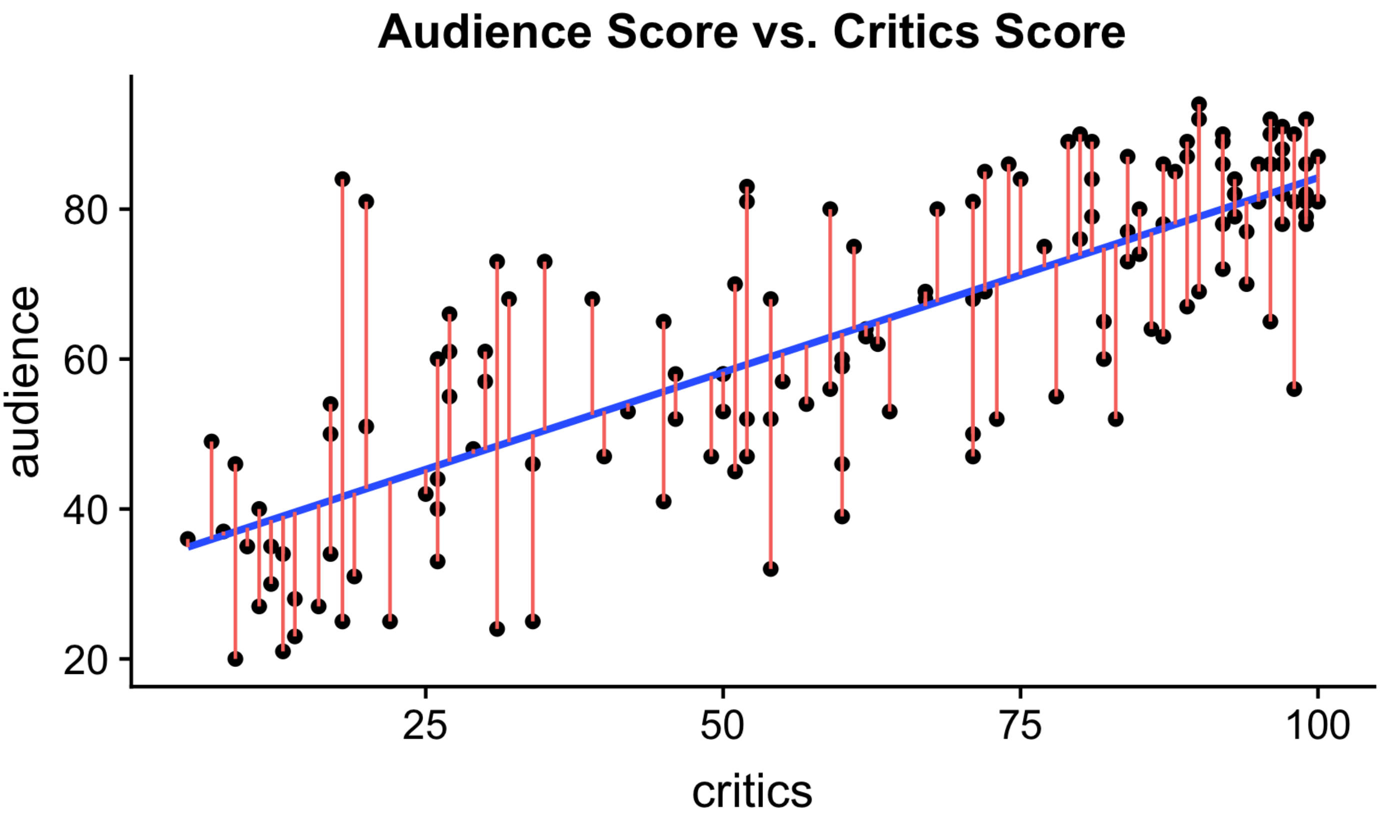
- For a single observation (x_i, y_i)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- We want to use the n observations $(x_1, y_1), \ldots, (x_n, y_n)$ to estimate β_0 and β_1. We will use *least-squares regression* estimates.

STA 210

Residuals



The residual is the difference between the observed and predicted response.

Residual Sum of Squares

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

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- The *residual sum of squares* is

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

- The least-squares regression approach chooses coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize RSS.

Estimating Coefficients

- **Slope:**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{s_y}{s_x}$$

such that r is the correlation between x And y.

- **Intercept:** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

Least-Squares Model

```
model <- lm(audience ~ critics, data = movie_scores)
tidy(model) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	32.316	2.343	13.795	0
critics	0.519	0.035	15.028	0

$$\widehat{\text{audience}} = 32.316 + 0.519 \times \text{critics}$$

STA 210

Interpreting Slope & Intercept

- **Slope:** Increase in the mean response for every one unit increase in the predictor variable
- **Intercept:** Mean response when the explanatory variable equals 0

- The regression equation for the Rotten Tomatoes data is

$$\widehat{\text{audience}} = 32.316 + 0.519 \times \text{critics}$$

Write the interpretation of the slope and intercept

STA 210

Nonsensical Intercept

- Sometimes it doesn't make sense to interpret the intercept
  - When predictor variable doesn't take values close to 0
  - When the intercept is negative even though the response variable should always be positive
- The intercept helps the line fit the data as closely as possible
- It is fine to have a nonsensical intercept if it helps the model give better overall predictions

Does it make sense to interpret the intercept?

- Example 1:
    - **Explanatory**: number of home runs in a baseball game
    - **Response**: attendance at the next baseball game

- Example 2:
    - **Explanatory**: height of a person
    - **Response**: weight of a person

Assessing Model Fit

R^2

We can use the coefficient of determination, R^2, to measure how well the model fits the data

- R^2 is the proportion of variation in Y that is explained by the regression line (reported as percentage)
- It is difficult to determine what's a "good" value of R^2. It depends on the context of the data.

Calculating R^2

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- **Total Sum of Squares:** Total variation in the Y's before fitting the regression

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2$$

- **Residual Sum of Squares (RSS):** Total variation in the Y's around the regression line (sum of squared residuals)

$$\text{RSS} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Rotten Tomatoes Data

```
rsquare(model,movie_scores)
```

```
## [1] 0.6106479
```

The critics score explains about 61.06% of the variation in audience scores on rottentomatoes.com.

STA 210

Checking Model Assumptions

Assumptions for Regression

1. **Linearity:** The plot of the mean value for y against x falls on a straight line

2. **Constant Variance:** The regression variance is the same for all values of x

3. **Normality:** For a given x, the distribution of y around its mean is Normal

4. **Independence:** All observations are independent

Checking Assumptions

We can use plots of the residuals to check the assumptions for regression.

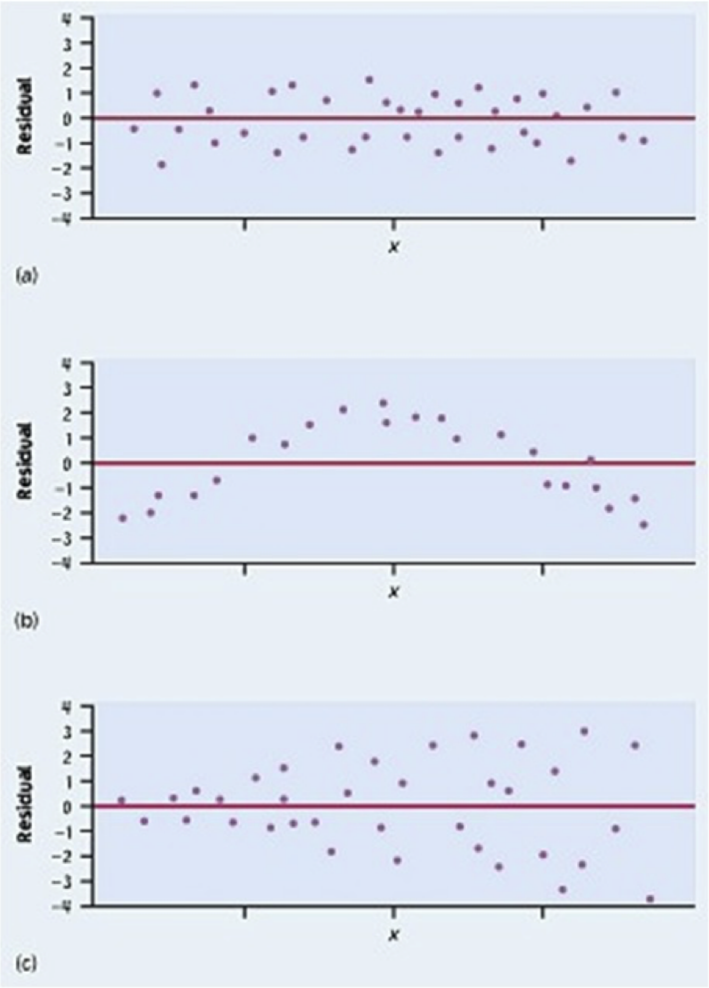
1. Scatterplot of Y vs. X (linearity).
    - Check this before fitting the regression model.
2. Plot of residuals vs. predictor variable (constant variance, linearity)
3. Histogram and Normal QQ-Plot of residuals (Normality)

Residuals vs. Predictor

- When all the assumptions are true, the values of the residuals reflect random (chance) error
- We can look at a plot of the residuals vs. the predictor variable
- There should be no distinguishable pattern in the residuals plot, i.e. the residuals should be randomly scattered
- A non-random pattern suggests assumptions might be violated

STA 210

Plots of Residuals



Ideal Residual Plot

Nonlinearity

Nonconstant Variance

STA 210

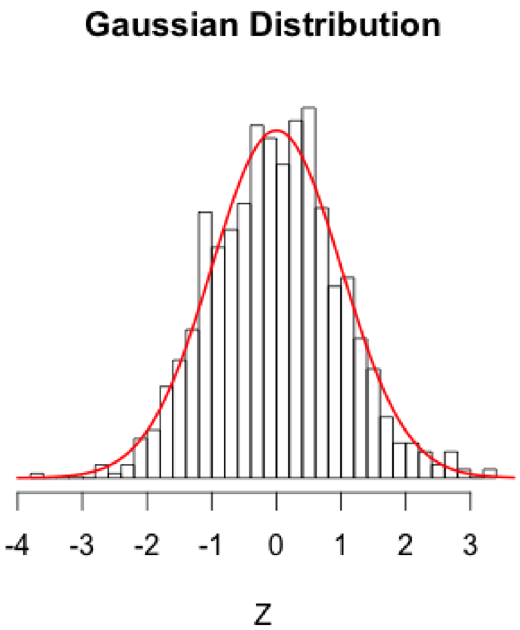
```r
movie_scores <- movie_scores %>% mutate(residuals=resid(mod
```

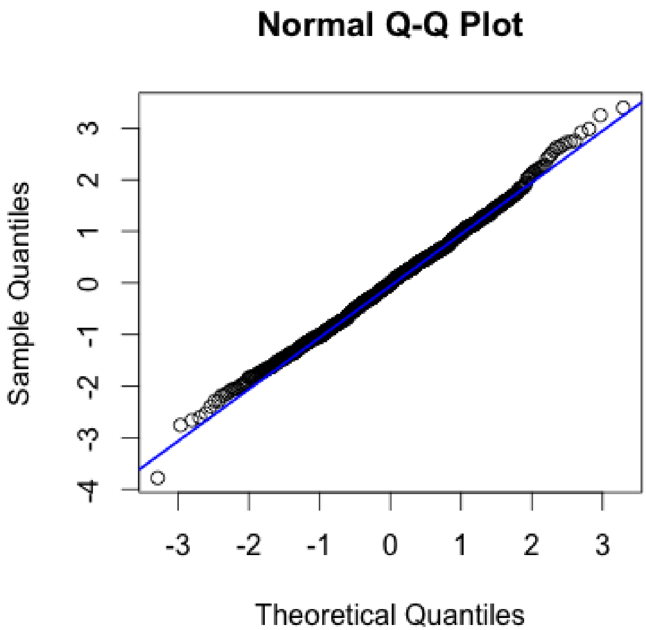
```r
ggplot(data=movie_scores,mapping=aes(x=critics, y=residuals
  geom_point() +
  geom_hline(yintercept=0,color="red")+
  labs(title="Residuals vs. Critics Score")
```

Checking Normality

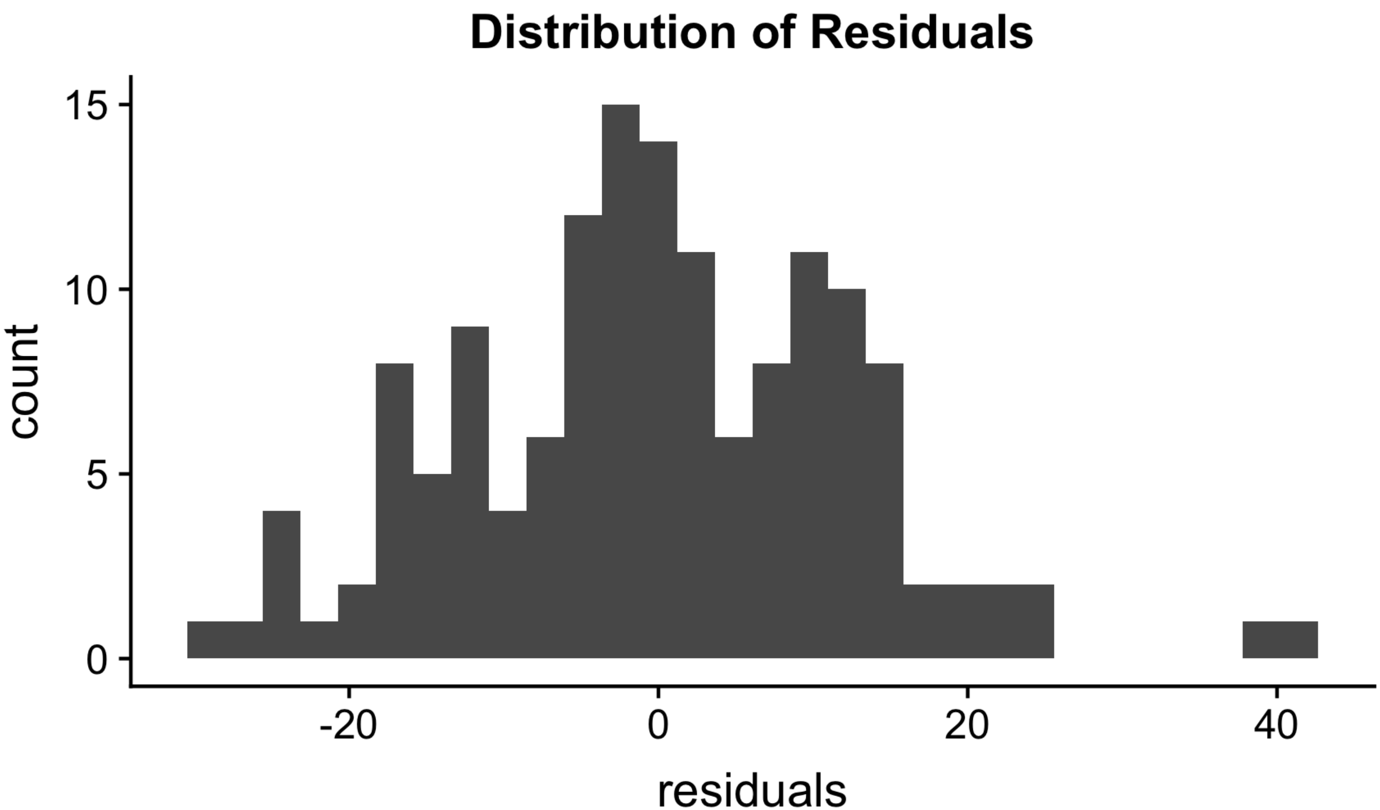
- Examine the distribution of the residuals to determine if the Normality assumption is satisfied
- Plot the residuals in a histogram and a Normal QQ plot to visualize their distribution and assess Normality
- Most inference methods for regression are robust to some departures from Normality

Normal QQ-Plot

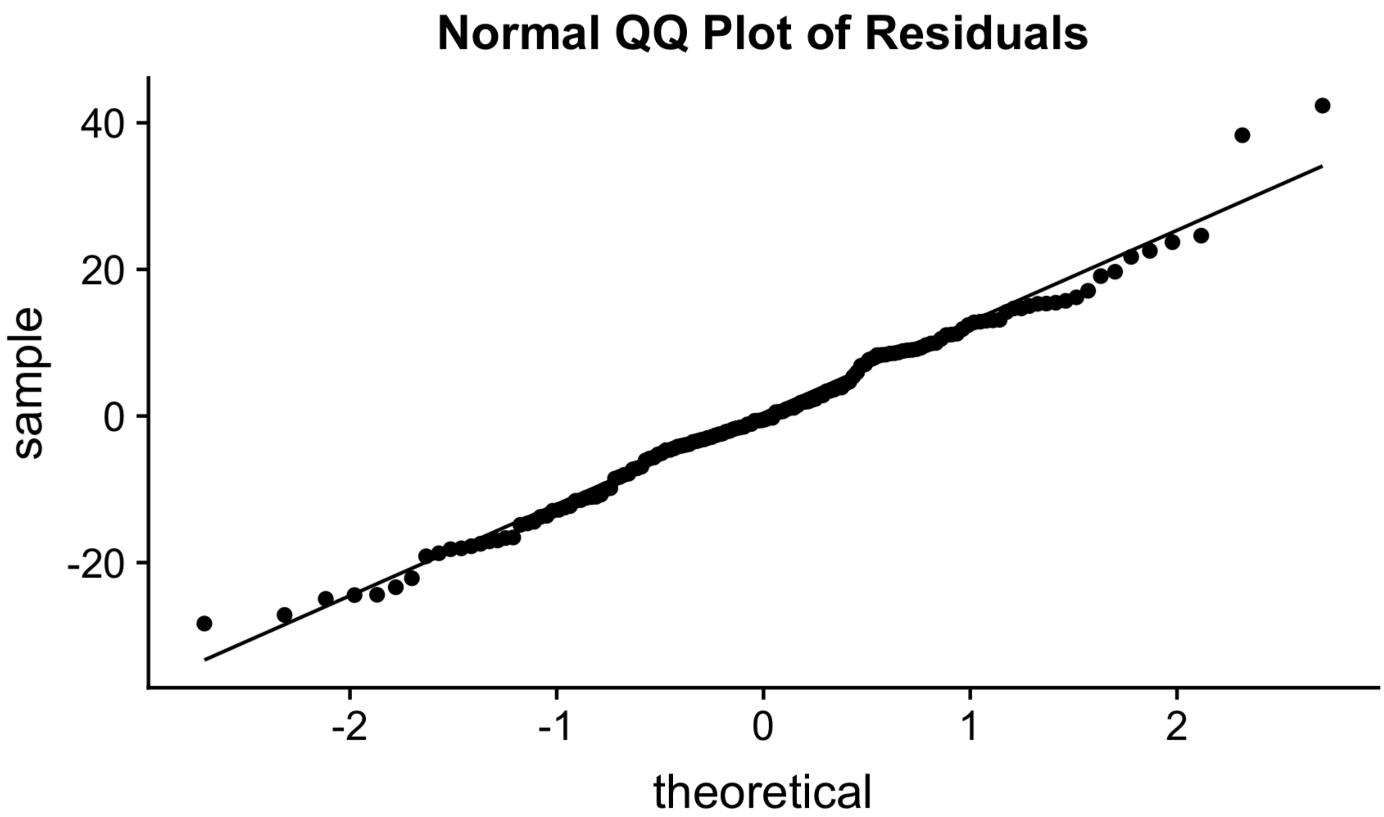




```
ggplot(data=movie_scores,mapping=aes(x=residuals)) +
  geom_histogram() +
  labs(title="Distribution of Residuals")
```



```r
ggplot(data=movie_scores,mapping=aes(sample=residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal QQ Plot of Residuals")
```



Checking Independence

- Often, we can conclude that the independence assumption is sufficiently met based on a description of the data and how it was collected.

- Two common violations of the independence assumption:

  - **Serial Effect**: If the data were collected over time, the residuals should be plotted in time order to determine if there is serial correlation

  - **Cluster Effect**: You can plot the residuals vs. a group identifier or use different markers (colors/shapes) in the residual plot to determine if there is a cluster effect.

STA 210

Recap

- Motivating Regression Analysis
- Simple Linear Regression
  - Estimating & interpreting coefficients
  - Assessing model fit: R^2
  - Residuals and model assumptions