Simple Linear Regression

Inference & Prediction

Dr. Maria Tackett

09.09.19

STA 210

[Click for PDF of slides]

Announcements

- Lab 02 due **tomorrow at 11:59p**

Check in

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

Today's Agenda

- Assessing model fit
- Model assumptions
- Inference for regression
- Prediction

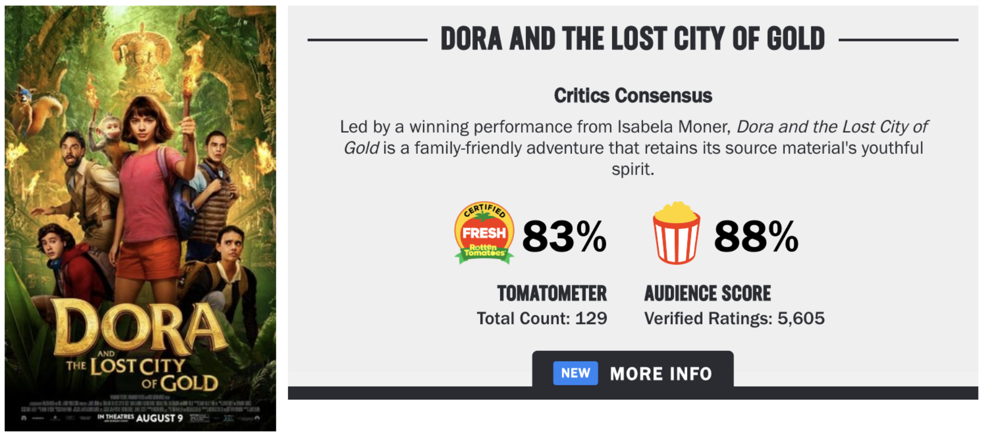
Packages and Data

```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)
```

rottentomatoes.com

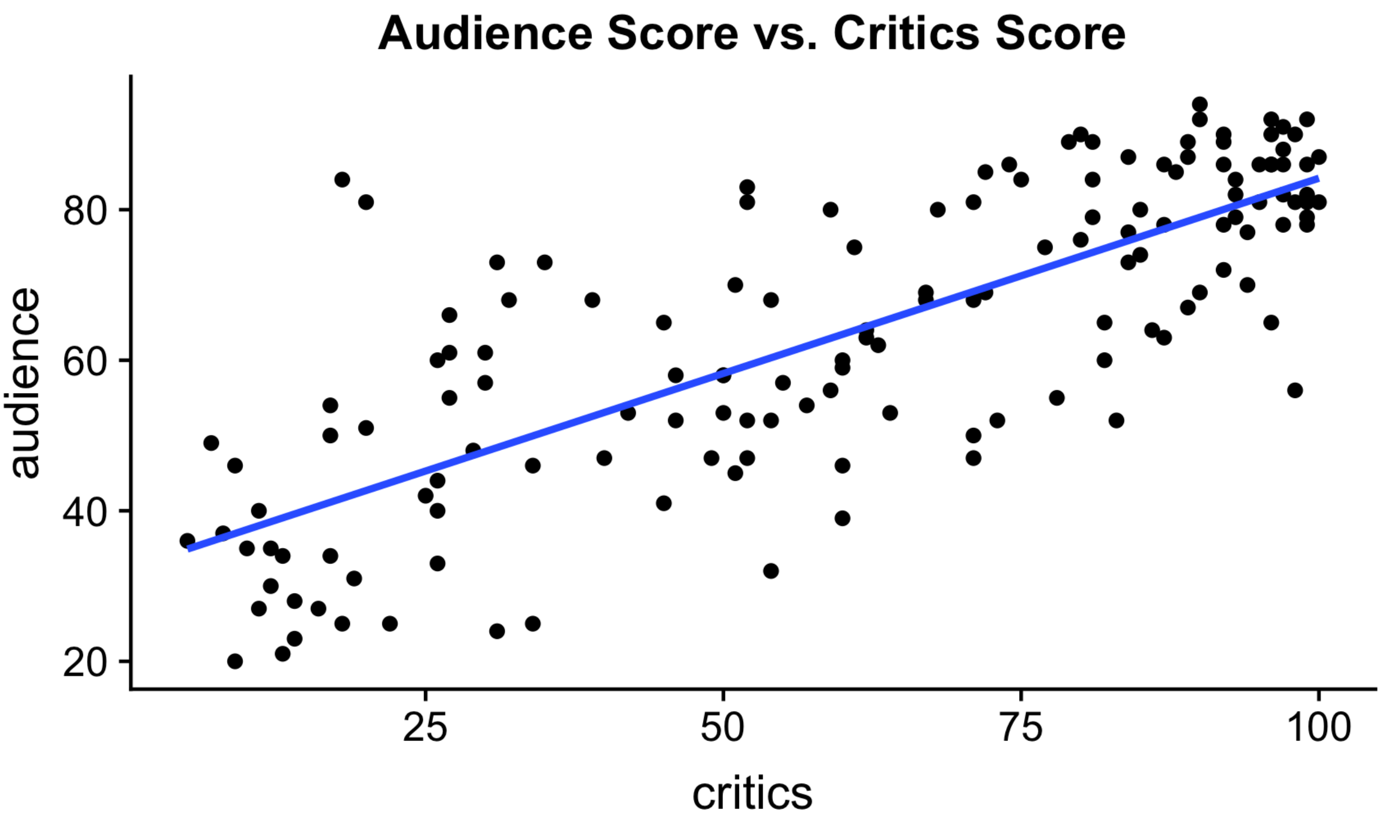
Can the ratings from movie critics be used to predict what movies the audience will like?



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`**: critics score for the film (0 - 100)
  - **`audience`**: Audience score for the film (0 - 100)

```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")
```



The 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

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

- **Slope:** For each additional percentage point in the critics score, the audience score is expected to increase by 0.519 percentage points on average.

- **Intercept:** If a movie gets a 0% from the critics, the audience score is expected to be 32.316%.

STA 210

Assessing Model Fit

R^2

- We can use the coefficient of determination, R^2, as one way to measure how well the model fits the data
  - specifically how well it explains variation in Y
- R^2 is the proportion of variation in Y that is explained by the regression line
  - R^2 values range from 0 to 1
  - Typically report R^2 as a percentage
- Ideally, we'll have R^2 close to 1; however, it is difficult to determine what exactly is a "good" value of R^2.
  - It depends on the context of the data.

STA 210

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 line

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

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

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

STA 210

Rotten Tomatoes Data

```
glance(model,movie_scores)$r.squared
```

```
## [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

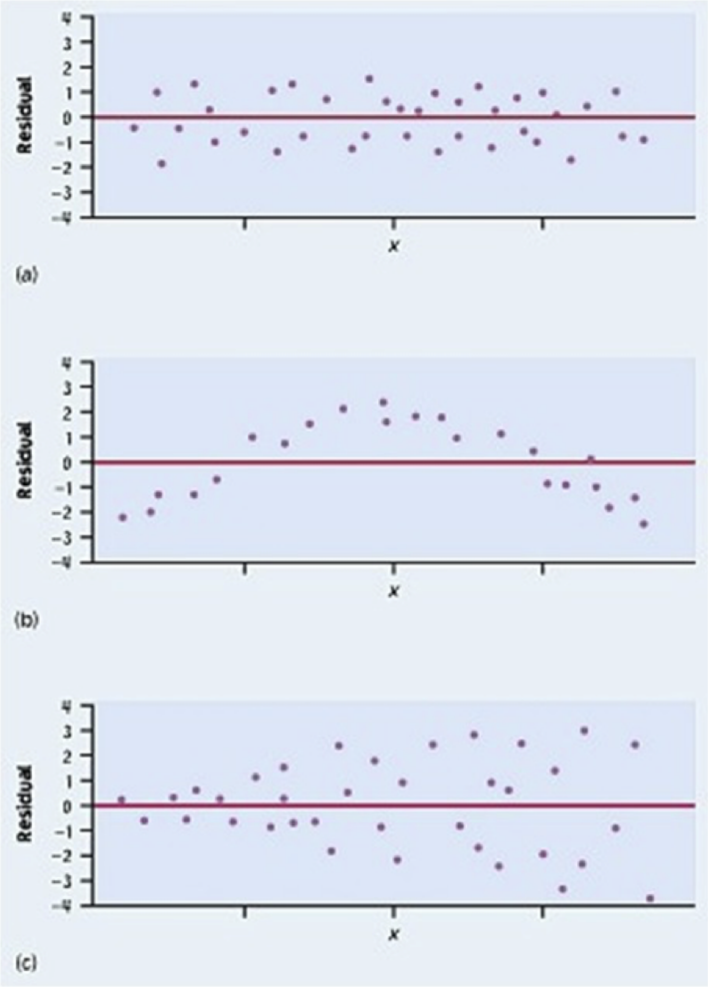
We can use plots of the residuals to check the assumptions for regression.

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

Plots of Residuals



Ideal Residual Plot

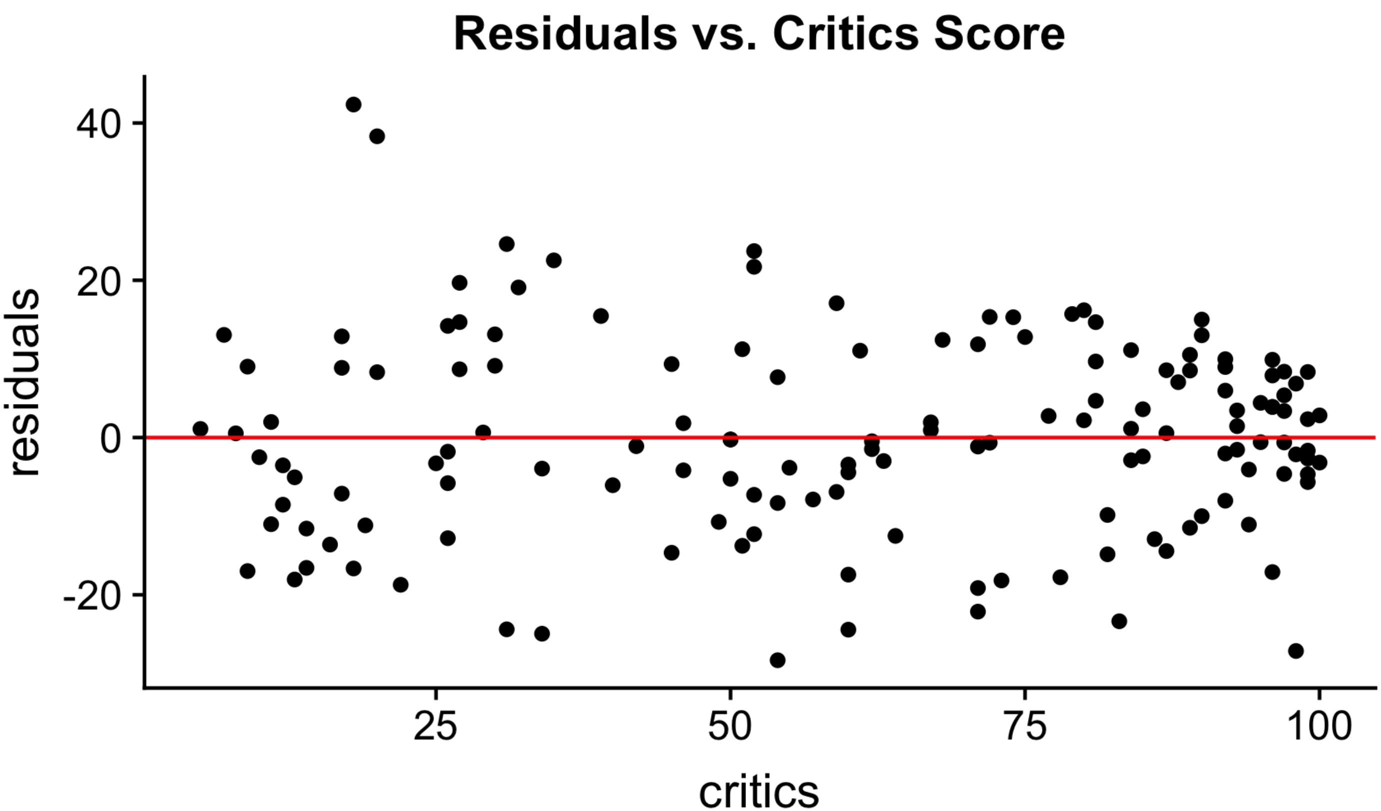
Nonlinearity

Nonconstant Variance

STA 210

```
movie_scores <- movie_scores %>%
  mutate(residuals = resid(model))
```

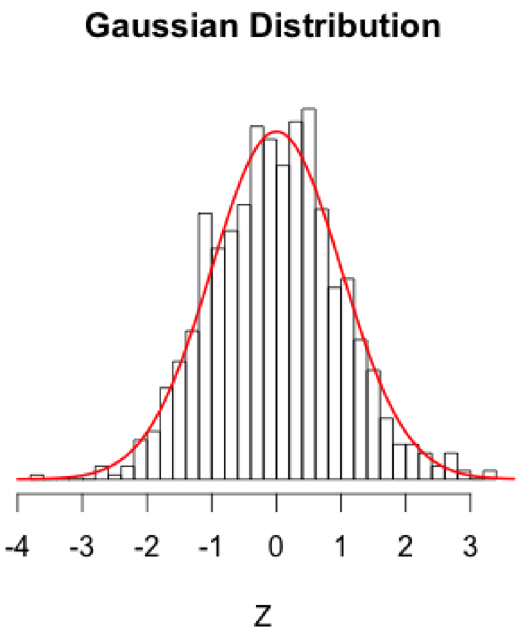
```
ggplot(data = movie_scores, mapping = aes(x = critics, y = residua
  geom_point() +
  geom_hline(yintercept = 0, color = "red")+
  labs(title = "Residuals vs. Critics Score")
```

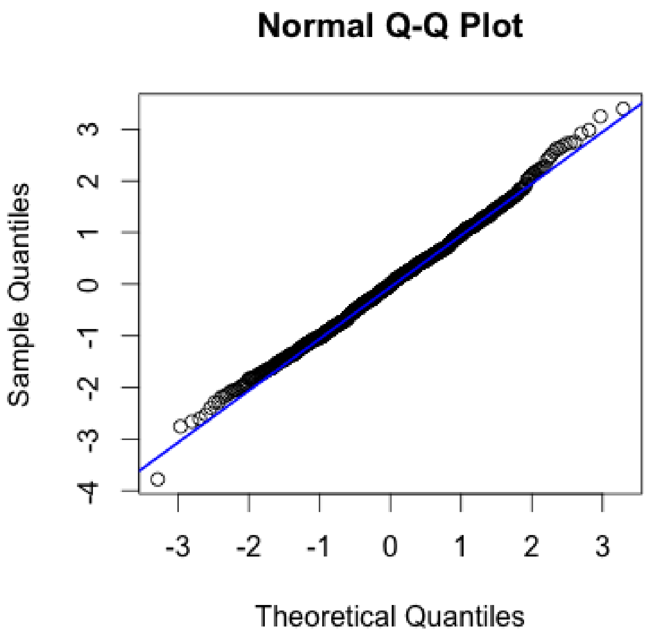


Checking Normality

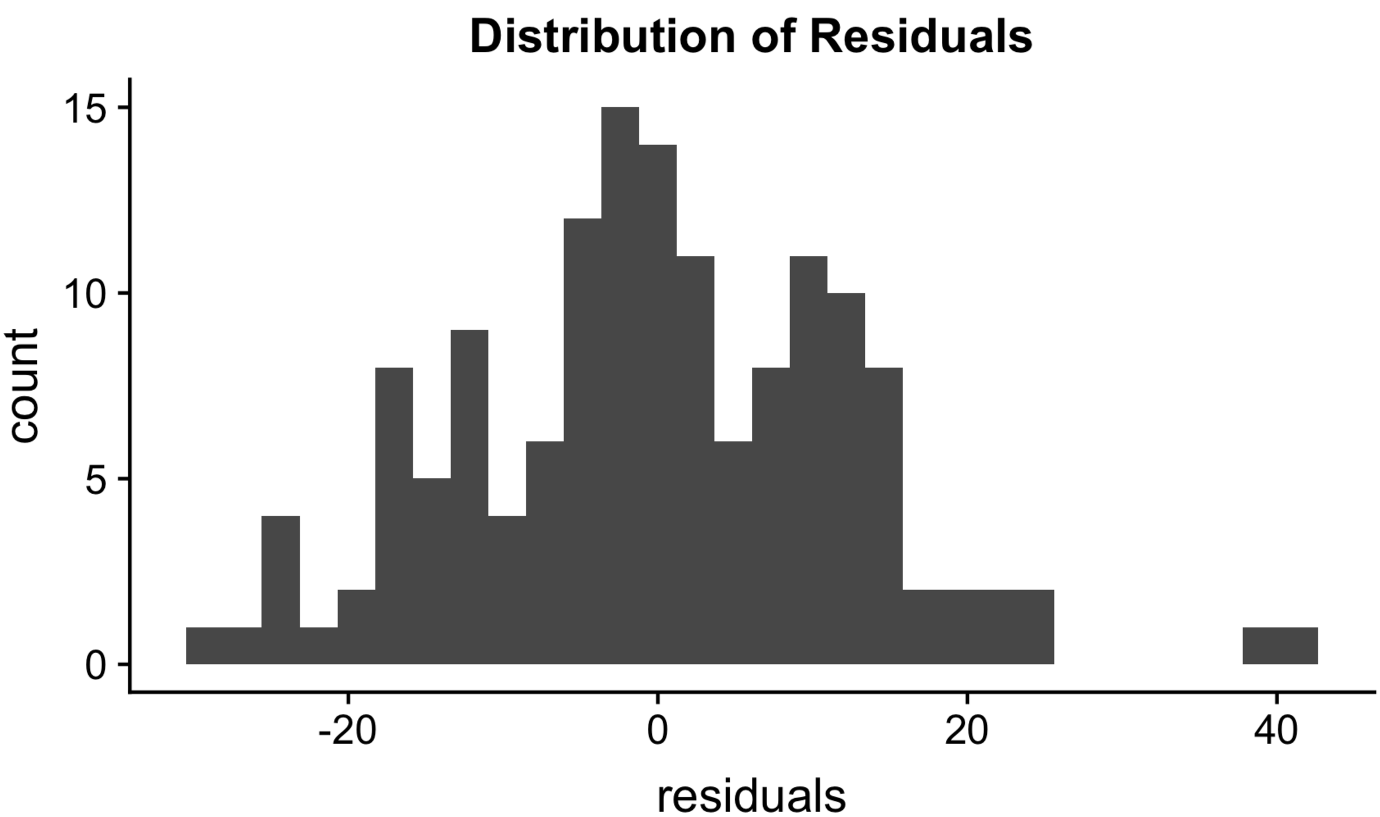
- Examine the distribution of the residuals to determine if the Normality assumption is satisfied
- Plot the residuals in a histogram and a Normal QQ plot to visualize their distribution and assess Normality
- Most inference methods for regression are robust to some departures from Normality

Normal QQ-Plot

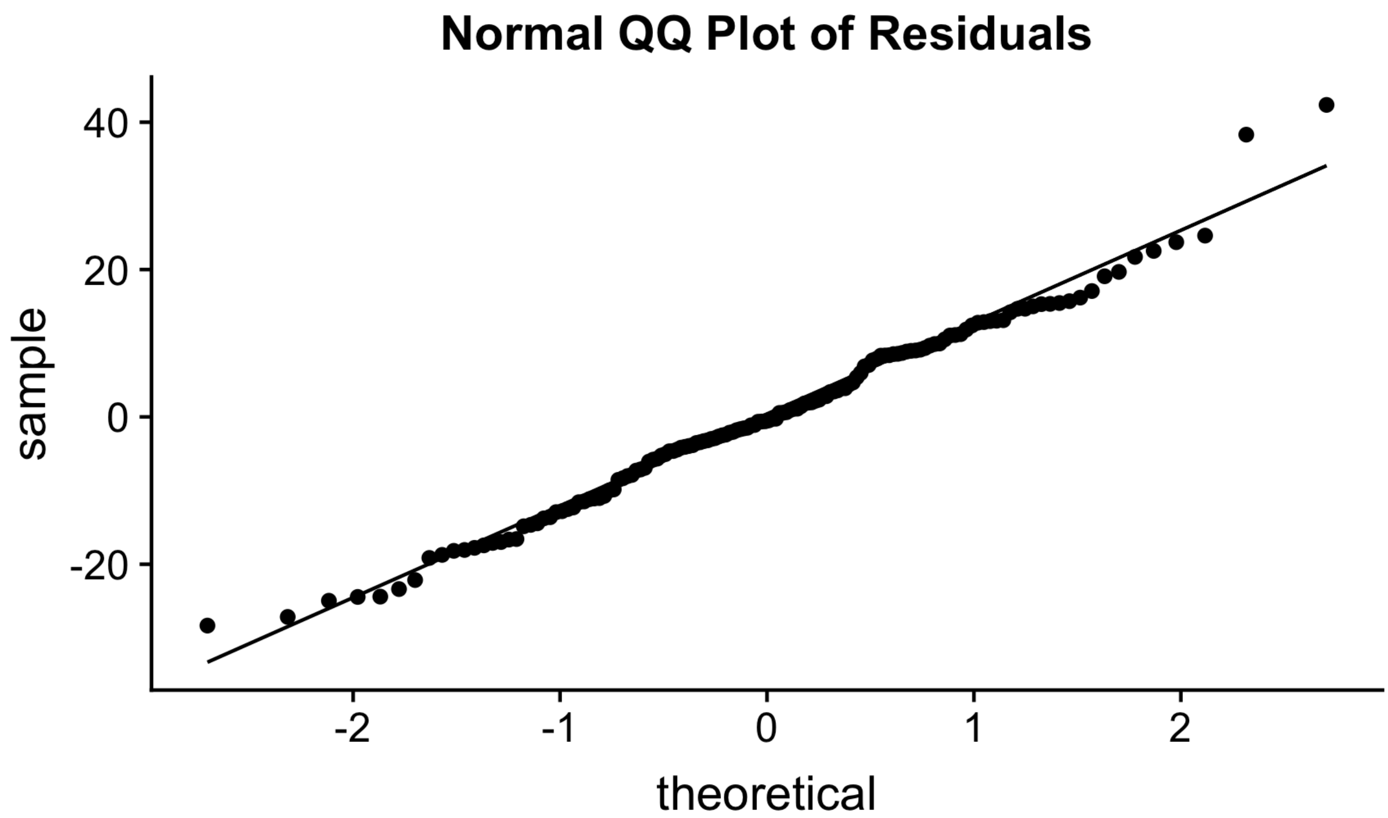




```r
ggplot(data = movie_scores, mapping = aes(x = residuals)) +
  geom_histogram() +
  labs(title = "Distribution of Residuals")
```



```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

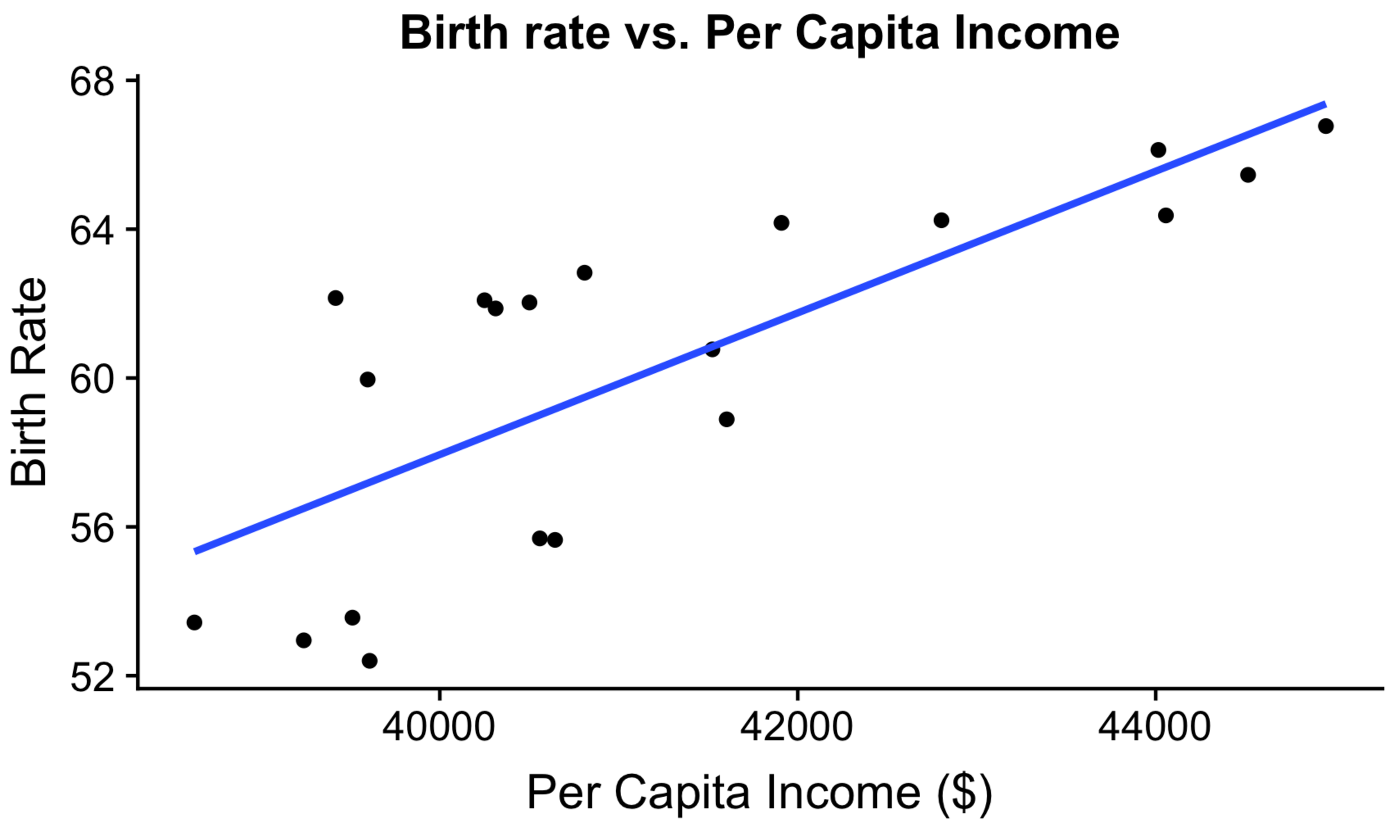
  - **Cluster Effect**: You can plot the residuals vs. a group identifier or use different markers (colors/shapes) in the residual plot to determine if there is a cluster effect.

STA 210

Example: Birth rate vs. Per Capita Income

- A [2011 study by Pew Research] looked at the economy's effect on birthrate in the United States.
- We will look at data for Virginia and Washington D.C. years 2000 - 2009
- Birth rate: Births per 100,000 women ages 15-44
- Per Capita Income: average income per person

```r
ggplot(data = pew_data, mapping = aes(x = percapitaincome, y = bi
  geom_point() +
  geom_smooth(method = "lm",se=FALSE) +
  labs(title = "Birth rate vs. Per Capita Income",
       x = "Per Capita Income ($)", y = "Birth Rate")
```



Birthrate vs. Per Capita Income

```
pew_model <- lm(birthrate ~ percapitaincome, data = pew_data)
tidy(pew_model) %>%
  kable(format = "markdown", digits = 3)
```

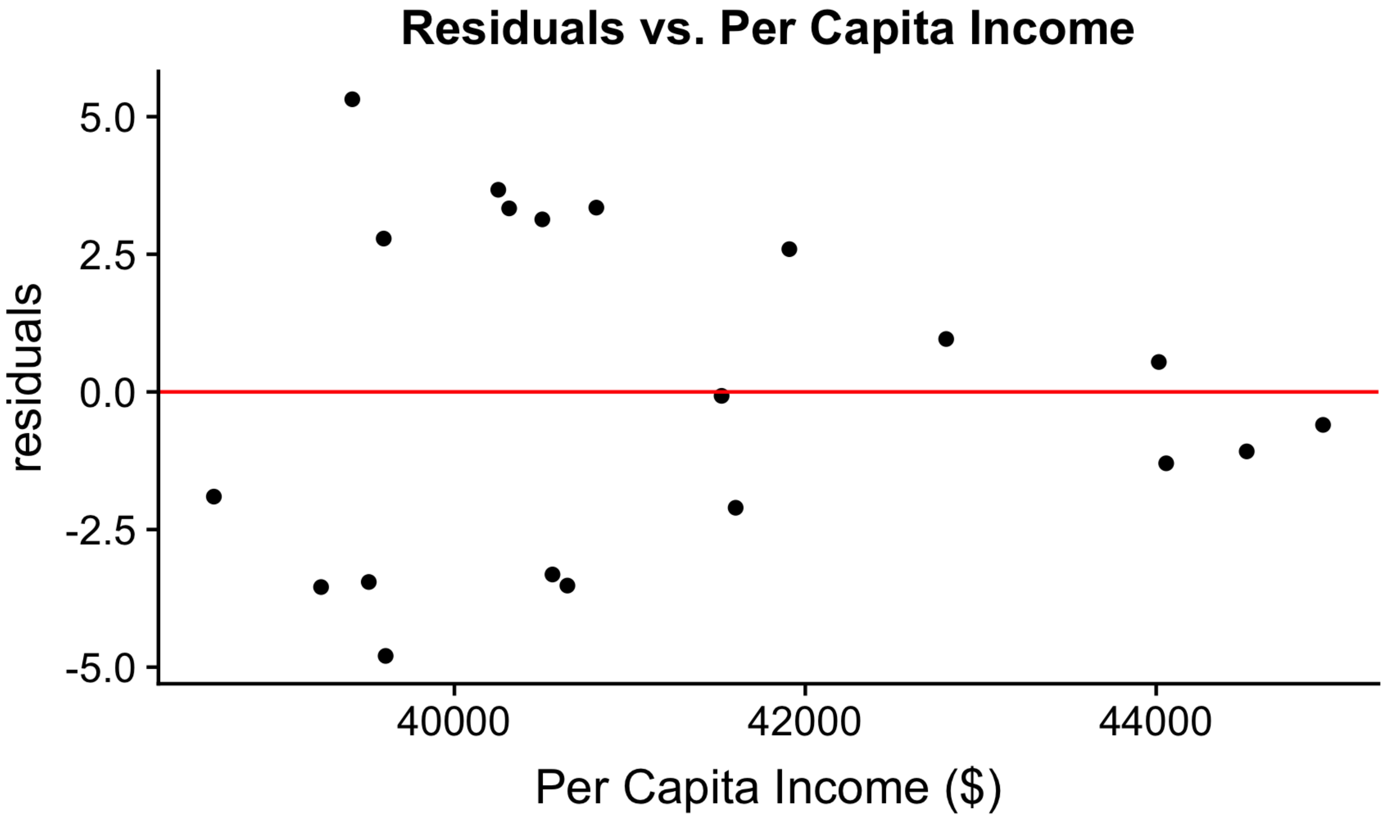
term	estimate	std.error	statistic	p.value
(Intercept)	-18.218	15.33	-1.188	0.25
percapitaincome	0.002	0.00	5.125	0.00

$$\widehat{\text{Birth Rate}} = -18.2 + 0.002 \times \text{ Per Capita Income}$$

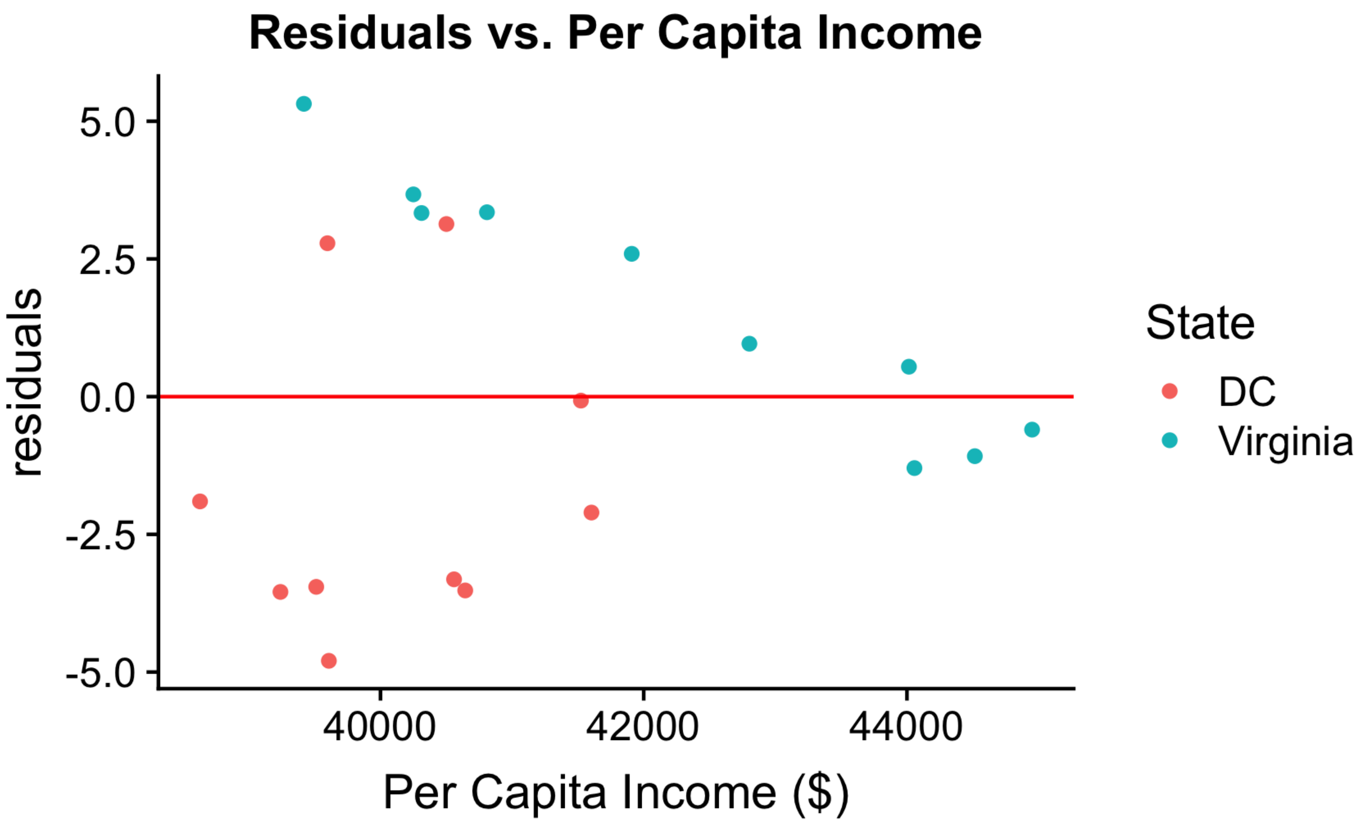
STA 210

Residuals vs. Explanatory Variable

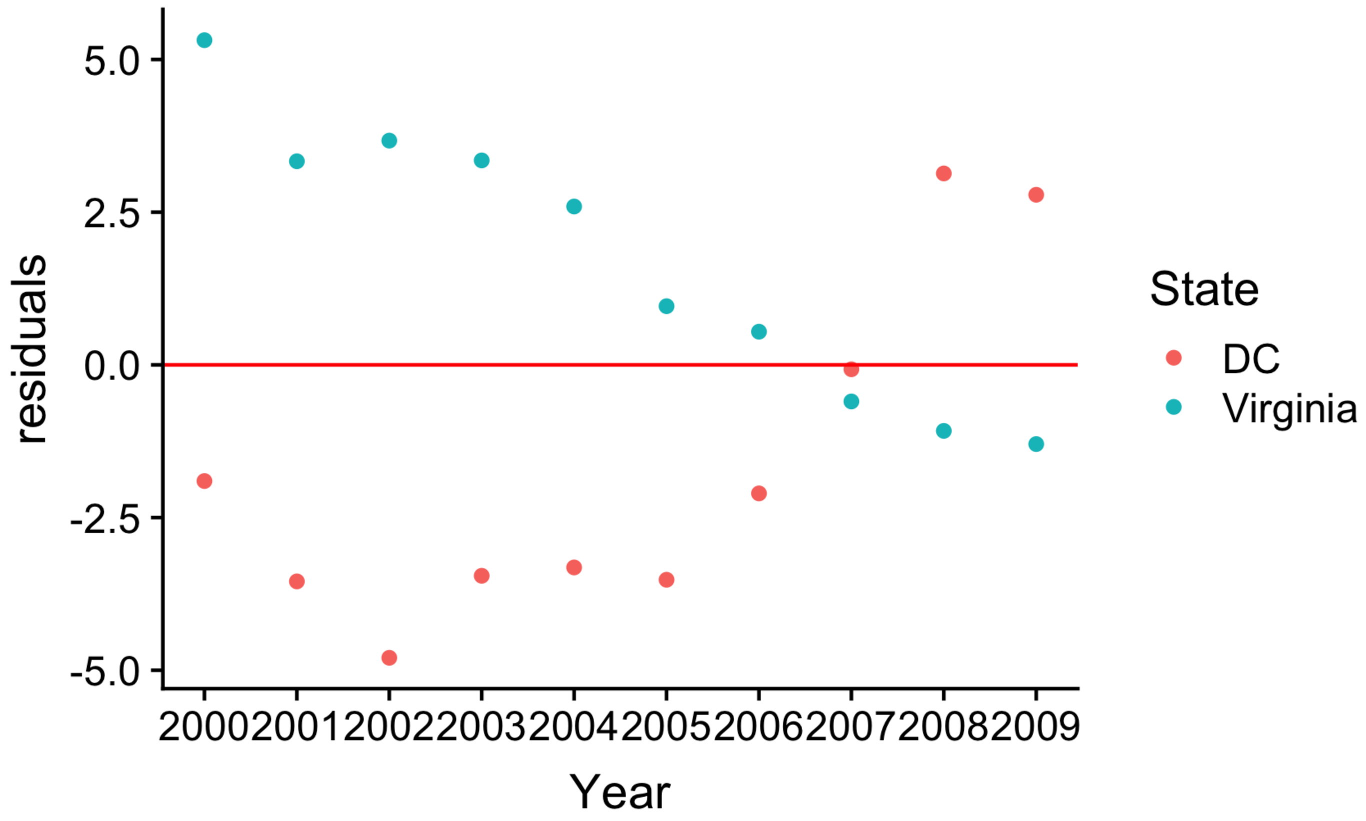
```
pew_data <- pew_data %>%
  mutate(residuals = resid(pew_model))
```



Residuals: Cluster Effect



Residuals: Serial Effect



STA 210

Inference for β_1

Questions of interest

In our example, we will treat the data as a random sample of movies from rottentomatoes.com

Questions of interest

- What is a plausible range of values of the true population slope for `critics`? (confidence interval)

- Is there truly a linear relationship between the critic and audience scores?

  - We estimated $\hat{\beta}_1 = 0.519$, but is there sufficient evidence to conclude that the true population slope β is different from 0? (hypothesis test)

STA 210

What is a plausible range of values of the true population slope for `critics`?

General form of the CI

- Let **SE** be the standard error of the statistic used to estimate the parameter of interest, then the general form of the confidence interval is

$$\text{Estimate} \pm \text{ (critical value) } \times \text{SE}$$

- *Note*: The critical value is determined by the distribution of the estimate (statistic) and the confidence level
- For the regression slope:
  - $\hat{\beta}_1$ is the statistic used to estimate the parameter, β_1
  - We will write the confidence interval as

$$\hat{\beta}_\mathbf{1} \pm \mathbf{t^* SE(\hat{\beta}_1)}$$

Confidence interval for β_1

- The confidence interval for the regression slope is

$$\hat{\beta}_1 \pm \mathbf{t^*SE(\hat{\beta}_1)}$$

- t^* is the critical value associated with the confidence level.
  - It is calculated from a t distribution with $n-2$ degrees of freedom
- $SE(\hat{\beta}_1)$ is the standard error for the slope

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}} = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}}$$

STA 210

What is $\hat{\sigma}$?

- Recall, the residual is the difference between the observed response the predicted response (the estimated mean)
  - The residual for the ith observation, (x_i, y_i), is

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

- The Residual Standard Error is the estimate of variation about the regression line
  - Also known as the **Root Mean Square Error (RMSE)**

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$$

Why *t*?

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma\sqrt{\frac{1}{(n-1)s_X^2}}\right)$$

- We don't know σ, so we use the estimate $\hat{\sigma}$ in our calculations. Therefore, we use the *t* distribution when we calculate the confidence interval (and conduct hypothesis tests) to account for the extra variability that's been introduced
- The critical value t^* is calculated from the *t(n-2)* distribution - the *t* distribution with *n-2* degrees of freedom.

STA 210

Movies data: Critical value

```
qt(0.975, 144)
```

```
## [1] 1.976575
```

Calculating the 95% CI for β_1

n	var.x	sigma	beta1	crit.val
146	910.156	12.538	0.519	1.977

Write the equation for the 95% confidence interval for β_1, the coefficient (slope) of `critics`.

Interpretation

```
model %>%
  tidy(conf.int=TRUE) %>%
  kable(format = "markdown", digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	32.316	2.343	13.795	0	27.685	36.946
critics	0.519	0.035	15.028	0	0.450	0.587

Interpret the 95% confidence interval for β_1, the coefficient (slope) of critics.

Is there truly a linear relationship between the critic and audience scores?

Recall: Outline of Hypothesis Test

1. State the hypotheses
2. Calculate the test statistic
3. Calculate the p-value
4. State the conclusion in the context of the problem

STA 210

1. State the hypotheses

- We are often interested in testing whether there is a significant linear relationship between the explanatory and response variable
- If there is no linear relationship between the two variables, the population regression slope, β_1, would equal 0
- We can test the hypotheses:

$$
\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_a &: \beta_1 \neq 0
\end{aligned}
$$

- This is the test conducted by the `lm()` function in R

STA 210

2. Calculate the test statistic

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Test Statistic:

$$\text{test statistic} = \frac{\text{Estimate} - \text{Hypothesized}}{SE}$$

$$= \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

STA 210

3. Calculate the p-value

p-value is calculated from a t distribution with $n-2$ degrees of freedom

$$\text{p-value} = P(t \geq |\text{test statistic}|)$$

Write the general interpretation of the p-value for tests of β_1.

4. State the conclusion

Magnitude of p-value	Interpretation
p-value < 0.01	strong evidence against H_0
0.01 < p-value < 0.05	moderate evidence against H_0
0.05 < p-value < 0.1	weak evidence against H_0
p-value > 0.1	effectively no evidence against H_0

Note: These are general guidelines. The strength of evidence depends on the context of the problem.

Movie data: Hypothesis test for β_1

```
model %>%
  tidy() %>%
  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

a. State the hypotheses in (1) words and (2) statistical notation.

b. What is the meaning of the test statistic in the context of the problem?

c. What is the meaning of the p-value in the context of the problem?

d. State the conclusion in context of the problem.

STA 210

Predictions

Predictions for New Observations

- We can use the regression model to predict for a response at x_0

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- Because the regression models produces the mean response for a given value of x_0, it will produce the same estimate whether we want to predict the mean response at x_0 or an individual response at x_0

Movies Data

What is the predicted audience score **for a movie** that has a critic score of 60%?

What is the predicted average audience score **for the subset of movies** that have a critic score of 60%?

Predictions for New Observations

- There is uncertainty in our predictions, so we need to calculate an a standard error (SE) to capture the uncertainty
- The SE is different depending on whether you are predicting an average value or an individual value
- SE is larger when predicting for an individual value than for an average value

Standard errors for predictions

Predicting the mean response

$$SE(\hat{\mu}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Predicting an individual response

$$SE(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Movie data: Predicting the mean response

We wish to predict the **mean** audience score for the subset of movies with a critics score of 60%.

```
x0 <- data.frame(critics = c(60))
predict.lm(model, x0, interval = "confidence", conf.level = 0.95)
```

Interpret the interval in the context of the data.

Movies data: Predicting an individual response

We wish to predict the **mean** audience score for the subset of movies with a critics score of 60%.

```
x0 <- data.frame(critics = c(60))
predict.lm(model, x0, interval = "prediction", conf.level = 0.95)
```

Interpret the interval in the context of the data.

STA 210