Simple Linear Regression

Inference & Prediction

Prof. Maria Tackett & Youngsoo Baek

01.27.20

Click for PDF of slides

Announcements

- Lab 02 due **Tuesday at 11:59p**
  - One person from each group submit the assignment on Gradescope.
  - Click to add all group members names to the submission.
- HW 01 due **Wed, Jan 29 at 11:59p**
- Fill out daily engagement survey (should receive it at the end of class)
- Reading for today
- No lecture on Wedneseday. Watch rstudio::conf Jan 29 - 30 and complete reflection by Jan 31 at 11:59p
  - Click here for details

Today's Agenda

- Model Assumptions
- Inference for regression
- Prediction (time permitting)

Packages and Data

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

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

rottentomatoes.com

What is the relationship between the critics' score and audience score for movies?

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 **83%** **88%**

TOMATOMETER Total Count: 129

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% **94%**

TOMATOMETER Total Count: 347

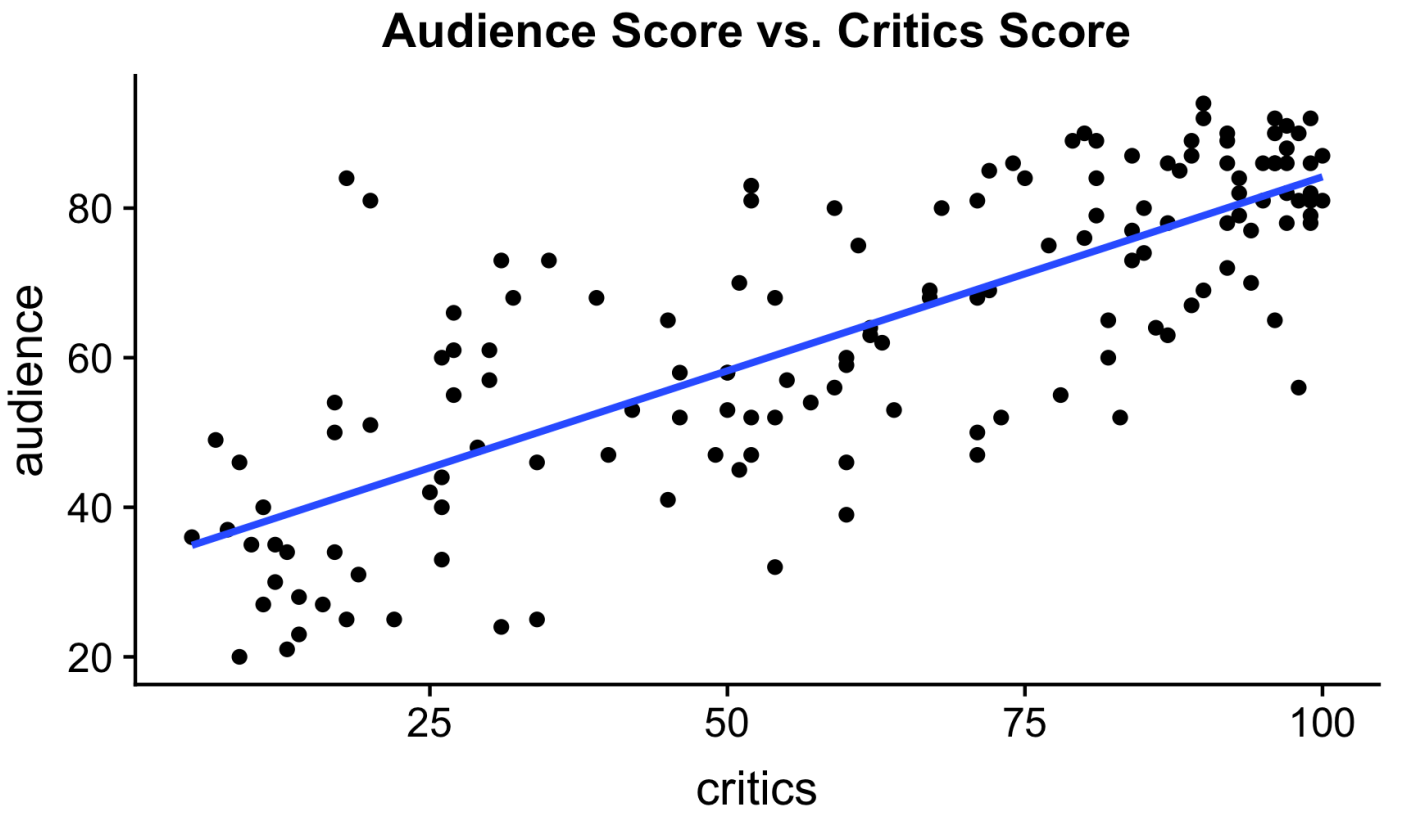
AUDIENCE SCORE Verified Ratings: 58,961

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

```
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

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

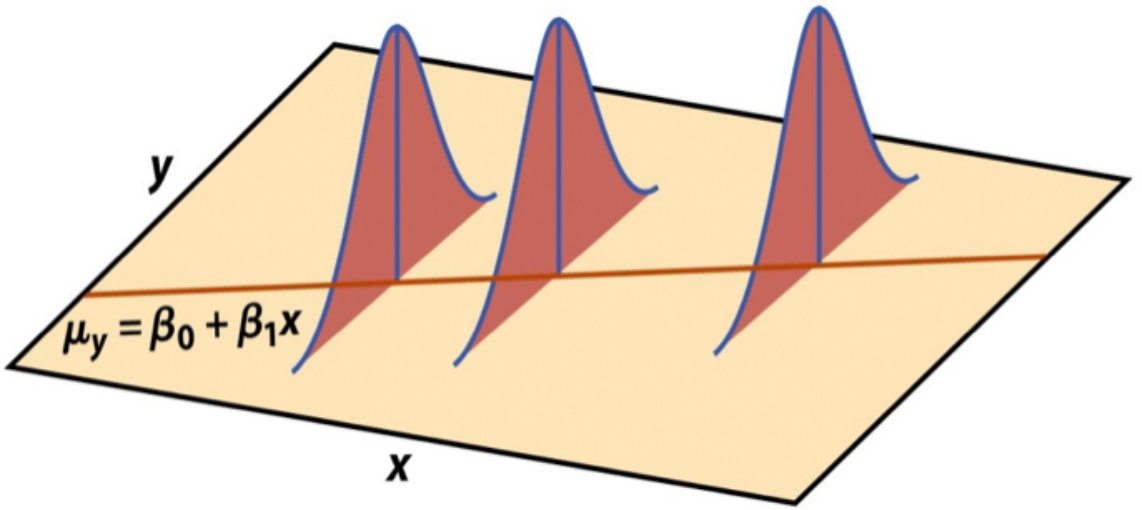
- **Slope:** For each additional percentage point in the critics score, the audience score is expected to increase by 0.519 percentage points on average.
- **Intercept:** If a movie gets a 0% from the critics, the audience score is expected to be 32.316%.

Checking Model Assumptions

Recall: The regression model

When we fit a simple linear regression model, we assume for every x_i the distribution of y is approximately Normal...

- with mean $\beta_0 + \beta_1 x_i$.
- and variance σ^2



Assumptions for Regression

When we fit a simple linear regression model, we must check that our data follows the relationship on the previous slide by checking the following assumptions...

1. **Linearity:** The plot of the mean value of y versus x falls on a straight line
2. **Constant Variance:** The variance of the distribution of y for a given value of x is σ^2, i.e. is the same for all values of x
3. **Normality:** For a given x, the distribution of y is Normal
4. **Independence:** All observations are independent

Checking assumptions

We will use various plots to check the assumptions for regression:

1. Scatterplot of y vs. x
    - Do this as part of the exploratory data analysis to see if there are any obvious departures from linearity
2. Plot of residuals vs. predictor variable
    - Check linearity condition
    - Check constant variance condition
3. Histogram and Normal QQ-Plot of residuals
    - Check Normality condition

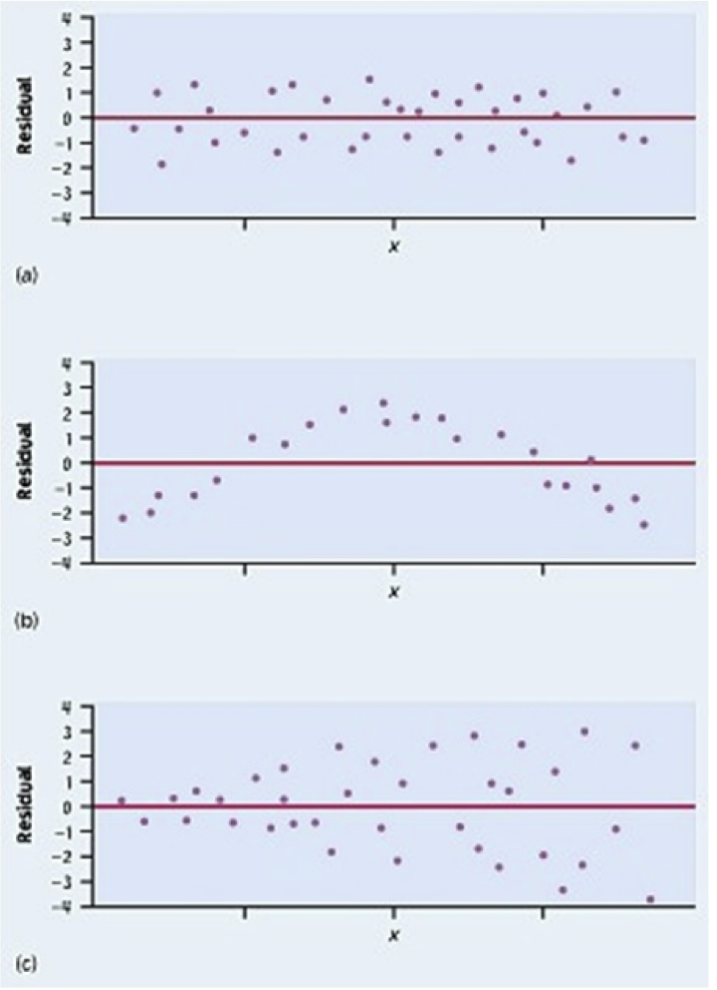
Linearity: Plot of residuals vs. predictor

- If the linear model, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, adequately describes the relationship between x and y, then the residuals should reflect random (chance) error
- To assess this, we can look at a plot of the residuals vs. the predictor variable
- **Linearity satisifed** if there is no distinguishable pattern in the residuals plot, i.e. the residuals should be randomly scattered
- A non-random pattern (e.g. a parabola) suggests a linear model does not adequately describe the relationship between x and y

Constant Variance: Plot of residuals vs. predictor

- If the spread of the distribution of y is equal for all values of x, then the spread of the residuals should be approximately equal for each value of x
- To assess this, we can look at a plot of the residuals vs. the predictor variable
- Constant variance satisifed if the vertical spread of the residuals is approximately equal as you move from left to right (i.e. there is no "fan" pattern)
- A fan pattern suggests the constant variance assumption is not satisified and transformation or some other remedy is required (more on this later in the semester)

Example residual plots



Ideal Residual Plot

Nonlinearity

Nonconstant Variance

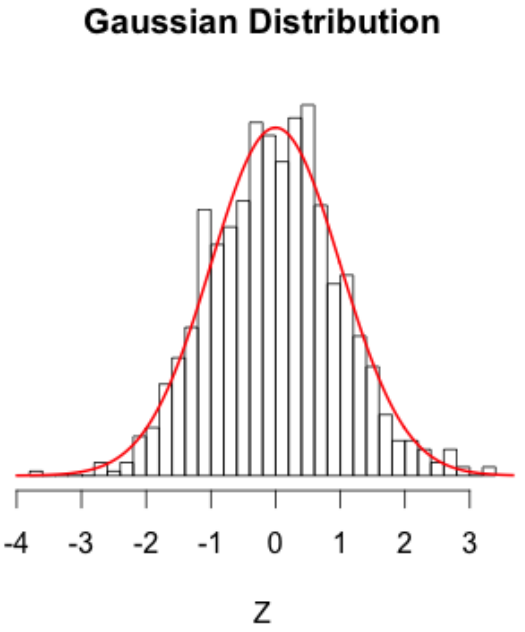
```
movie_scores <- movie_scores %>%
  mutate(residuals = resid(model))
```

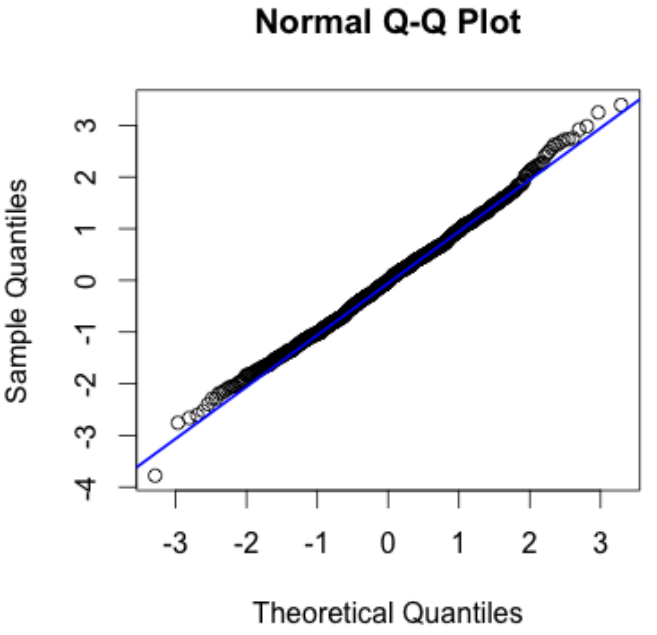
```
ggplot(data = movie_scores, mapping = aes(x = critics, y = residua
  geom_point() +
  geom_hline(yintercept = 0, color = "red")+
  labs(title = "Residuals vs. Critics Score")
```

Normality: Histogram & Normal QQ-Plot

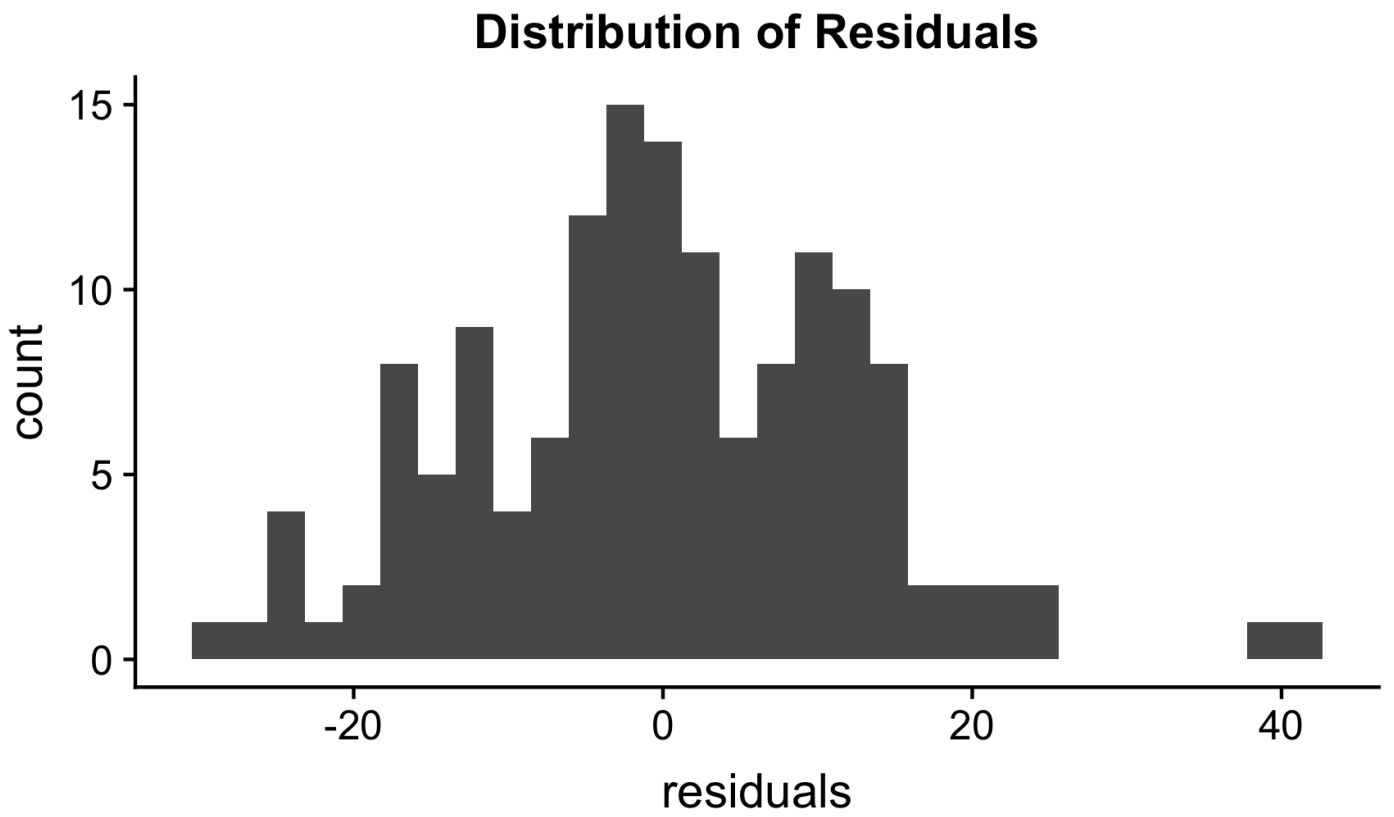
- The linear model assumes that the distribution of y is Normal for every value of x.
- This is impossible to check in practice, so we will look at the overall distribution of the residuals to assess if the Normality assumption is satisified
- Normaltiy satisified if a histogram of the residuals is approximately Normal
  - Can also check that a Normal QQ-plot falls along a diagonal line
- Note: Most inference methods for regression are robust to some departures from Normality

Understanding the Normal QQ plot

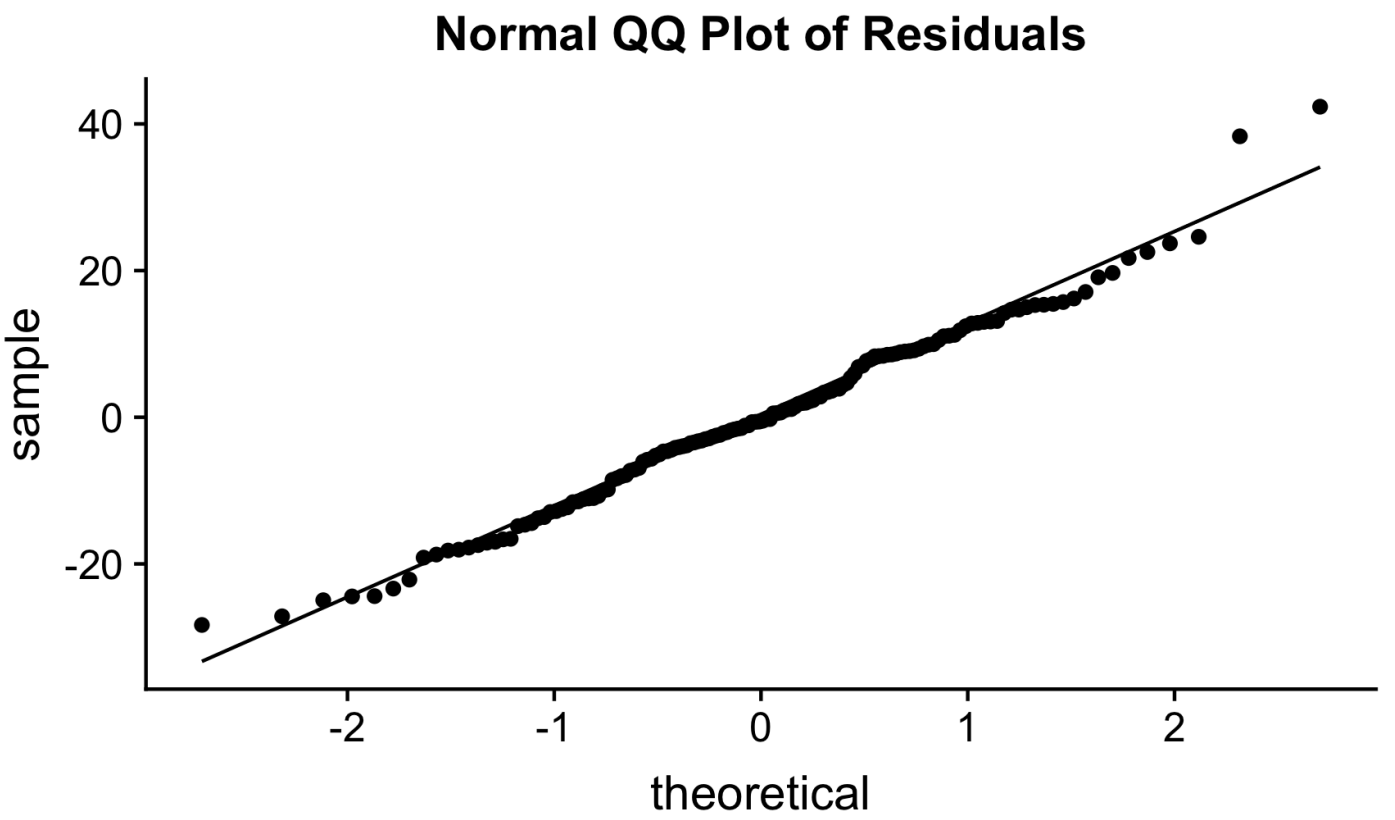




```
ggplot(data = movie_scores, mapping = aes(x = residuals)) +
  geom_histogram() +
  labs(title = "Distribution of Residuals")
```



```
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**: If there are subgroups represented in the data that are not accounted for in the model (e.g. type of movie), you can color the points in the residual plots by group to see if the model systematically over or under predicts for a particular subgroup

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 actually a linear relationship between `critics` and `audience`, or is the relationshp we observed due to random chance?
  - 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)

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

General form of the confidence interval

- 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

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

Confidence interval for β_1

- The confidence interval for the regression slope is

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

- 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

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

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 its 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 (same reason we use *t* when doing inference for a mean)
- The critical value t^* is calculated from the *t(n-2)* distribution - the *t* distribution with *n-2* degrees of freedom.

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 actually a linear relationship between `critics` and `audience`, or is the relationshp we observed due to random chance?

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

1. State the hypotheses

- We are often interested in testing whether there is a statistically significant linear relationship between the predictor and response variables
- If there is actually no linear relationship between the two variables, the population regression slope, β_1, would equal 0
- Therefore, let's test the hypotheses:

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

- These are the hypotheses corresponding to the output from the `lm` function

2. Calculate the test statistic

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

Test Statistic:

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

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

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 definition 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

Notes:

- These are general guidelines. The strength of evidence depends on the context of the problem.
- Don't just rely on the reject/fail to reject conclusion from the hypothesis test to draw conclusions about the relationship between x and y. Use the EDA and confidence intervals to provide more context about the implications of your results.

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.

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

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

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.

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.

How does the predicted value compare to the prediction for an individual movie? How does the interval compare?