Inference for Simple Linear Regression

Prof. Wells

STA 209, 5/8/23

Outline

In this lecture, we will. . .

- Review framework for linear regression
- Discuss inference procedures for linear models
- Review conditions for regression on linear models

Section 1

Simple Linear Regression

Review of Simple Linear Regression

- Previously, we used linear regression to analyze the relationship between two quantitative variables

Review of Simple Linear Regression

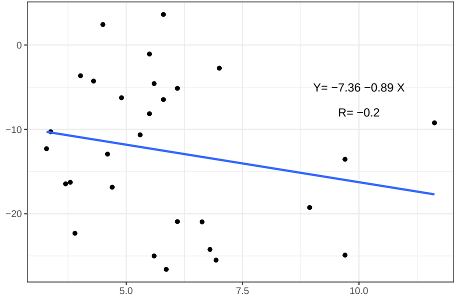
- Previously, we used linear regression to analyze the relationship between two quantitative variables
    - The strength and direction of the linear relationship is summarized by the correlation coefficient R

Review of Simple Linear Regression

- Previously, we used linear regression to analyze the relationship between two quantitative variables
  - The strength and direction of the linear relationship is summarized by the correlation coefficient R
  - The linear model $\hat{Y} = \beta_0 + \beta_1 X$ can be used to make predictions about Y using the values of X.

Review of Simple Linear Regression

- Previously, we used linear regression to analyze the relationship between two quantitative variables
  - The strength and direction of the linear relationship is summarized by the correlation coefficient R
  - The linear model $\hat{Y} = \beta_0 + \beta_1 X$ can be used to make predictions about Y using the values of X.



Linear Models in R

- To fit a linear model in R, use the `lm` function

```
my_mod <- lm(Y ~ X, data = my_data)
```

Linear Models in R

- To fit a linear model in R, use the `lm` function

```
my_mod <- lm(Y ~ X, data = my_data)
```

- To view coefficients of the model, use `get_regression_table` from `moderndive`

```
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 X            -0.89     0.835     -1.07   0.296    -2.60    0.824
```

Linear Models in R

- To fit a linear model in R, use the `lm` function

```
my_mod <- lm(Y ~ X, data = my_data)
```

- To view coefficients of the model, use `get_regression_table` from `moderndive`

```
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9      3.21
## 2 X            -0.89     0.835     -1.07   0.296    -2.60     0.824
```

- Correlation can be computed using `summarize` and `cor`:

```
my_data %>% summarize(R = cor(X,Y))
```

```
## # A tibble: 1 x 1
##        R
##    <dbl>
## 1 -0.201
```

Linear Models in R

- To fit a linear model in R, use the `lm` function

```
my_mod <- lm(Y ~ X, data = my_data)
```

- To view coefficients of the model, use `get_regression_table` from `moderndive`

```
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 X            -0.89     0.835     -1.07   0.296    -2.60    0.824
```

- Correlation can be computed using `summarize` and `cor`:

```
my_data %>% summarize(R = cor(X,Y))
```

```
## # A tibble: 1 x 1
##        R
##    <dbl>
## 1 -0.201
```

- We can fit a linear model to any data set we want.

Linear Models in R

- To fit a linear model in R, use the `lm` function

```
my_mod <- lm(Y ~ X, data = my_data)
```

- To view coefficients of the model, use `get_regression_table` from `moderndive`

```
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 X            -0.89     0.835     -1.07   0.296    -2.60    0.824
```

- Correlation can be computed using `summarize` and `cor`:

```
my_data %>% summarize(R = cor(X,Y))
```

```
## # A tibble: 1 x 1
##        R
##    <dbl>
## 1 -0.201
```

- We can fit a linear model to any data set we want.
    - But if we just have a *sample* of data, any trend we detect doesn't necessarily demonstrate that the trend exists in the *population*.

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

  - **Parameters** of interest:
    - β_0 (intercept)
    - β_1 (slope)
    - ρ (correlation)
    - σ (standard deviation of residuals)

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

  - **Parameters** of interest:
    - β_0 (intercept)
    - β_1 (slope)
    - ρ (correlation)
    - σ (standard deviation of residuals)

- But in general, we won't ever be able to know the true values of these parameters. So we estimate them based on sample data.

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

  - **Parameters** of interest:
    - β_0 (intercept)
    - β_1 (slope)
    - ρ (correlation)
    - σ (standard deviation of residuals)

- But in general, we won't ever be able to know the true values of these parameters. So we estimate them based on sample data.

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

Statistical Inference for Regression

Goal: Use *statistics* calculated from data to make inferences about the nature of *parameters*

- For regression, we can propose a model for the relationship between explanatory variable X and response variable Y:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

  - **Parameters** of interest:
    - β_0 (intercept)
    - β_1 (slope)
    - ρ (correlation)
    - σ (standard deviation of residuals)

- But in general, we won't ever be able to know the true values of these parameters. So we estimate them based on sample data.

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

  - **Statistics** from sample:
    - $\hat{\beta}_0$ (intercept)
    - $\hat{\beta}_1$ (slope)
    - R (correlation)
    - $\hat{\sigma}$ (standard error of residuals)

Midterm Elections

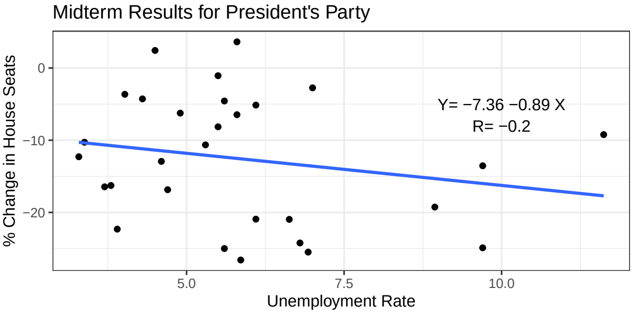
- Elections for the U.S. House of Representatives occur every two years, while elections for the U.S. president occurs every 4 years.
  - House elections in the middle of a Presidential term are called **midterm elections**.

Midterm Elections

- Elections for the U.S. House of Representatives occur every two years, while elections for the U.S. president occurs every 4 years.
  - House elections in the middle of a Presidential term are called **midterm elections**.
- One political theory suggests that high unemployment rate corresponds to worse performance by the President's party in midterm elections.

Midterm Elections

- Elections for the U.S. House of Representatives occur every two years, while elections for the U.S. president occurs every 4 years.
  - House elections in the middle of a Presidential term are called **midterm elections**.
- One political theory suggests that high unemployment rate corresponds to worse performance by the President's party in midterm elections.



Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020

Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020
  - Results during the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively.

Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020
  - Results during the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively.
- Our data is *not* a sample from historical midterm elections

Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020
  - Results during the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively.
- Our data is *not* a sample from historical midterm elections
- But we can treat the many effects complicated effects that influence midterm performance as random variables
  - We can create a model for midterm performance, and treat our data as a random sample from the collection of all theoretical midterm election results according to this model

Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020
  - Results during the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively.
- Our data is *not* a sample from historical midterm elections
- But we can treat the many effects complicated effects that influence midterm performance as random variables
  - We can create a model for midterm performance, and treat our data as a random sample from the collection of all theoretical midterm election results according to this model
- Not every random sample from this model will be have the same regression statistics (slope, intercept, correlation, standard deviation of residuals)

Unemployment Model

- Our data consists of results for (almost) all midterm elections between 1900 and 2020
  - Results during the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively.
- Our data is *not* a sample from historical midterm elections
- But we can treat the many effects complicated effects that influence midterm performance as random variables
  - We can create a model for midterm performance, and treat our data as a random sample from the collection of all theoretical midterm election results according to this model
- Not every random sample from this model will be have the same regression statistics (slope, intercept, correlation, standard deviation of residuals)
- We're interested in assessing how much these statistics may change, just due to the randomness in this model

Section 2

Hypothesis Tests

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: There is no linear relationship between Unemployment X and Percent Change in Midterm Seats Y
- **Alternative Hypothesis**: There is a negative linear relationship between Unemployment X and Midterm Results Y

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 < 0$$

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: There is no linear relationship between Unemployment X and Percent Change in Midterm Seats Y
- **Alternative Hypothesis**: There is a negative linear relationship between Unemployment X and Midterm Results Y

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 < 0$$

Method

- If there is no linear relationship, then the pairing between X and Y is superficial and we can shuffle the values of Y among the values of X to simulate a similar data set:

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: There is no linear relationship between Unemployment X and Percent Change in Midterm Seats Y
- **Alternative Hypothesis**: There is a negative linear relationship between Unemployment X and Midterm Results Y

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 < 0$$

Method

- If there is no linear relationship, then the pairing between X and Y is superficial and we can shuffle the values of Y among the values of X to simulate a similar data set:
  - For each midterm election, record unemployment rate, but randomly choose percent change in house seats from among all recorded percent changes (without replacement)

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: There is no linear relationship between Unemployment X and Percent Change in Midterm Seats Y
- **Alternative Hypothesis**: There is a negative linear relationship between Unemployment X and Midterm Results Y

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 < 0$$

Method

- If there is no linear relationship, then the pairing between X and Y is superficial and we can shuffle the values of Y among the values of X to simulate a similar data set:
  - For each midterm election, record unemployment rate, but randomly choose percent change in house seats from among all recorded percent changes (without replacement)
  - Compute the slope of the regression model for this simulated data set

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: There is no linear relationship between Unemployment X and Percent Change in Midterm Seats Y
- **Alternative Hypothesis**: There is a negative linear relationship between Unemployment X and Midterm Results Y

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 < 0$$

Method

- If there is no linear relationship, then the pairing between X and Y is superficial and we can shuffle the values of Y among the values of X to simulate a similar data set:
  - For each midterm election, record unemployment rate, but randomly choose percent change in house seats from among all recorded percent changes (without replacement)
  - Compute the slope of the regression model for this simulated data set
  - Repeat several times to assess variability in slope assuming H_0 is true

A Few Shuffles

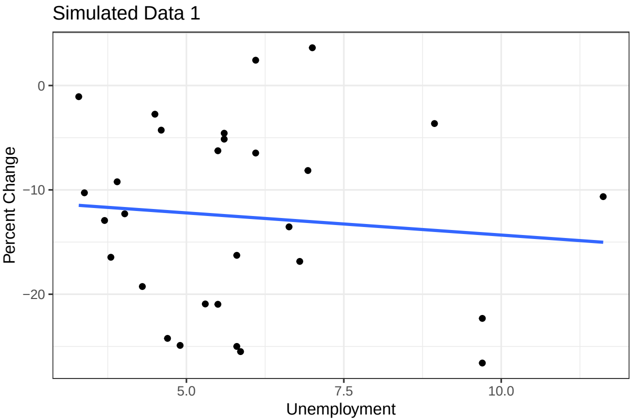
```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(1, type = "permute")
```

```
## # A tibble: 6 x 2
##   house_change unemp
##          <dbl> <dbl>
## 1        -10.6  11.6
## 2        -19.3   4.3
## 3         -1.07  3.29
## 4        -25.5   5.86
## 5        -13.5   6.63
## 6        -10.3   3.38
```

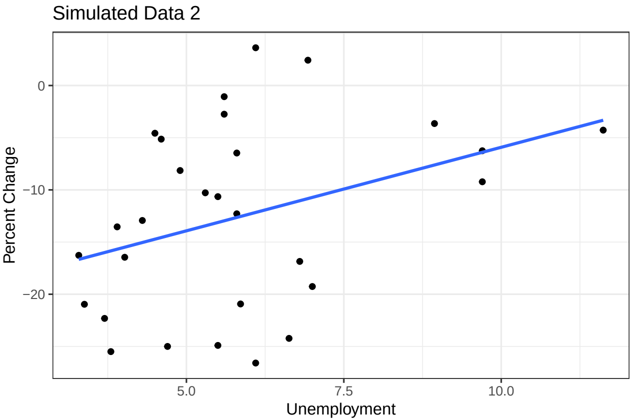
```
## # A tibble: 6 x 2
##   house_change unemp
##          <dbl> <dbl>
## 1         -4.28 11.6
## 2        -12.9   4.3
## 3        -16.3   3.29
## 4        -20.9   5.86
## 5        -24.2   6.63
## 6        -21.0   3.38
```

```
## # A tibble: 6 x 2
##   house_change unemp
##          <dbl> <dbl>
## 1        -16.3  11.6
## 2         -9.22  4.3
## 3        -10.6   3.29
## 4         -4.57  5.86
## 5        -12.9   6.63
## 6         -2.75  3.38
```

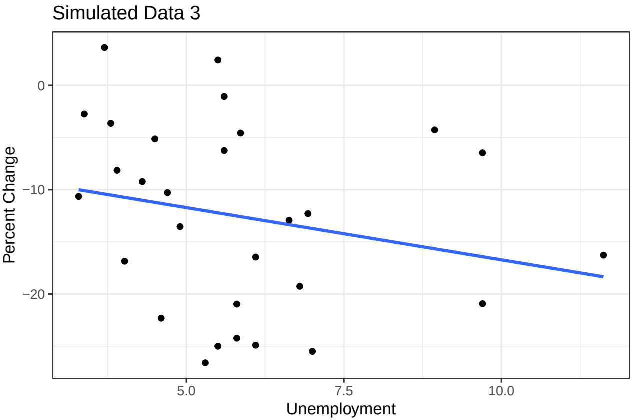
Scatterplots of Synthetic Data I



Scatterplots of Synthetic Data II



Scatterplots of Synthetic Data III



Note: location of individual points change, but general clusters do not.

Calculate Statistics

Now we generate 1000 replicates, and compute the slope of the regression line for each

Calculate Statistics

Now we generate 1000 replicates, and compute the slope of the regression line for each

```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(1000, type = "permute")
  calculate( stat = "slope")
```

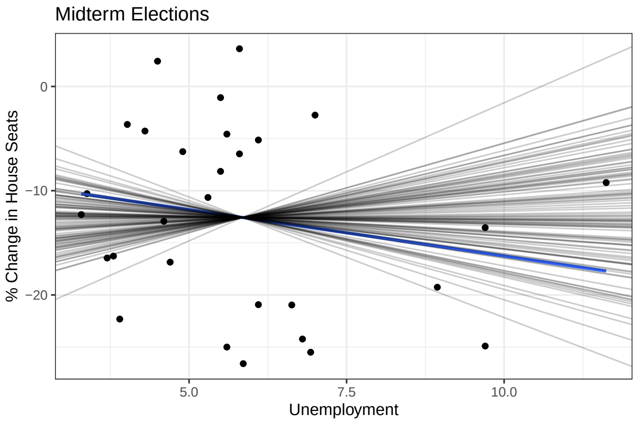
Calculate Statistics

Now we generate 1000 replicates, and compute the slope of the regression line for each

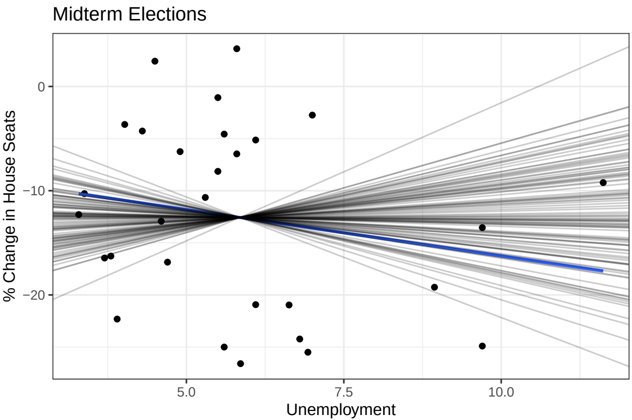
```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  hypothesize(null = "independence") %>%
  generate(1000, type = "permute")
  calculate( stat = "slope")
```

```
## Response: house_change (numeric)
## Explanatory: unemp (numeric)
## Null Hypothesis: independence
## # A tibble: 6 x 2
##   replicate    stat
##       <int>   <dbl>
## 1         1 -0.105
## 2         2 -1.23
## 3         3  0.0265
## 4         4 -0.931
## 5         5  0.600
## 6         6 -0.0527
```

Visualizing 1000 Slopes

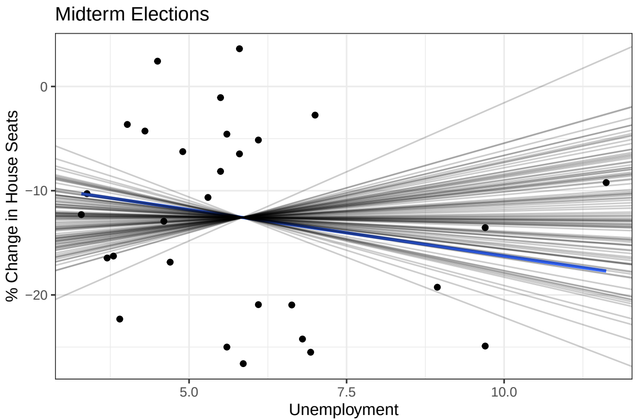


Visualizing 1000 Slopes



- Most lines are approximately horizontal. But some have positive or negative slope.

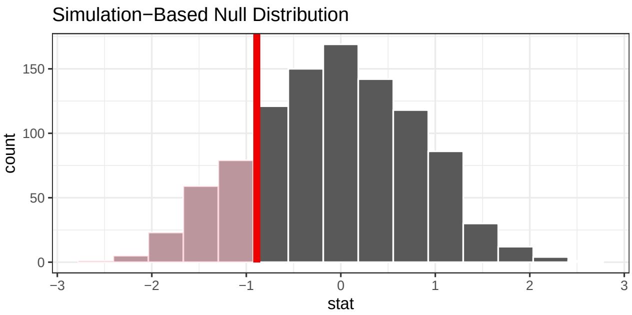
Visualizing 1000 Slopes



- Most lines are approximately horizontal. But some have positive or negative slope.
- The linear regression line for the original data is shown in blue.

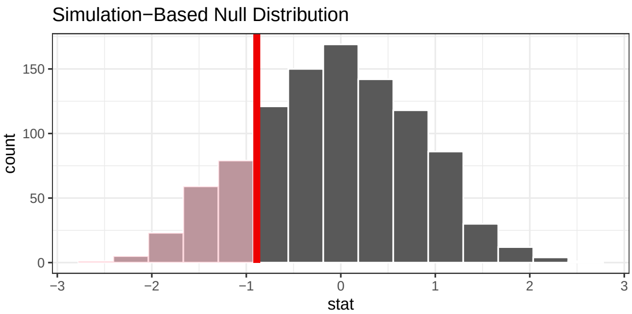
The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.89, direction = "left")
```



The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.89, direction = "left")
```




```
null_slope %>% get_p_value(obs_stat = -0.89, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.179
```

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
    - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
    - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats
- Does this mean there is **no** relationship between Unemployment and Change in House Seats?

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
  - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats
- Does this mean there is **no** relationship between Unemployment and Change in House Seats?
  - No! Failing to reject H_0 is not the same as showing that H_0 is true.

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
  - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats
- Does this mean there is **no** relationship between Unemployment and Change in House Seats?
  - No! Failing to reject H_0 is not the same as showing that H_0 is true.
  - Perhaps there is a small effect, but our sample size was insufficient to detect it

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
  - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats
- Does this mean there is **no** relationship between Unemployment and Change in House Seats?
  - No! Failing to reject H_0 is not the same as showing that H_0 is true.
  - Perhaps there is a small effect, but our sample size was insufficient to detect it
  - Perhaps there is an effect, but it is non-linear

Conclusion

With a P-value of 0.179, which is greater than $\alpha = 0.05$, we fail to reject H_0

- A slope like this is consistent with those arising due to chance if there were no relationship between Unemployment and Change in House Seats.
  - The data does not provide evidence of a linear relationship between Unemployment and Change in House Seats
- Does this mean there is **no** relationship between Unemployment and Change in House Seats?
  - No! Failing to reject H_0 is not the same as showing that H_0 is true.
  - Perhaps there is a small effect, but our sample size was insufficient to detect it
  - Perhaps there is an effect, but it is non-linear
  - Perhaps there is an effect, but it is masked by other confounding variabes.

Section 3

Confidence Intervals

Confidence Intervals for Linear Models

- A hypothesis test allows us to assess the strength of evidence of a claim, while a confidence interval allows us to assess the magnitude of an effect.

Confidence Intervals for Linear Models

- A hypothesis test allows us to assess the strength of evidence of a claim, while a confidence interval allows us to assess the magnitude of an effect.
- Suppose Percent Change in Seats could be perfectly predicted by Unemployment Rate (with no deviations or errors). What slope would we expect to find in the linear regression model?

Confidence Intervals for Linear Models

- A hypothesis test allows us to assess the strength of evidence of a claim, while a confidence interval allows us to assess the magnitude of an effect.
- Suppose Percent Change in Seats could be perfectly predicted by Unemployment Rate (with no deviations or errors). What slope would we expect to find in the linear regression model?
    - It's impossible to say without knowing the variability in the unemployment and percent change data.
    - Reminder: slope tells us the average increase in the response variable per unit increase in the explanatory variable

Confidence Intervals for Linear Models

- A hypothesis test allows us to assess the strength of evidence of a claim, while a confidence interval allows us to assess the magnitude of an effect.
- Suppose Percent Change in Seats could be perfectly predicted by Unemployment Rate (with no deviations or errors). What slope would we expect to find in the linear regression model?
  - It's impossible to say without knowing the variability in the unemployment and percent change data.
  - Reminder: slope tells us the average increase in the response variable per unit increase in the explanatory variable
- If we want to estimate the strength of the linear relationship between the two variables, we should instead create a confidence interval for the correlation R.

Bootstrapping for confidence intervals

- To approximate variability in the correlation statistic R, we create a bootstrap sample by resampling the paired data and then calculation correlation
    - This corresponds to sampling with replacement from the columns of the original sample

Bootstrapping for confidence intervals

- To approximate variability in the correlation statistic R, we create a bootstrap sample by resampling the paired data and then calculation correlation
  - This corresponds to sampling with replacement from the columns of the original sample

```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  generate(1, type = "bootstrap")
```

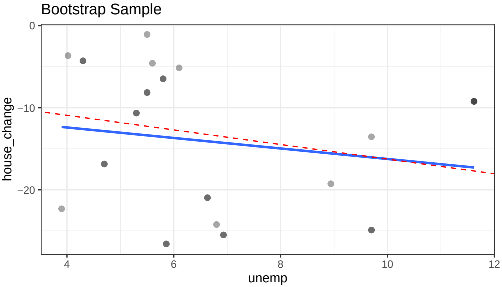
```
## # A tibble: 6 x 2
##   house_change unemp
##          <dbl> <dbl>
## 1        -13.5   9.7
## 2         -6.47  5.8
## 3        -10.6   5.3
## 4        -25.5   6.93
## 5        -26.6   5.86
## 6        -19.3   8.94
```

Bootstrapping for confidence intervals

- To approximate variability in the correlation statistic R, we create a bootstrap sample by resampling the paired data and then calculation correlation
  - This corresponds to sampling with replacement from the columns of the original sample

```r
midterms_house %>%
  specify(house_change ~ unemp) %>%
  generate(1, type = "bootstrap")
```

```
## # A tibble: 6 x 2
##   house_change unemp
##          <dbl> <dbl>
## 1        -13.5   9.7
## 2         -6.47  5.8
## 3        -10.6   5.3
## 4        -25.5   6.93
## 5        -26.6   5.86
## 6        -19.3   8.94
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.175
```



- Dashed red line indicates regression line for original sample
- Darker points correspond to observations included in bootstrap more than once

Bootstrap Distribution for correlation

Now we generate 1000 replicates, and compute the correlation for each

Bootstrap Distribution for correlation

Now we generate 1000 replicates, and compute the correlation for each

```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  generate(1000, type = "bootstrap") %>%
  calculate(stat = "correlation")
```

Bootstrap Distribution for correlation

Now we generate 1000 replicates, and compute the correlation for each

```
midterms_house %>%
  specify(house_change ~ unemp) %>%
  generate(1000, type = "bootstrap") %>%
  calculate(stat = "correlation")
```

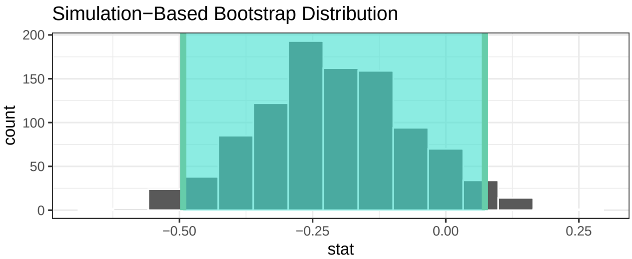
```
## Response: house_change (numeric)
## Explanatory: unemp (numeric)
## # A tibble: 6 x 2
##   replicate    stat
##       <int>   <dbl>
## 1         1 -0.305
## 2         2 -0.0639
## 3         3 -0.0805
## 4         4 -0.0308
## 5         5 -0.193
## 6         6 -0.322
```

The Bootstrap Distribution for R

A 95% confidence interval for correlation ρ is

```
boot_slope %>% get_ci(level = .95, type = "percentile")
```

```
##   lower_ci upper_ci
## 1    -0.49    0.073
```

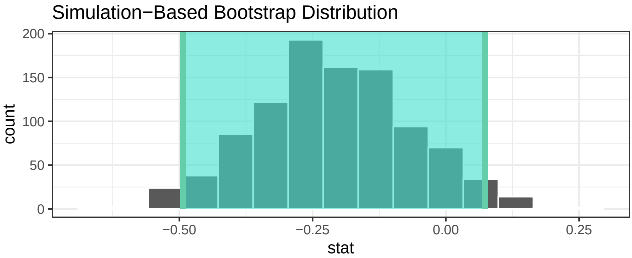


The Bootstrap Distribution for R

A 95% confidence interval for correlation ρ is

```
boot_slope %>% get_ci(level = .95, type = "percentile")
```

```
##   lower_ci upper_ci
## 1    -0.49    0.073
```



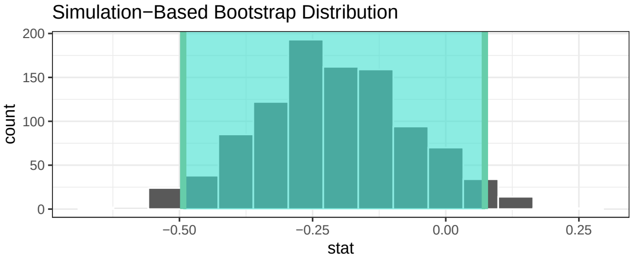
- The original sample had correlation $R = \text{-}0.2$
    - It is possible that Unemployment and Percent Change has between moderately negative correlation (-0.49) and very weak positive correlation (0.07).

The Bootstrap Distribution for R

A 95% confidence interval for correlation ρ is

```
boot_slope %>% get_ci(level = .95, type = "percentile")
```

```
##   lower_ci upper_ci
## 1    -0.49    0.073
```



- The original sample had correlation $R = $ -0.2
  - It is possible that Unemployment and Percent Change has between moderately negative correlation (-0.49) and very weak positive correlation (0.07).
  - It's also plausible that the two variables have 0 correlation.

Section 4

Conditions for Inference

Conditions for Inference: LINE!

In order to responsibly use linear regression for prediction or inference, we require:

Conditions for Inference: LINE!

In order to responsibly use linear regression for prediction or inference, we require:

1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot/residual plot

Conditions for Inference: LINE!

In order to responsibly use linear regression for prediction or inference, we require:

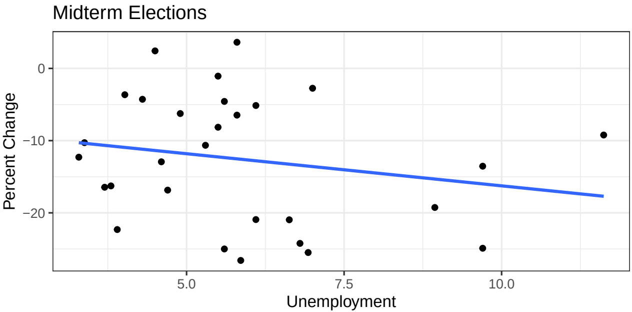
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
   - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be Normally distributed. (**Normal**)
   - Check using histogram of residuals

Conditions for Inference: LINE!

In order to responsibly use linear regression for prediction or inference, we require:

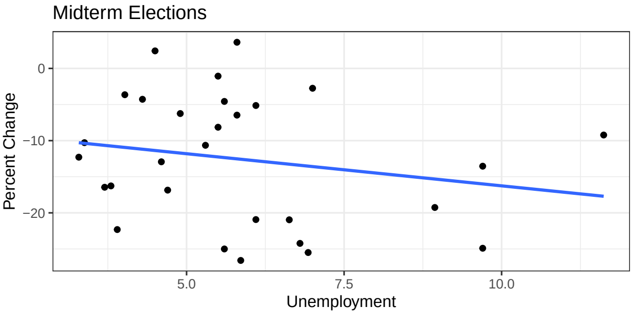
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
   - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be Normally distributed. (**Normal**)
   - Check using histogram of residuals
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)
   - Check using residual plot.

Checking Conditions: Linear



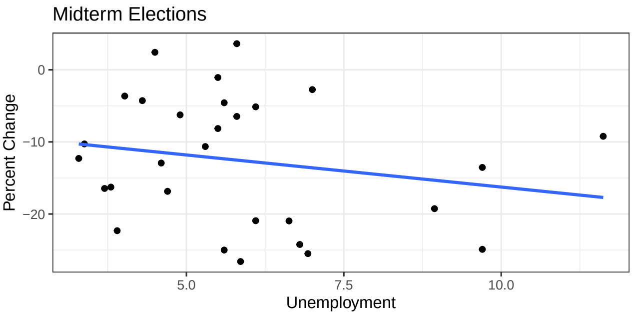
- Data is not tightly clustered around line of best fit

Checking Conditions: Linear



- Data is not tightly clustered around line of best fit
  - But this doesn't mean data is not linear. Just that residuals have high variance

Checking Conditions: Linear



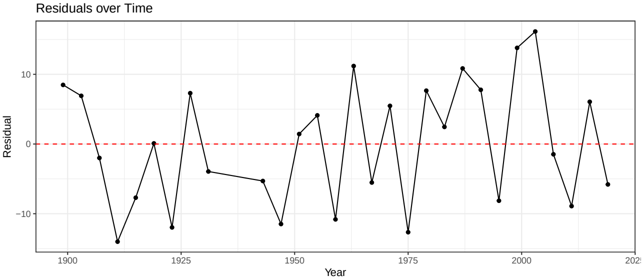
- Data is not tightly clustered around line of best fit
  - But this doesn't mean data is not linear. Just that residuals have high variance
  - Scatterplot does not show signs of **NON**-linear relationship

Checking Conditions: Independence

- The assumption that observations are independent is the most important for inference, but also most difficult to check.
  - Data representing repeated observations over time is particular susceptible to dependence
  - Consecutive observations in a time interval may have unwanted correlation

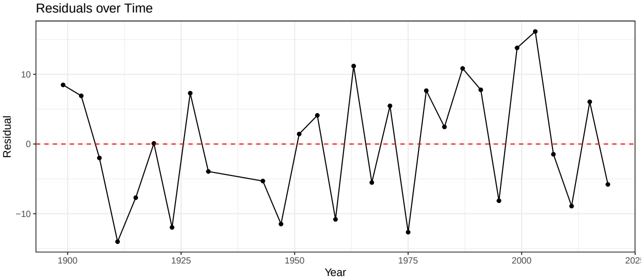
Checking Conditions: Independence

- The assumption that observations are independent is the most important for inference, but also most difficult to check.
  - Data representing repeated observations over time is particular susceptible to dependence
  - Consecutive observations in a time interval may have unwanted correlation



Checking Conditions: Independence

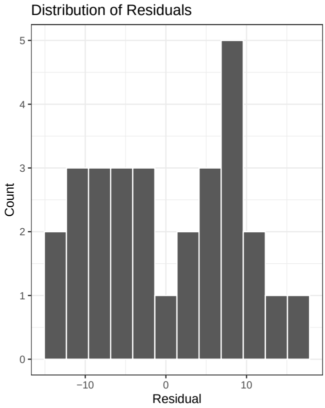
- The assumption that observations are independent is the most important for inference, but also most difficult to check.
  - Data representing repeated observations over time is particular susceptible to dependence
  - Consecutive observations in a time interval may have unwanted correlation



- Here, variables over time do not show strong consistent patterns

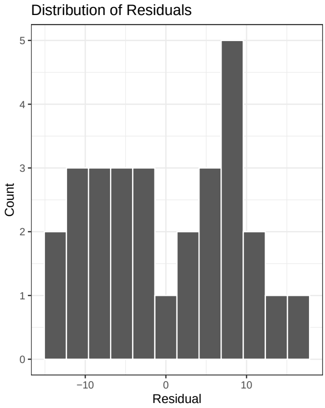
Checking Conditions: Normal

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes( x = residual))+geom_histogram(bins = 12, color = "white")
```



Checking Conditions: Normal

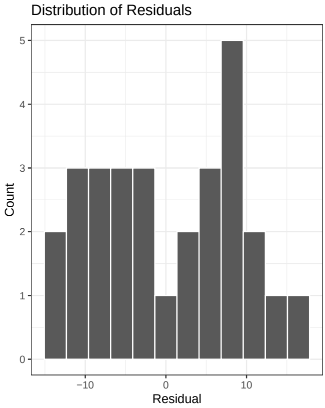
```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes( x = residual))+geom_histogram(bins = 12, color = "white")
```



- The distribution appears somewhat symmetric, although with some evidence of bimodality

Checking Conditions: Normal

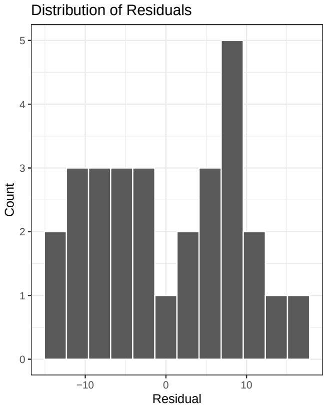
```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes( x = residual))+geom_histogram(bins = 12, color = "white")
```



- The distribution appears somewhat symmetric, although with some evidence of bimodality
- This provides some evidence residuals are not Normally distributed.

Checking Conditions: Normal

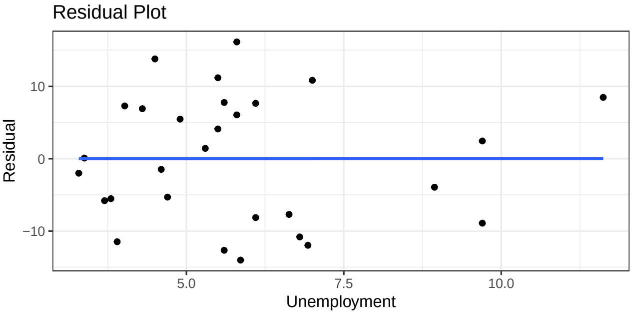
```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes( x = residual))+geom_histogram(bins = 12, color = "white")
```



- The distribution appears somewhat symmetric, although with some evidence of bimodality
- This provides some evidence residuals are not Normally distributed.
- This doesn't mean we discard analysis entirely, but we should be more cautious about inferential conclusions

Checking Conditions: Equal Variability

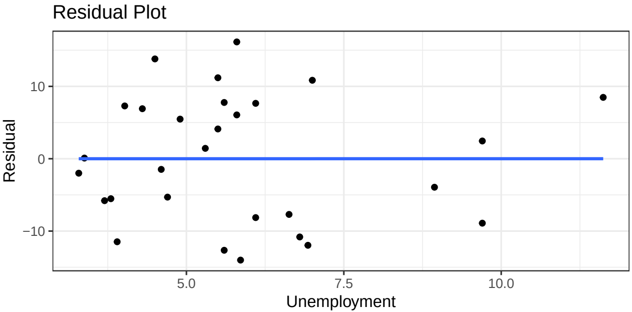
```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes(x = unemp, y = residual))+geom_point()
```



Residuals appear to have constant variability for unemployment between 2 and 7.5

Checking Conditions: Equal Variability

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
mod_residuals <- get_regression_points(my_mod)

ggplot(mod_residuals, aes(x = unemp, y = residual))+geom_point()
```



Residuals appear to have constant variability for unemployment between 2 and 7.5

- However, data with unemployment greater than 8 is relatively sparse, making it more difficult to assess variability

Section 5

Theory-Based Methods

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(DON'T MEMORIZE!)}$$

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(DON'T MEMORIZE!)}$$

  - In practice, we estimate β_0, β_1 in the formula using $\hat{\beta}_0, \hat{\beta}_1$.

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(DON'T MEMORIZE!)}$$

  - In practice, we estimate β_0, β_1 in the formula using $\hat{\beta}_0, \hat{\beta}_1$.
- We perfom a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{\hat{\beta}_1 - 0}{SE}$$

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \text{(DON'T MEMORIZE!)}$$

  - In practice, we estimate β_0, β_1 in the formula using $\hat{\beta}_0, \hat{\beta}_1$.
- We perfom a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{\hat{\beta}_1 - 0}{SE}$$

- And we create a confidence interval for β_1 using

$$\text{sample stat} \pm t^* \cdot SE \quad = \quad \hat{\beta}_1 \pm t^* \cdot SE$$

Inference for Slope

- Can we make inference about the slope β_1 of a linear model without using simulation?
  - We need to know the *mean*, *standard error*, and *shape* of the sampling distribution for $\hat{\beta}_1$
- If LINE conditions are satisfied, then $\hat{\beta}_1$ is Normally distributed with mean β_1.
  - And the standard error is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2}\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(DON'T MEMORIZE!)}$$

  - In practice, we estimate β_0, β_1 in the formula using $\hat{\beta}_0, \hat{\beta}_1$.
- We perfom a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{\hat{\beta}_1 - 0}{SE}$$

- And we create a confidence interval for β_1 using

$$\text{sample stat} \pm t^* \cdot SE \quad = \quad \hat{\beta}_1 \pm t^* \cdot SE$$

- In both cases, the reference distribution is the t-distribution with $n-2$ degrees of freedom.

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?
    - Yes! Using the `lm` function in R.

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?
    - Yes! Using the `lm` function in R.

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 unemp        -0.89     0.835     -1.07   0.296    -2.60    0.824
```

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?
    - Yes! Using the `lm` function in R.

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165  -17.9      3.21
## 2 unemp        -0.89     0.835     -1.07   0.296   -2.60     0.824
```

- The theory-based standard error is `std_error`, the test statistic is `statistic`, and the corresponding p-value in the t-distribution with n-2 df is `p_value`.

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?
  - Yes! Using the `lm` function in R.

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 unemp        -0.89     0.835     -1.07   0.296    -2.60    0.824
```

- The theory-based standard error is `std_error`, the test statistic is `statistic`, and the corresponding p-value in the t-distribution with n-2 df is `p_value`.
- The upper and lower bounds for the 95% confidence interval are `lower_ci` and `upper_ci`

Calculating test statistics and confidence intervals

- Can we get test statisics and confidence intervals for β_1 *without* tedious calculation?
  - Yes! Using the `lm` function in R.

```
my_mod <- lm(house_change ~ unemp, data = midterms_house)
get_regression_table(my_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    -7.36     5.16      -1.43   0.165   -17.9     3.21
## 2 unemp        -0.89     0.835     -1.07   0.296    -2.60    0.824
```

- The theory-based standard error is `std_error`, the test statistic is `statistic`, and the corresponding p-value in the t-distribution with n-2 df is `p_value`.
- The upper and lower bounds for the 95% confidence interval are `lower_ci` and `upper_ci`
- The table also gives similar information for the intercept and hypothesis test $H_0 : \beta_0 = 0$ (but this is less useful in practice)

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

- To test the hypothesis $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, use the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$$

where t follows the t-distribution with $n - 2$ degrees of freedom.

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

- To test the hypothesis $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, use the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$$

  where t follows the t-distribution with $n - 2$ degrees of freedom.

- There is a formula for confidence intervals, but it is considerably more complicated.

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

- To test the hypothesis $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, use the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$$

  where t follows the t-distribution with $n - 2$ degrees of freedom.

- There is a formula for confidence intervals, but it is considerably more complicated.
  - This is because the sampling distribution for R is highly skewed unless R is close to 0

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

- To test the hypothesis $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, use the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$$

  where t follows the t-distribution with $n - 2$ degrees of freedom.

- There is a formula for confidence intervals, but it is considerably more complicated.
  - This is because the sampling distribution for R is highly skewed unless R is close to 0
  - Therefore, we can't use the Normal approximation for R unless either the sample size is very large, or R is close to 0.

Inference for Correlation

- Suppose we are interested in investigating the correlation ρ between two variables
- The standard error for the sample correlation R when $\rho = 0$ is

$$SE(R) = \sqrt{\frac{1 - R^2}{n - 2}}$$

- To test the hypothesis $H_0 : \rho = 0$ against $H_a : \rho \neq 0$, use the test statistic

$$t = \frac{\text{sample stat} - \text{null value}}{SE} = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$$

  where t follows the t-distribution with $n - 2$ degrees of freedom.

- There is a formula for confidence intervals, but it is considerably more complicated.
  - This is because the sampling distribution for R is highly skewed unless R is close to 0
  - Therefore, we can't use the Normal approximation for R unless either the sample size is very large, or R is close to 0.
  - This is one situation where the simulation-based method clearly outperforms the theory-based method