Inference for Linear Regression

Nate Wells

Math 141, 4/18/22

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

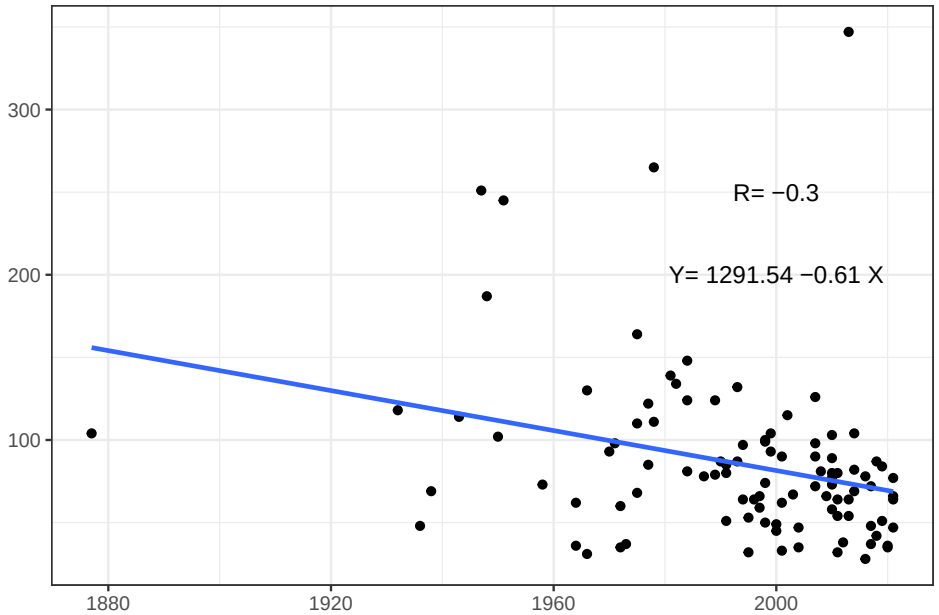
- Previously, we used linear regression to analyze the relationship between two quantitative variables
  - The strength and direction of the linear relationship is summarized by the correlation coefficient R

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 1292.      394.         3.28   0.001  509.    2074.
## 2 X           -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

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 1292.      394.         3.28   0.001  509.     2074.
## 2 X           -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

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

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

```
## # A tibble: 1 x 1
##         R
##     <dbl>
## 1 -0.295
```

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 1292.      394.         3.28   0.001  509.     2074.
## 2 X           -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

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

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

```
## # A tibble: 1 x 1
##         R
##     <dbl>
## 1 -0.295
```

- 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 1292.      394.         3.28   0.001  509.     2074.
## 2 X           -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

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

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

```
## # A tibble: 1 x 1
##         R
##     <dbl>
## 1 -0.295
```

- 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.

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

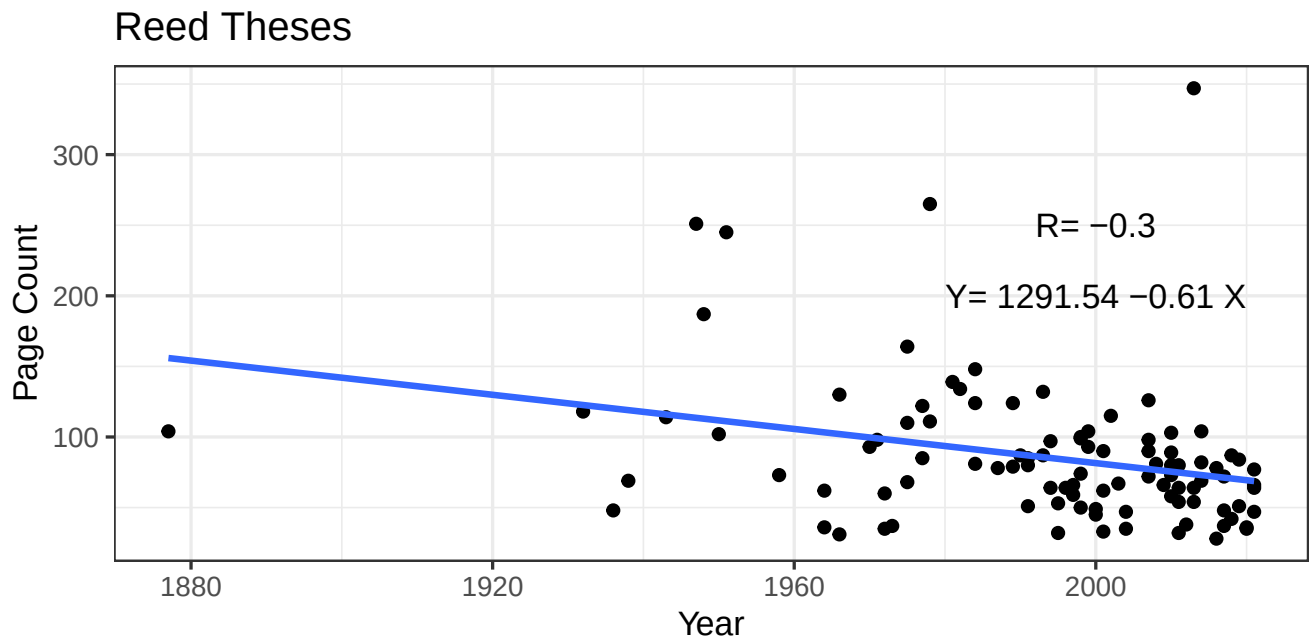
  - **Statistics** from sample:
    - $\hat{\beta}_0$ (intercept)
    - $\hat{\beta}_1$ (slope)
    - R (correlation)
    - $\hat{\sigma}$ (standard error of residuals)

Reed Thesis

- Earlier this year, Math 141 students collected data on several hundred senior theses from thesis tower.

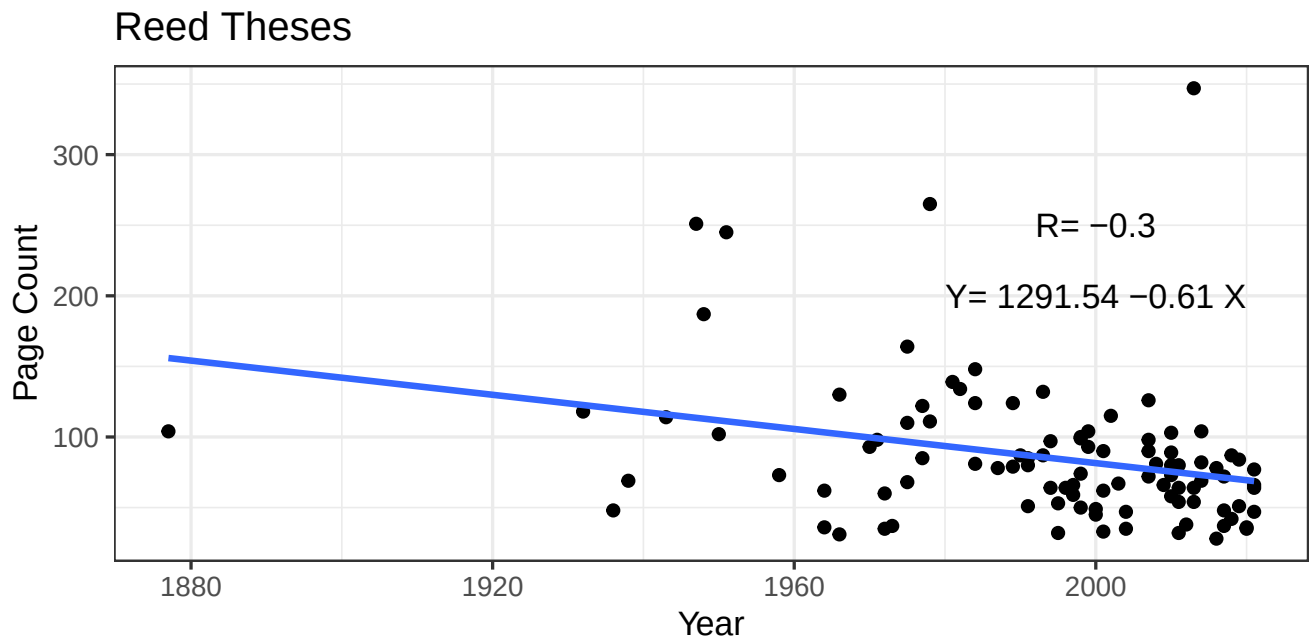
Reed Thesis

- Earlier this year, Math 141 students collected data on several hundred senior theses from thesis tower.
  - Page Count and Year Published for several of these theses are shown below:



Reed Thesis

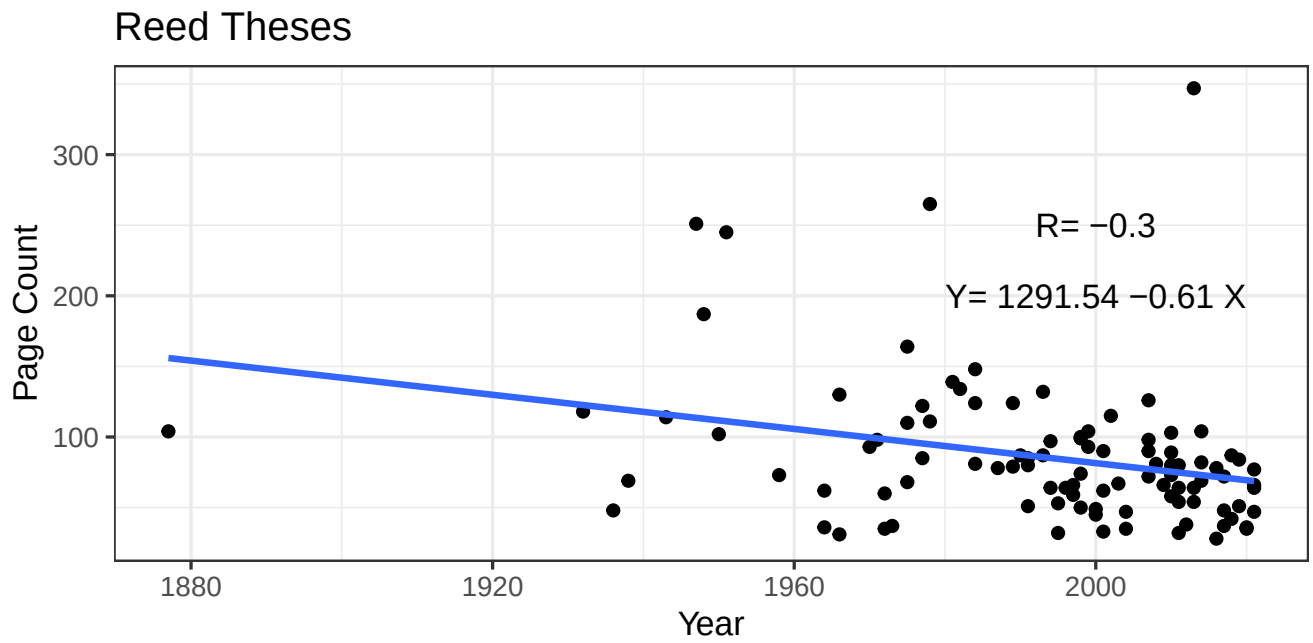
- Earlier this year, Math 141 students collected data on several hundred senior theses from thesis tower.
  - Page Count and Year Published for several of these theses are shown below:



- But this is just a sample of data. Would a different sample produce a different regression line?

Reed Thesis

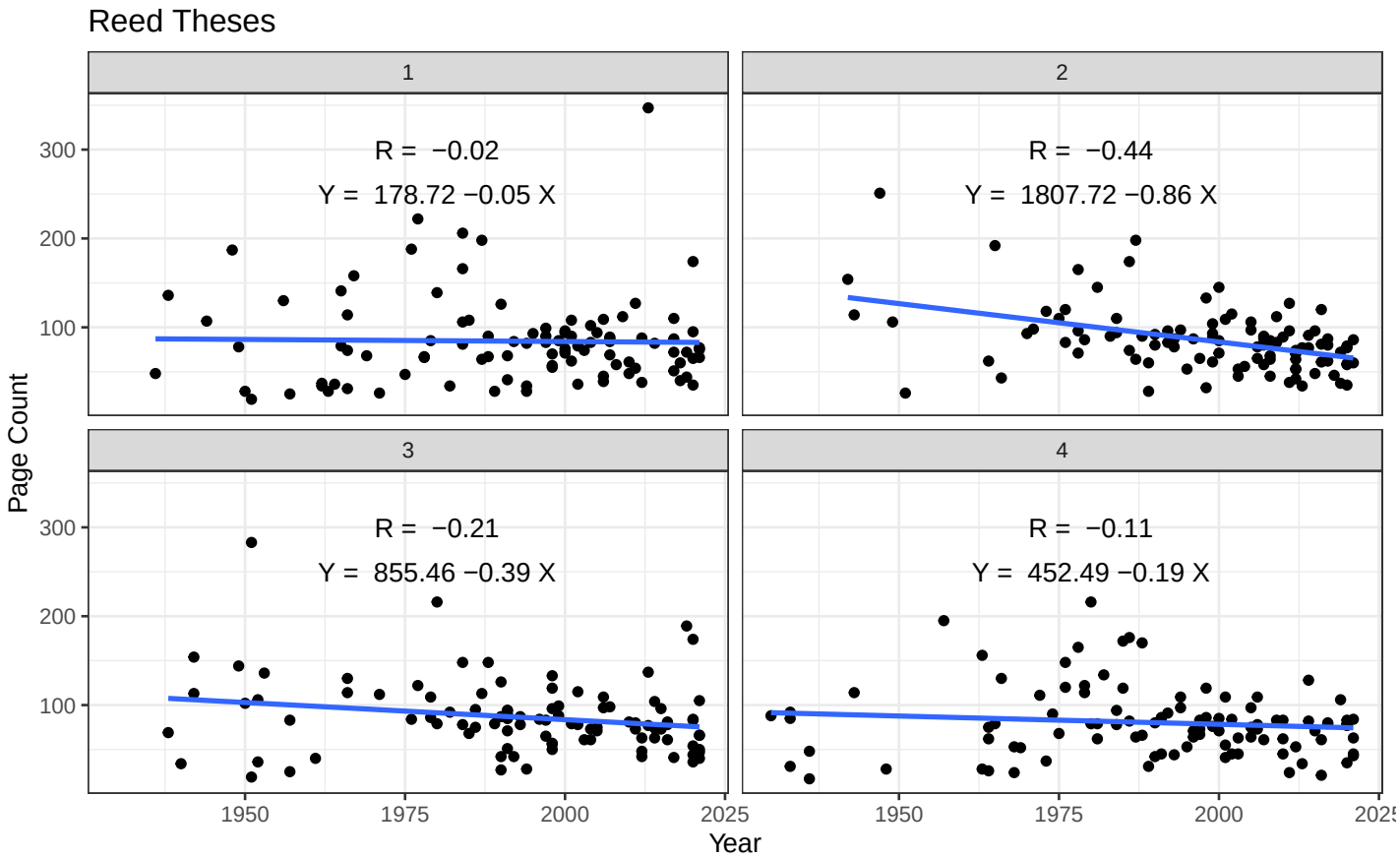
- Earlier this year, Math 141 students collected data on several hundred senior theses from thesis tower.
  - Page Count and Year Published for several of these theses are shown below:



- But this is just a sample of data. Would a different sample produce a different regression line?
  - Almost certainly!

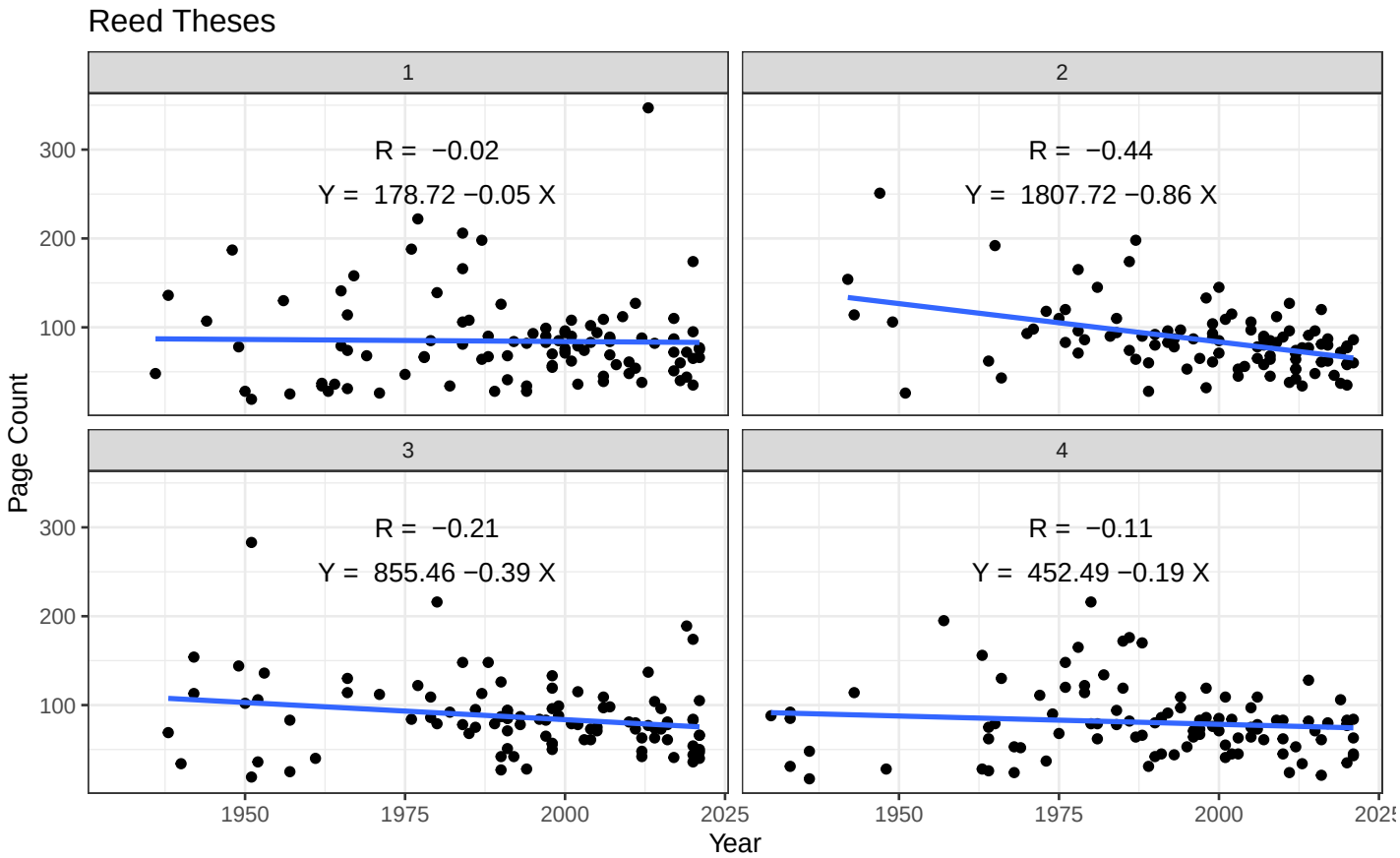
Reed Thesis, More Samples

- Here are several more samples:



Reed Thesis, More Samples

- Here are several more samples:



- By how much will regression statistics (slope, intercept, standard deviation, correlation) change, just due to random sampling?

Section 2

Hypothesis Tests

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: Year X and Page Count Y are uncorrelated
- **Alternative Hypothesis**: Page Count and Year are negatively correlated

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

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: Year X and Page Count Y are uncorrelated
- **Alternative Hypothesis**: Page Count and Year are negatively correlated

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

Method

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

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: Year X and Page Count Y are uncorrelated
- **Alternative Hypothesis**: Page Count and Year are negatively correlated

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

Method

- If there is no relationship, then the pairing between X and Y is artificial and we can shuffle the values of Y among the values of X to produce a similar data set:
  - For each thesis, record the year of publications, but randomly choose a page count from among all recorded page counts (without replacement)

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: Year X and Page Count Y are uncorrelated
- **Alternative Hypothesis**: Page Count and Year are negatively correlated

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

Method

- If there is no relationship, then the pairing between X and Y is artificial and we can shuffle the values of Y among the values of X to produce a similar data set:
  - For each thesis, record the year of publications, but randomly choose a page count from among all recorded page counts (without replacement)
  - Compute the slope of the regression model for this synthetic data set

Hypothesis Tests for Regression

Hypotheses

- **Null Hypothesis**: Year X and Page Count Y are uncorrelated
- **Alternative Hypothesis**: Page Count and Year are negatively correlated

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

Method

- If there is no relationship, then the pairing between X and Y is artificial and we can shuffle the values of Y among the values of X to produce a similar data set:
  - For each thesis, record the year of publications, but randomly choose a page count from among all recorded page counts (without replacement)
  - Compute the slope of the regression model for this synthetic data set
  - Repeat several times to assess variability in slope assuming H_0 is true

A Few Shuffles

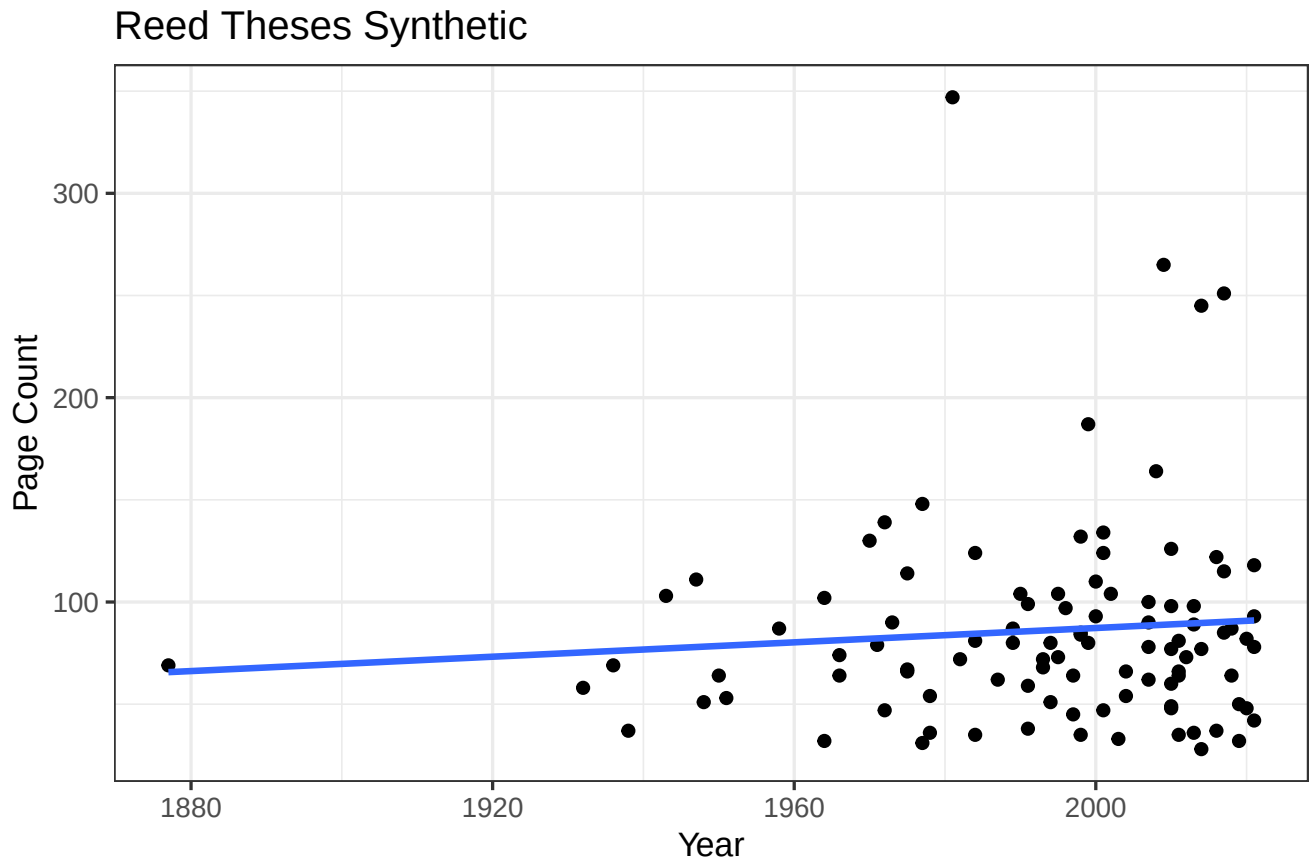
```
theses_samp %>%
  specify(n_pages~year) %>%
  hypothesize(null = "independence") %>%
  generate(1, type = "permute")
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1      48  2020         1
## 2      54  1978         1
## 3     124  2001         1
## 4      36  2013         1
## 5     124  1984         1
## 6      90  2007         1
```

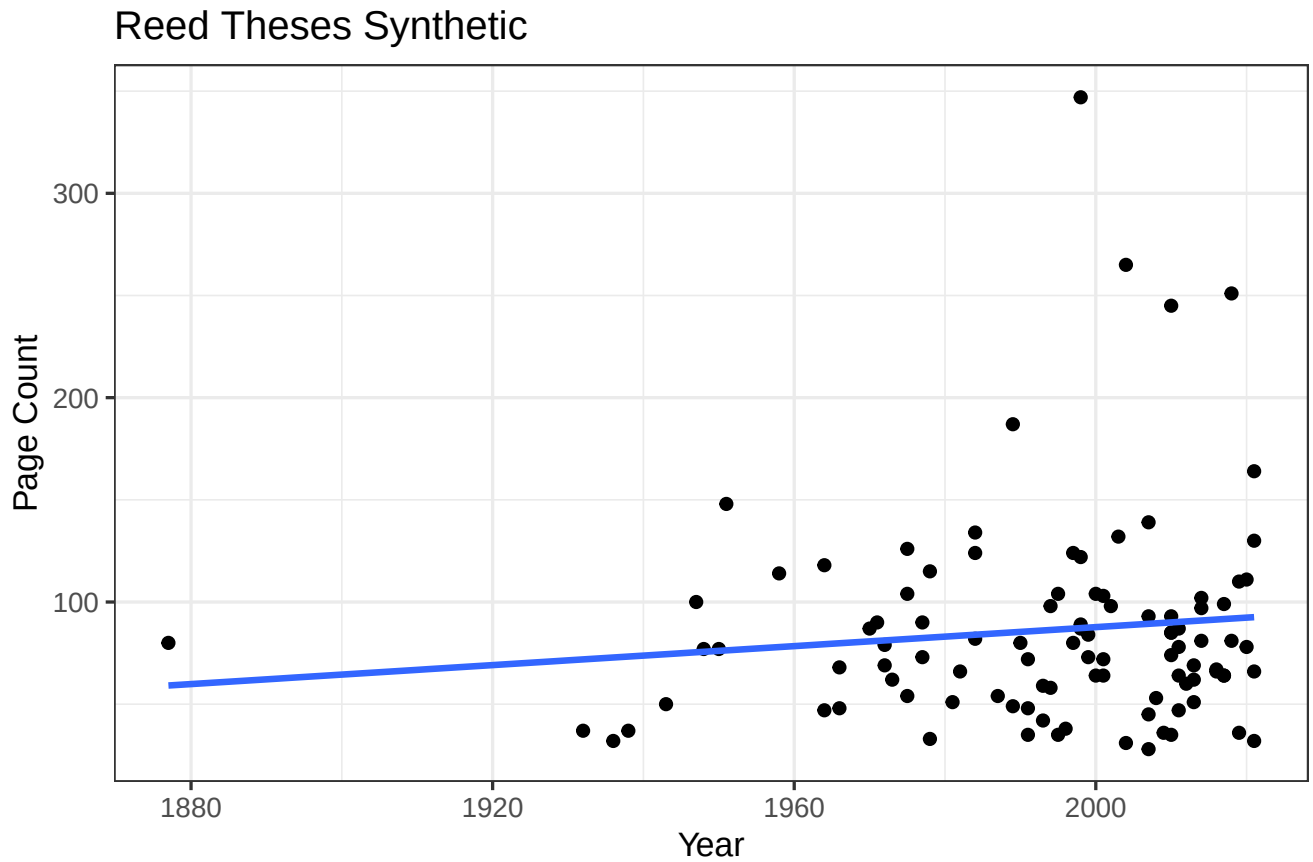
```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1      78  2020         1
## 2     115  1978         1
## 3      64  2001         1
## 4      51  2013         1
## 5      82  1984         1
## 6      45  2007         1
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1      36  2020         1
## 2      38  1978         1
## 3      87  2001         1
## 4      32  2013         1
## 5      45  1984         1
## 6      97  2007         1
```

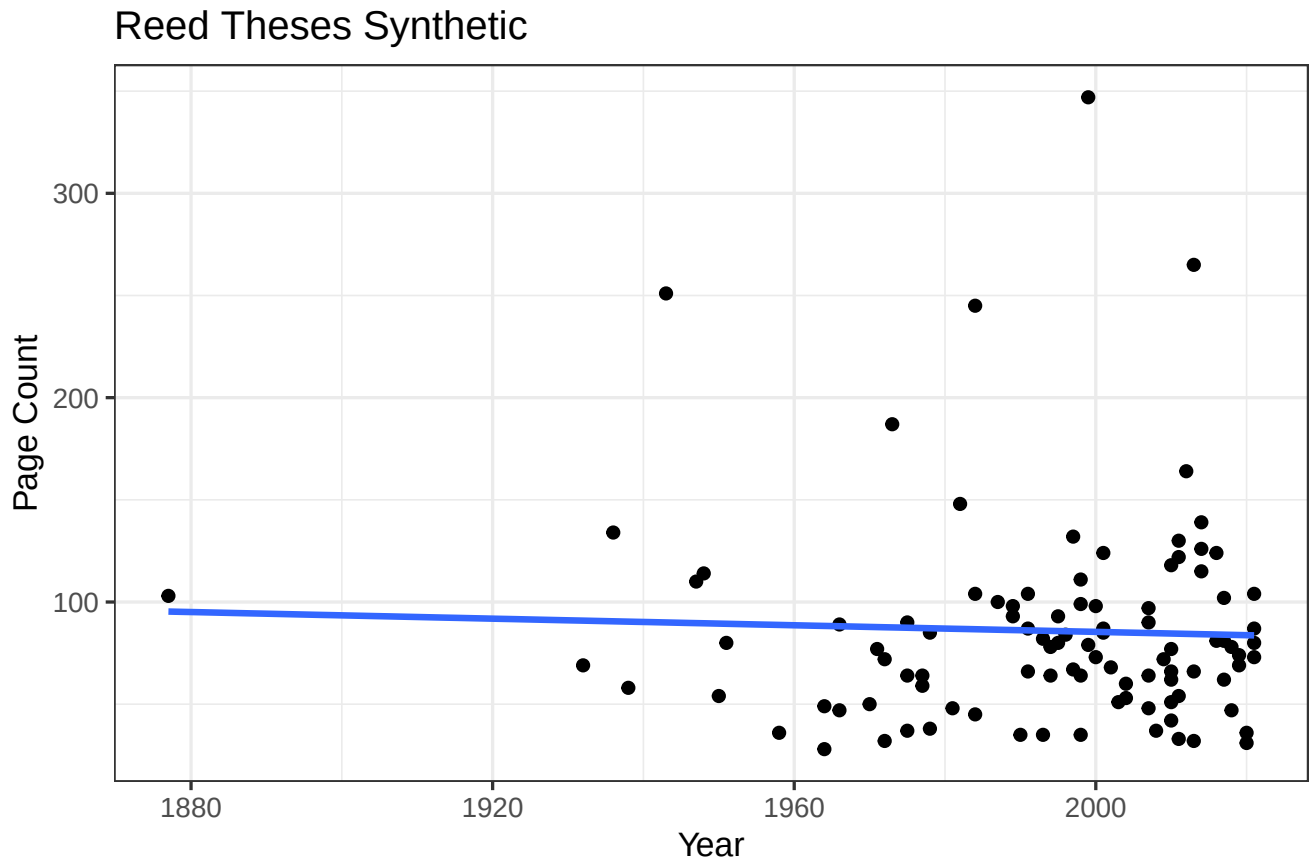
Scatterplots of Synthetic Data I



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

```
theses_samp %>%
  specify(n_pages~year) %>%
  hypothesize(null = "independence") %>%
  generate(1000, type = "permute") %>%
  calculate( stat = "slope")
```

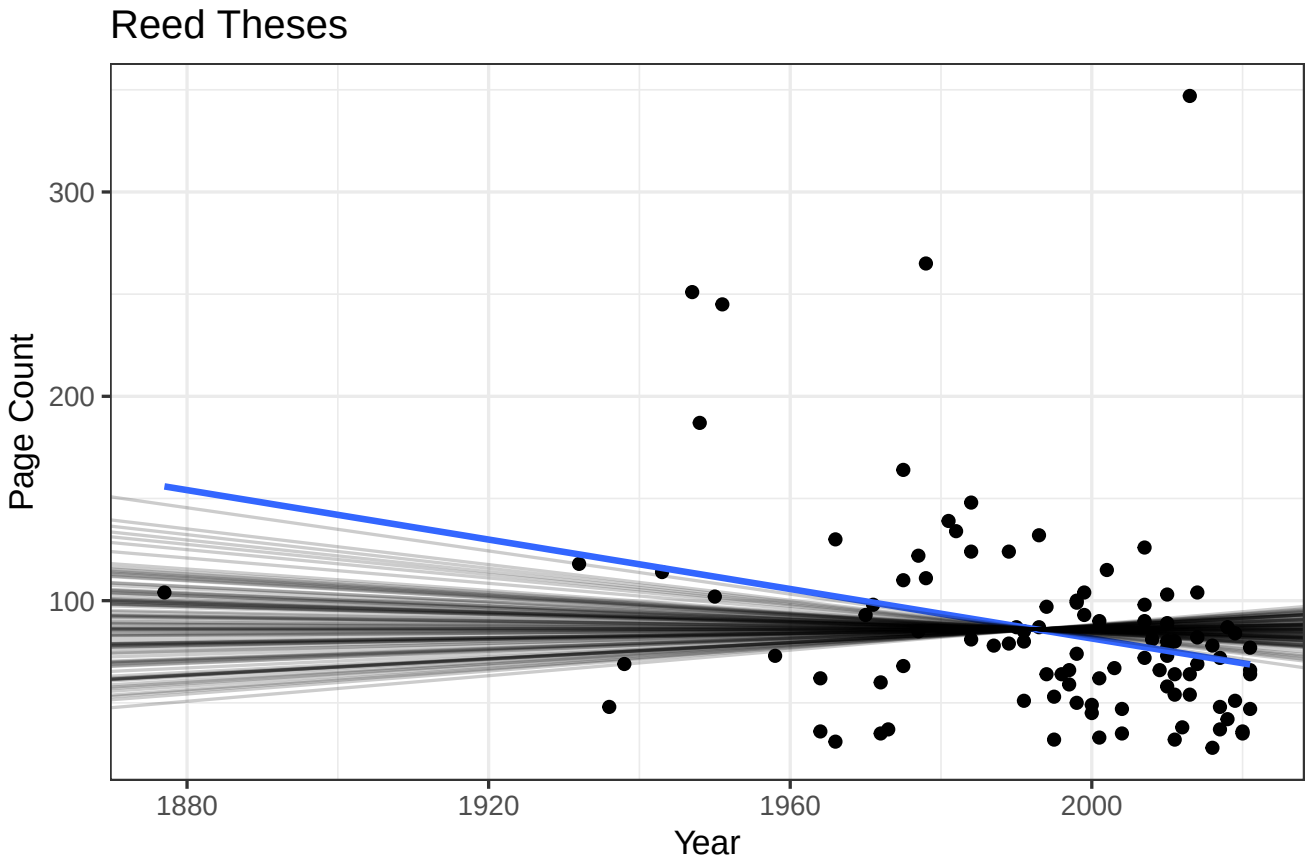
Calculate Statistics

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

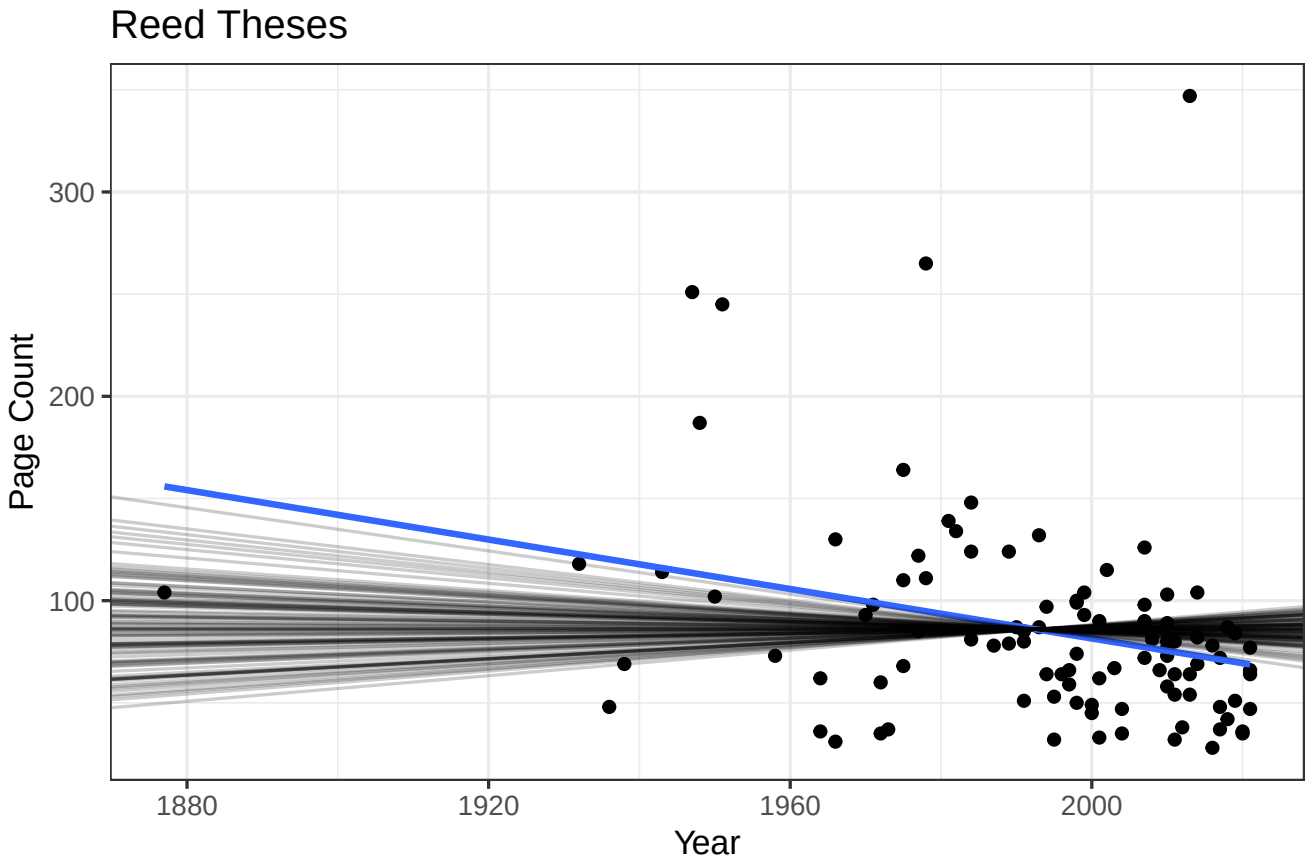
```
theses_samp %>%
  specify(n_pages~year) %>%
  hypothesize(null = "independence") %>%
  generate(1000, type = "permute") %>%
  calculate( stat = "slope")
```

```
## Response: n_pages (numeric)
## Explanatory: year (numeric)
## Null Hypothesis: independence
## # A tibble: 6 x 2
##   replicate     stat
##       <int>    <dbl>
## 1         1 -0.225
## 2         2  0.262
## 3         3 -0.219
## 4         4  0.00218
## 5         5 -0.00447
## 6         6 -0.146
```

Visualizing 1000 Slopes

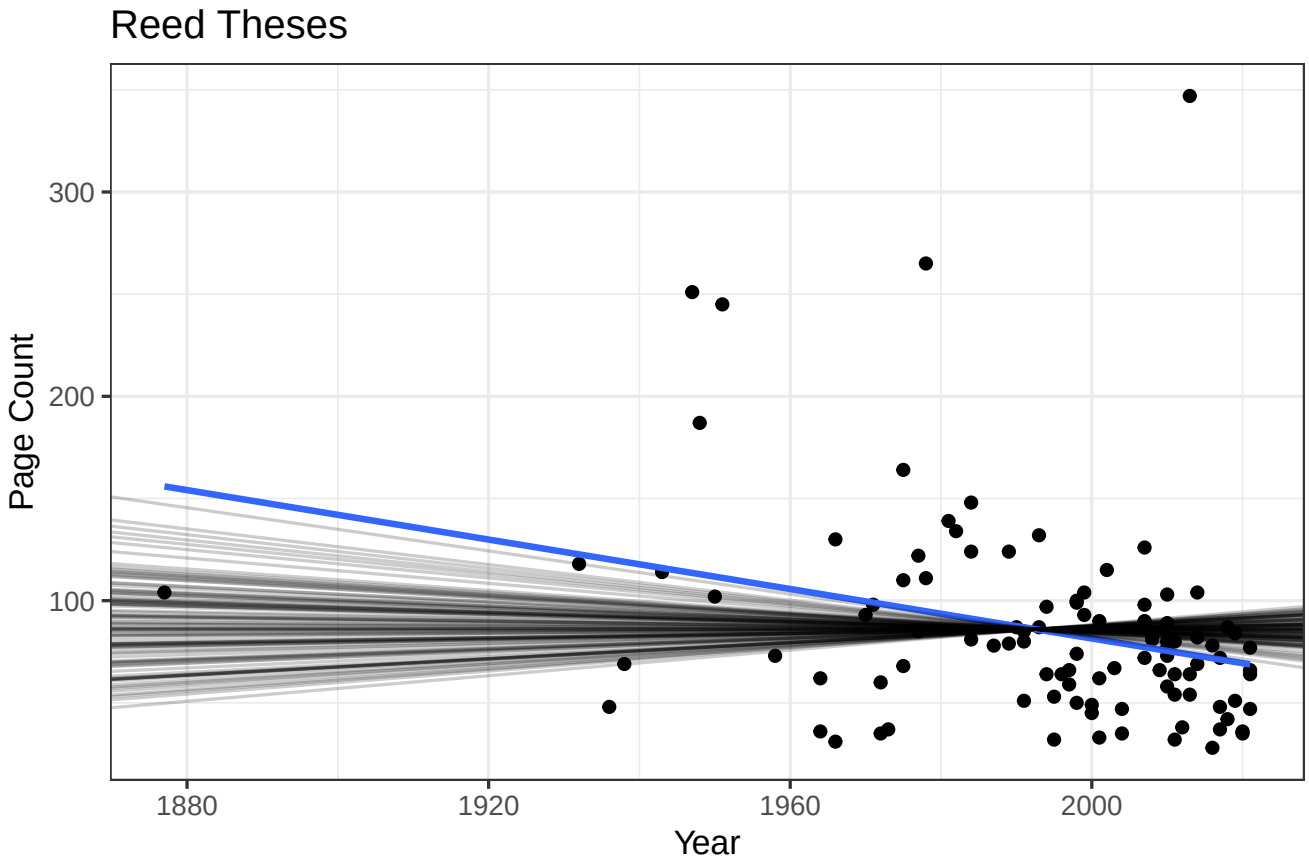


Visualizing 1000 Slopes



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

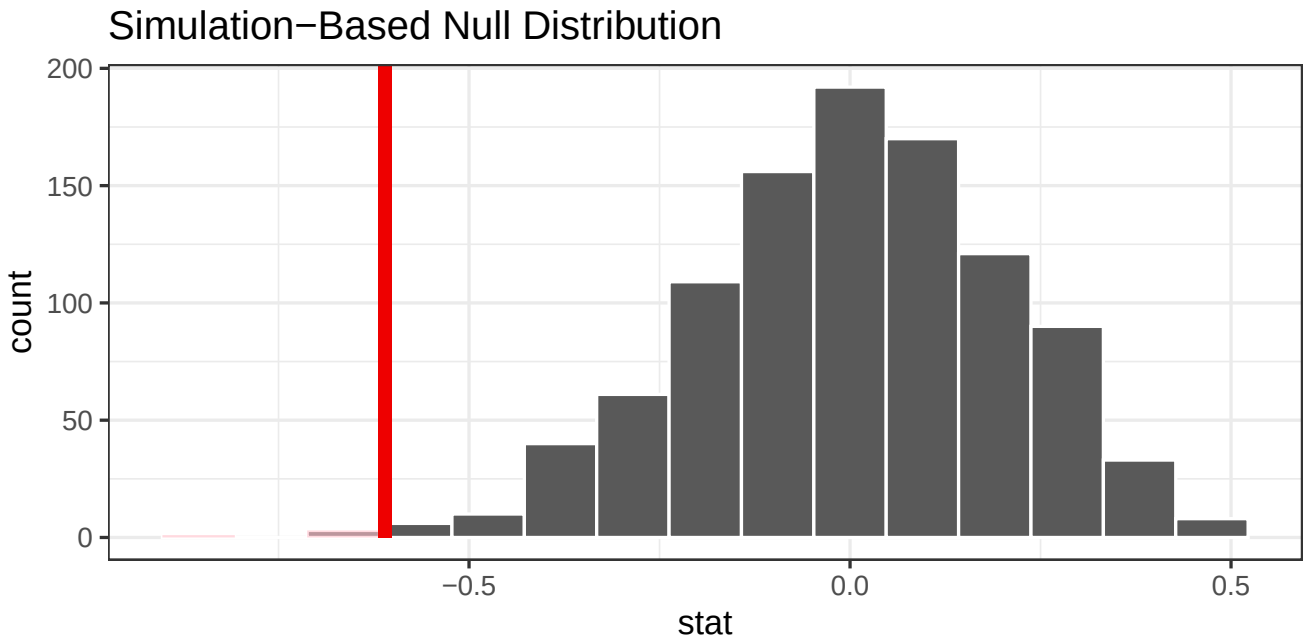
Visualizing 1000 Slopes



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

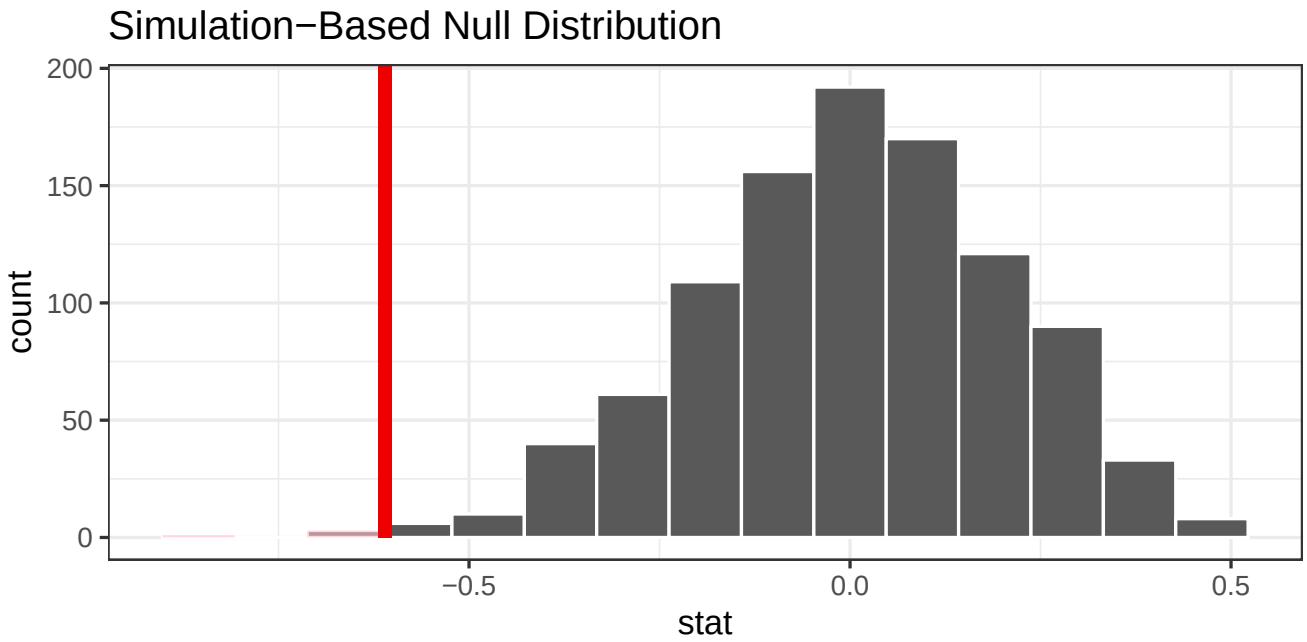
The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.61, direction = "left")
```



The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.61, direction = "left")
```




```
null_slope %>% get_p_value(obs_stat = -0.61, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.004
```

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.
- Is decreased page count **caused** by decreasing standards over time? Very uncertain.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.
- Is decreased page count **caused** by decreasing standards over time? Very uncertain.
  - Perhaps changes in typesetting explain difference.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.
- Is decreased page count **caused** by decreasing standards over time? Very uncertain.
  - Perhaps changes in typesetting explain difference.
  - Perhaps different divisions have different typical lengths of theses, and divisional representation has changed over time.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.
- Is decreased page count **caused** by decreasing standards over time? Very uncertain.
  - Perhaps changes in typesetting explain difference.
  - Perhaps different divisions have different typical lengths of theses, and divisional representation has changed over time.
  - Even if page count has truly decreased on average, page count doesn't necessarily indicate quality or standards.

Conclusion

With a P-value less than $\alpha = 0.01$, we reject H_0 in favor of H_a.

- A slope like this is unlikely to have arisen due to chance if there were no relationship between Year and Page Count.
- The data does indeed suggest Page Count and Year are negatively correlated.
- Is decreased page count **caused** by decreasing standards over time? Very uncertain.
  - Perhaps changes in typesetting explain difference.
  - Perhaps different divisions have different typical lengths of theses, and divisional representation has changed over time.
  - Even if page count has truly decreased on average, page count doesn't necessarily indicate quality or standards.
  - Perhaps conditions for inference were not met!

Section 3

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:

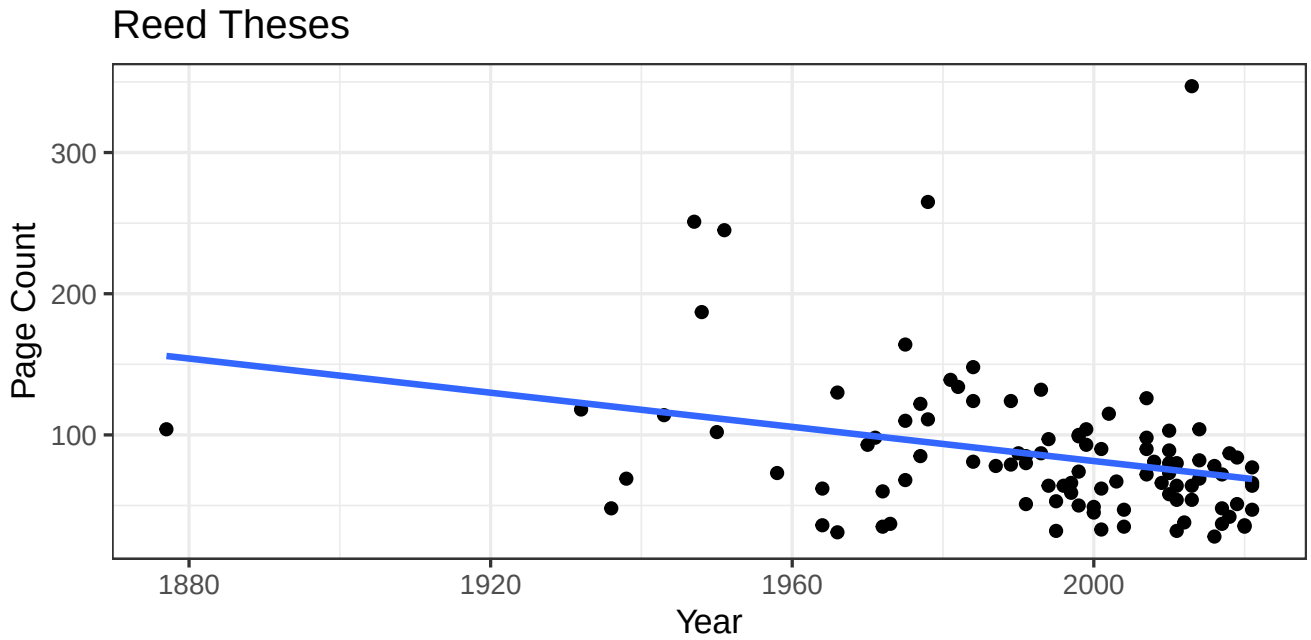
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
   - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
   - Check using histogram of residuals

Conditions for Inference: LINE!

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

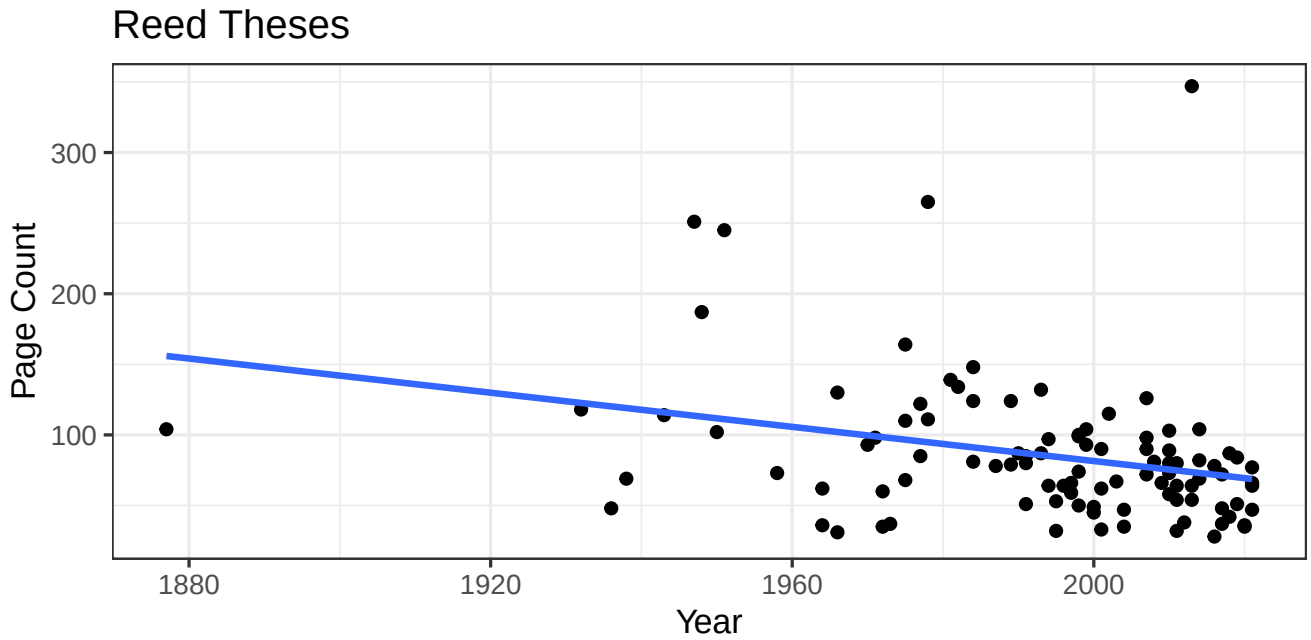
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
   - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
   - Check using scatterplot/residual plot, as well as sample design
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**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

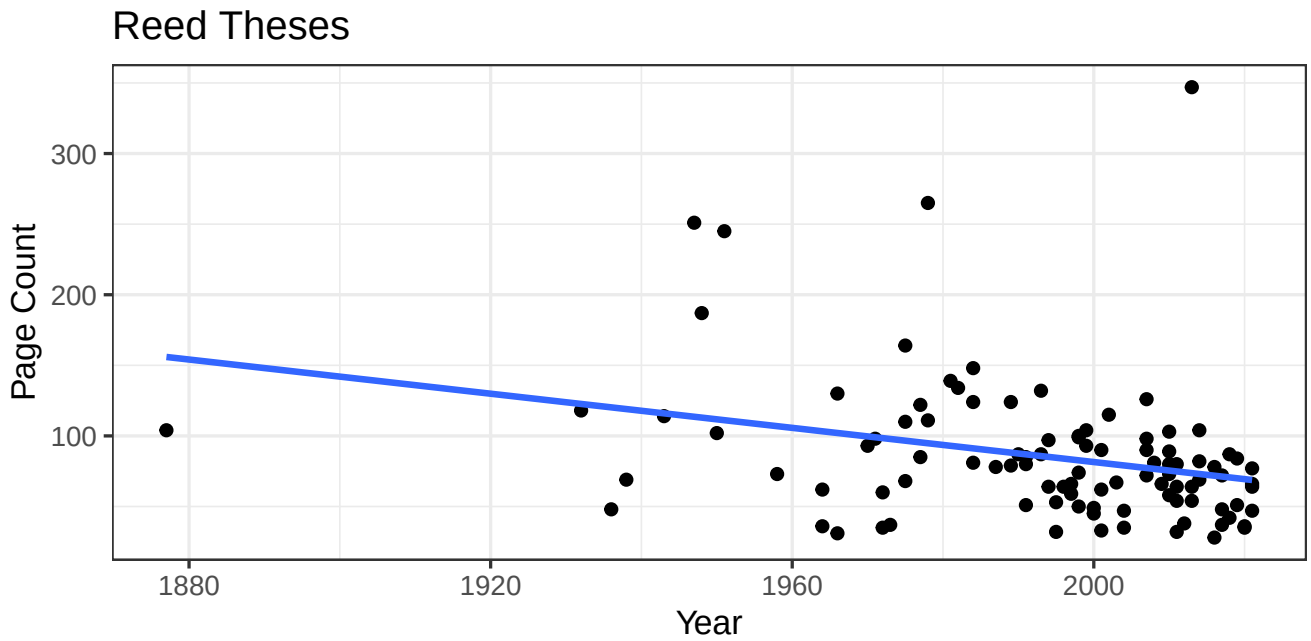
Checking Conditions: Linear



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

```
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.295
```

Checking Conditions: Independence

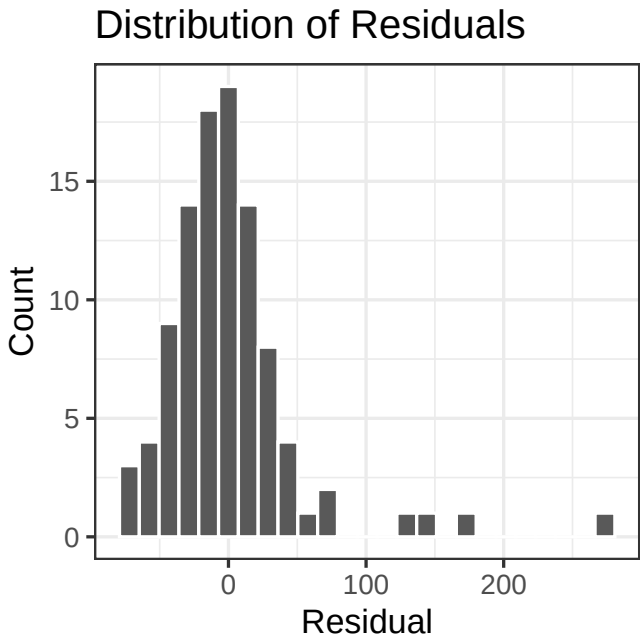
- When students were tasked with sampling theses, they were asked to consider whether their method represented an SRS. Here are some methods used:

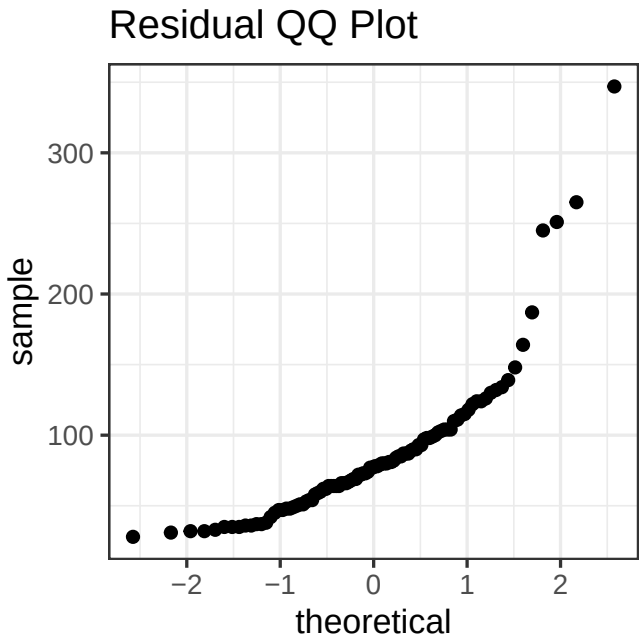
Checking Conditions: Independence

- When students were tasked with sampling theses, they were asked to consider whether their method represented an SRS. Here are some methods used:

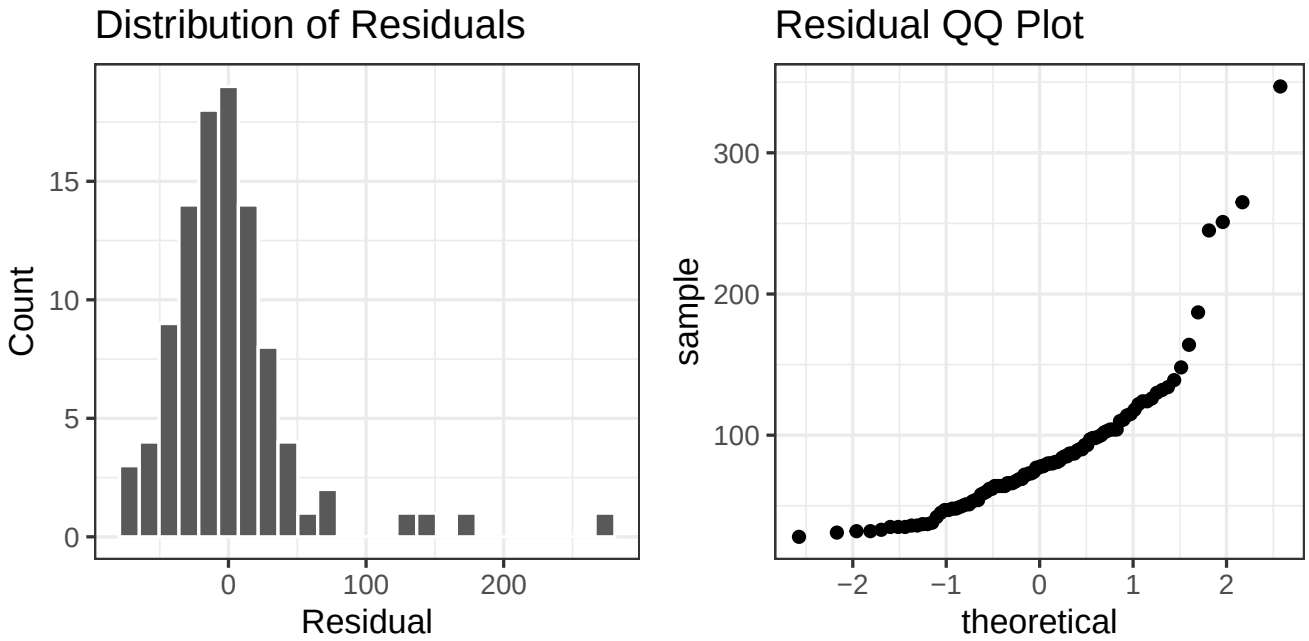
1. Sort theses in the online library catalog by year published and title. Generate 10 random numbers between 1 and 16159, and use these to select theses from catalog.
2. Use the library database with no order specified. Randomly generate a letter of the alphabet and pick the first thesis in the list whose title included the letter.
3. Generate 3 random letters of the alphabet, and choose 10 theses whose author's last name begins with the given letter.
4. Divide the thesis tower into 6 sections of approx. equal size. Randomly choose 1 section using 6-sided die. Then randomly choose a shelf in this section, followed by a row, and then a thesis on the row (using appropriately sized dice)

Checking Conditions: Normal



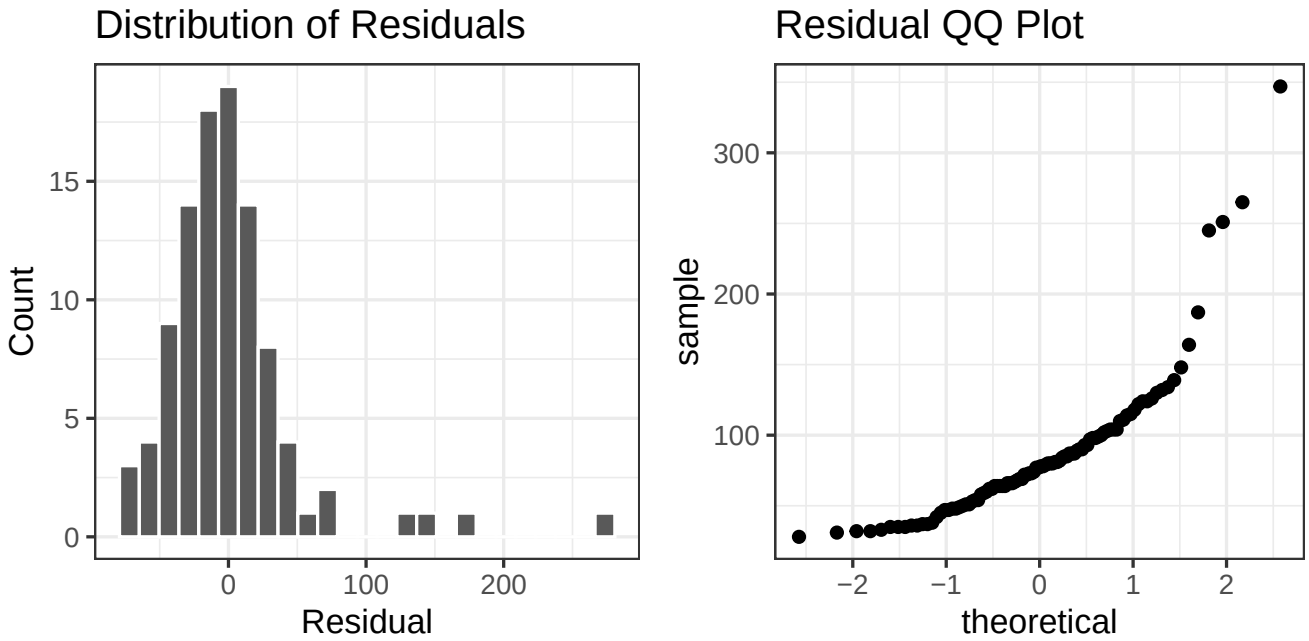


Checking Conditions: Normal



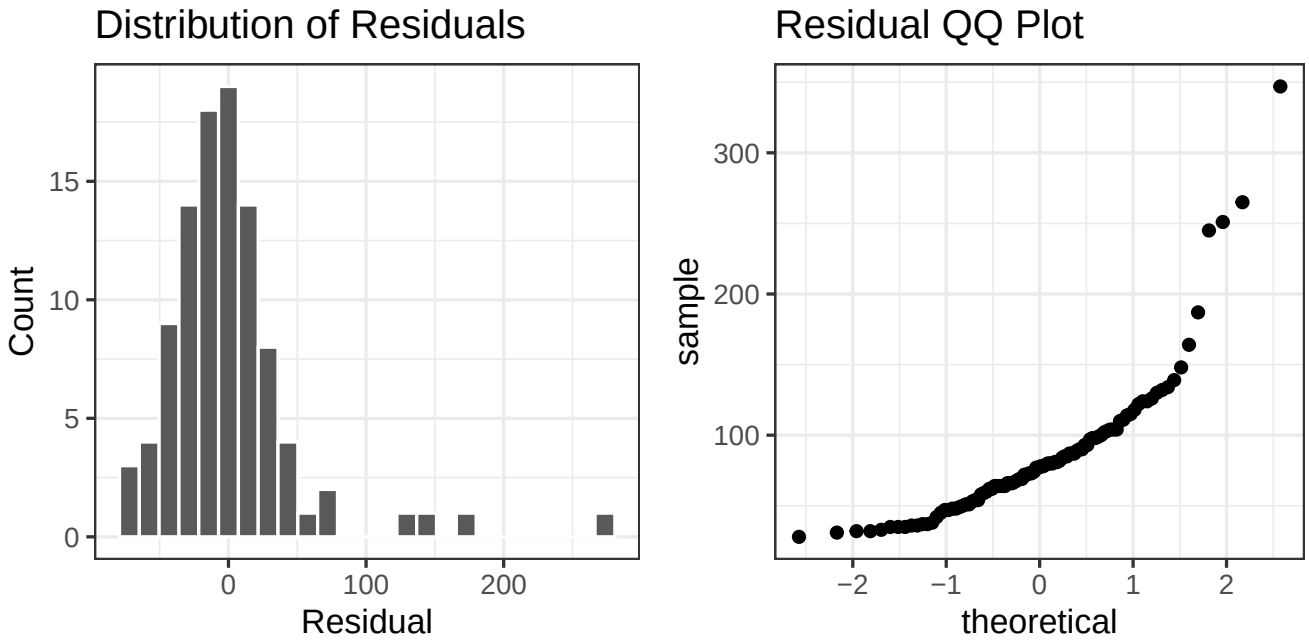
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.

Checking Conditions: Normal



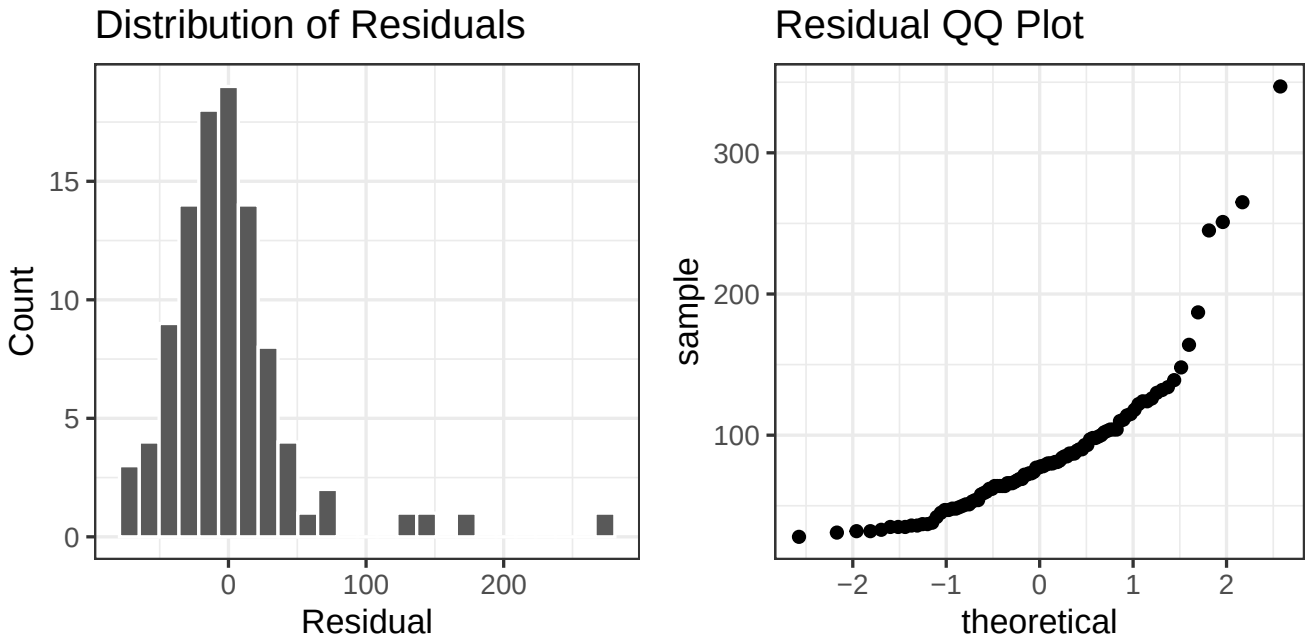
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.

Checking Conditions: Normal



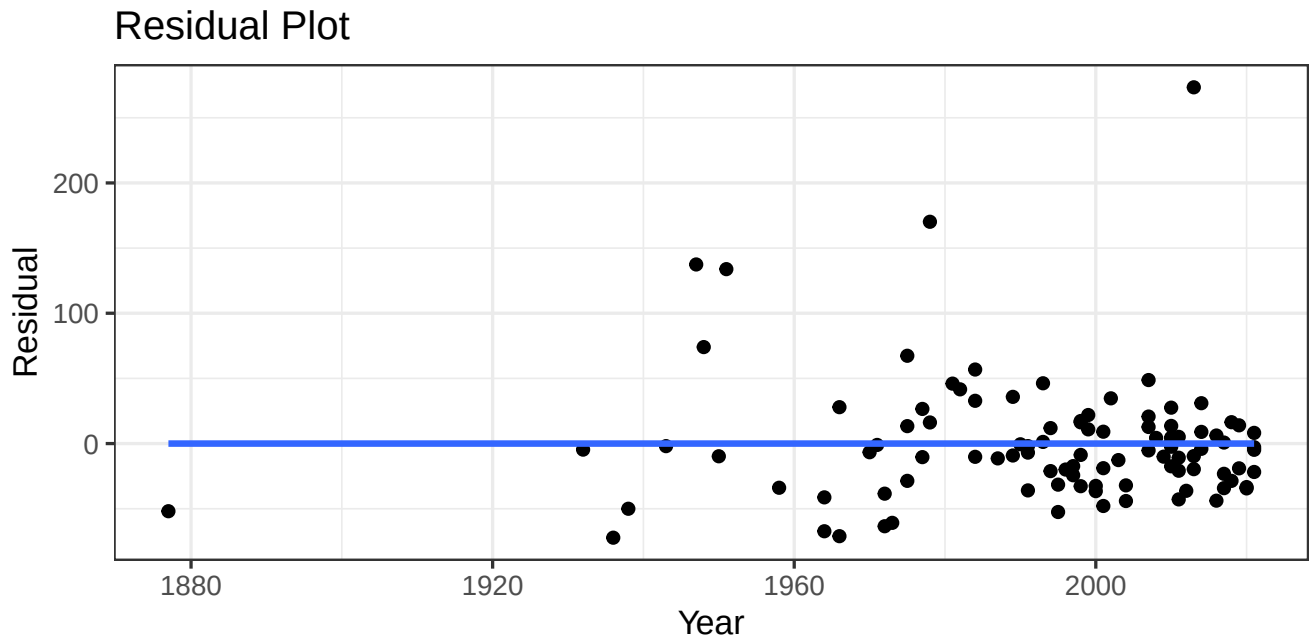
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.
- Do we discard conclusions entirely?

Checking Conditions: Normal



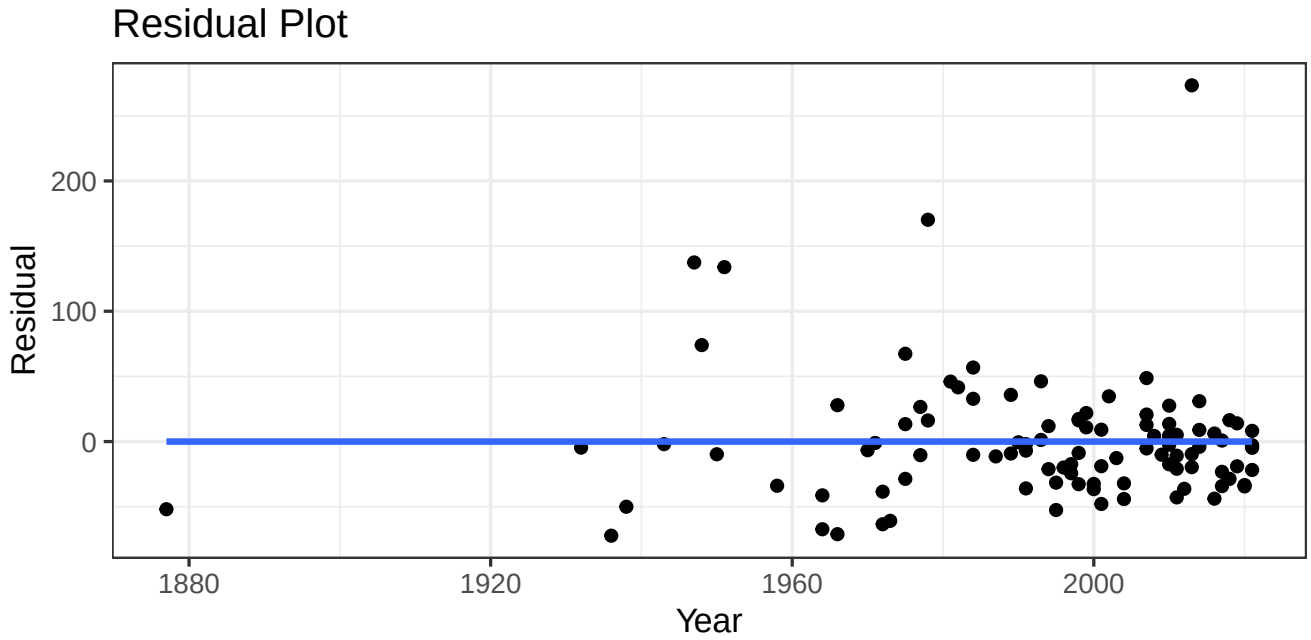
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.
- Do we discard conclusions entirely?
  - No. But this does warrant further research.

Checking Conditions: Equal Variability



Residuals appear to have constant variability between 1975 and 2020

Checking Conditions: Equal Variability



Residuals appear to have constant variability between 1975 and 2020

- However, theses prior to 1975 appear to have more spread (and almost all outliers come from this region of sparser data)

Section 4

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 page count can be perfectly predicted by year (with no deviations or errors). What slope would we expect to find in the 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 page count can be perfectly predicted by year (with no deviations or errors). What slope would we expect to find in the regression model?
  - It's hard to say without knowing the variability in the year and in the page count data.
  - Remember that 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 page count can be perfectly predicted by year (with no deviations or errors). What slope would we expect to find in the regression model?
  - It's hard to say without knowing the variability in the year and in the page count data.
  - Remember that 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 variablity 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 variablity 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

```
theses_samp %>%
  specify(n_pages~year) %>%
  generate(1, type = "bootstrap")
```

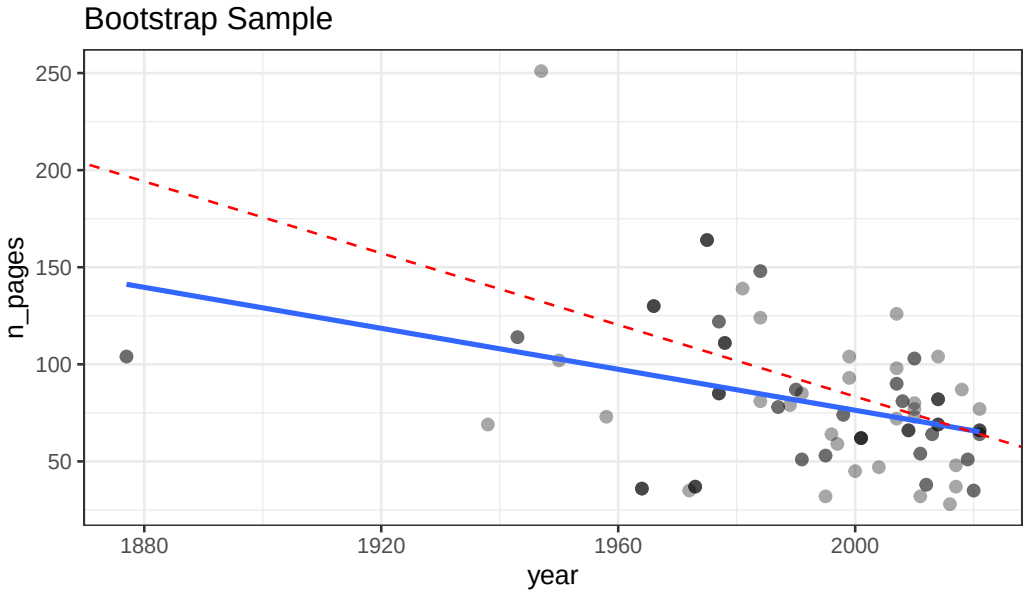
```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   replicate n_pages  year
##       <int>   <dbl> <dbl>
## 1         1      51  1991
## 2         1      78  1987
## 3         1     103  2010
## 4         1      81  2008
## 5         1      36  1964
## 6         1      37  1973
```

Bootstrapping for confidence intervals

- To approximate variablity 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

```
theses_samp %>%
  specify(n_pages~year) %>%
  generate(1, type = "bootstrap")
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   replicate n_pages  year
##       <int>   <dbl> <dbl>
## 1         1      51  1991
## 2         1      78  1987
## 3         1     103  2010
## 4         1      81  2008
## 5         1      36  1964
## 6         1      37  1973
## # A tibble: 1 x 2
##   replicate    cor
##       <int>  <dbl>
## 1         1 -0.382
```



- 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

```
theses_samp %>%
  specify(n_pages~year) %>%
  generate(1000, type = "bootstrap") %>%
  calculate(stat = "correlation")
```

Bootstrap Distribution for correlation

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

```
theses_samp %>%
  specify(n_pages~year) %>%
  generate(1000, type = "bootstrap") %>%
  calculate(stat = "correlation")
```

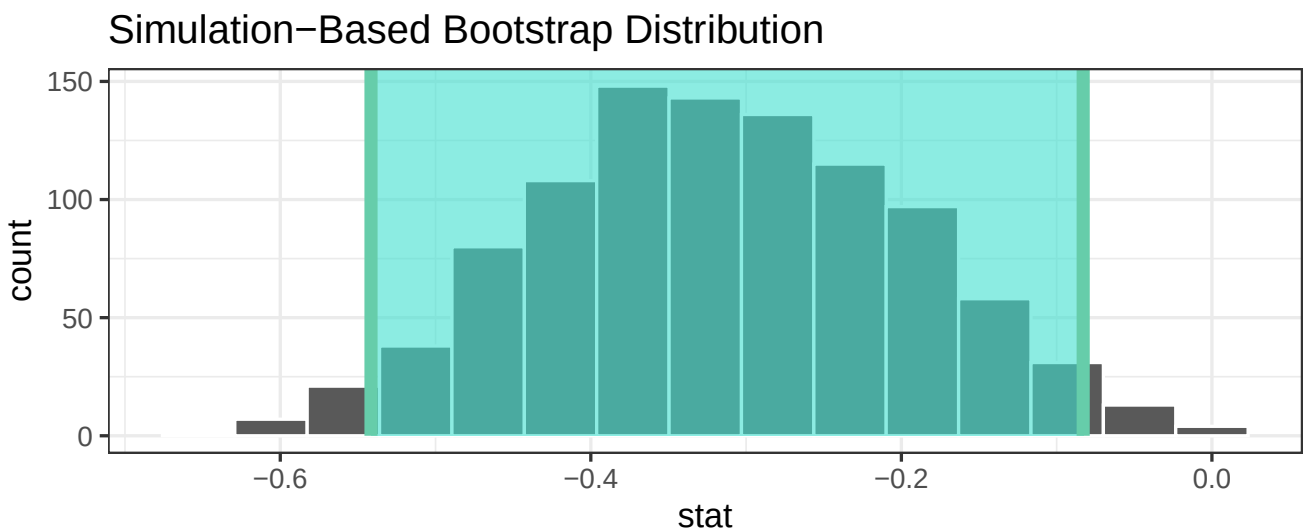
```
## Response: n_pages (numeric)
## Explanatory: year (numeric)
## # A tibble: 6 x 2
##   replicate    stat
##       <int>   <dbl>
## 1         1 -0.294
## 2         2 -0.242
## 3         3 -0.235
## 4         4 -0.0830
## 5         5 -0.268
## 6         6 -0.407
```

The Bootstrap Distribution for *R*

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.542  -0.0829
```

```
boot_slope %>% visualize()+shade_ci(endpoints =correlation_ci)
```

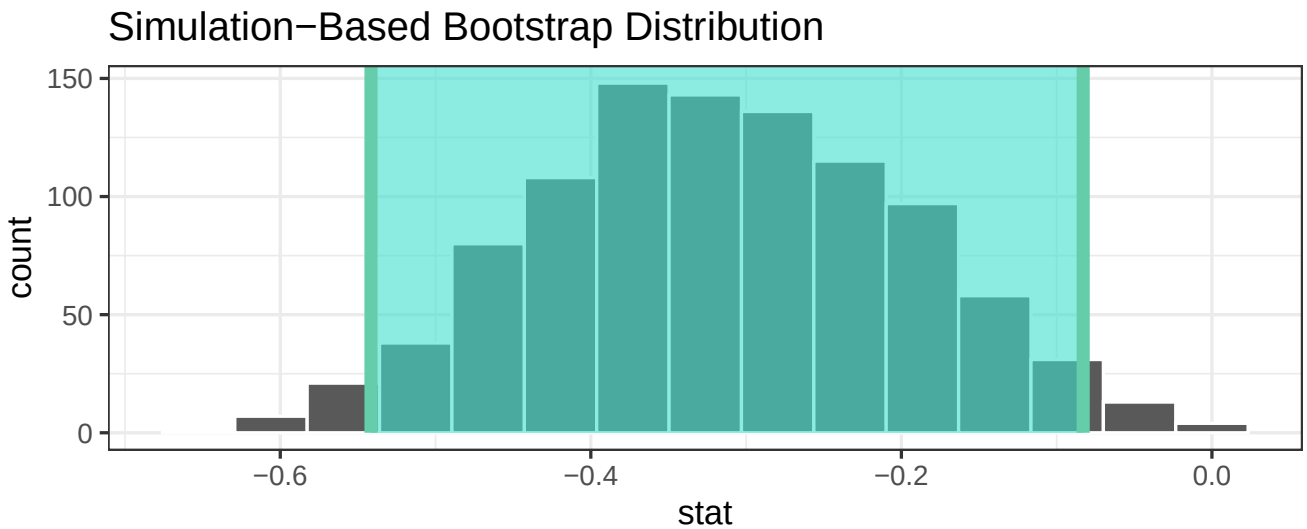


The Bootstrap Distribution for R

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.542  -0.0829
```

```
boot_slope %>% visualize()+shade_ci(endpoints =correlation_ci)
```



- The original sample had correlation R = -0.3
  - It is possible the true relationship between page count and year has between very weak (-0.08) and moderate (-0.54) negative correlation.