Inference for Regression

Nate Wells

Math 141, 4/26/21

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

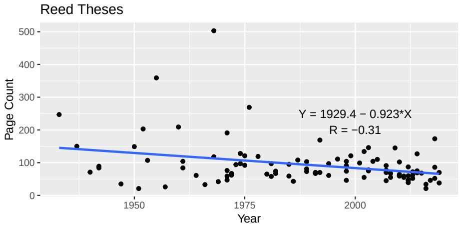
- Previously, we used linear regression to analyze the relationship between two quantitative variables
    - The strength and direction of the linear relationship is summarized by the correlation coefficient R

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.

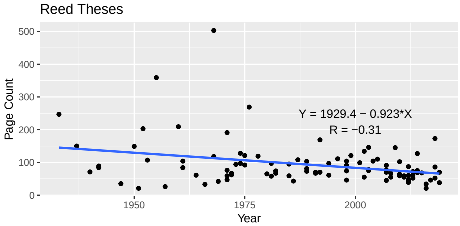
Review of Simple Linear Regression

- 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.



Review of Simple Linear Regression

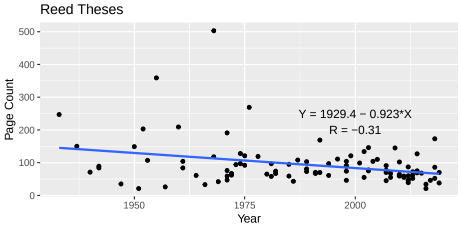
- Previously, we used linear regression to analyze the relationship between two quantitative variables
  - The strength and direction of the linear relationship is summarized by the correlation coefficient R
  - The linear model $\hat{Y} = \beta_0 + \beta_1 X$ can be used to make predictions about Y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



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

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.



- 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 *statisics* calculated from data to make inferences about the nature of *parameters*

Statistical Inference for Regression

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

For regression:

- **Parameters**: β_0, β_1
- **Statistics**: b_0, b_1

Statistical Inference for Regression

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

For regression:

- **Parameters**: β_0, β_1
- **Statistics**: b_0, b_1

Classic tools of inference:

- **Confidence Intervals** to estimate values
- **Hypothesis Tests** to assess claims about values

Statistical Inference for Regression

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

For regression:

- **Parameters**: β_0, β_1
- **Statistics**: b_0, b_1

Classic tools of inference:

- **Confidence Intervals** to estimate values
- **Hypothesis Tests** to assess claims about values

Our sample data represents a shadow of the true population.

Statistical Inference for Regression

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

For regression:

- **Parameters**: β_0, β_1
- **Statistics**: b_0, b_1

Classic tools of inference:

- **Confidence Intervals** to estimate values
- **Hypothesis Tests** to assess claims about values

Our sample data represents a shadow of the true population.

- Like shadows, certain features may accentuated or compressed compared to the genuine article

Statistical Inference for Regression

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

For regression:

- **Parameters**: β_0, β_1
- **Statistics**: b_0, b_1

Classic tools of inference:

- **Confidence Intervals** to estimate values
- **Hypothesis Tests** to assess claims about values

Our sample data represents a shadow of the true population.

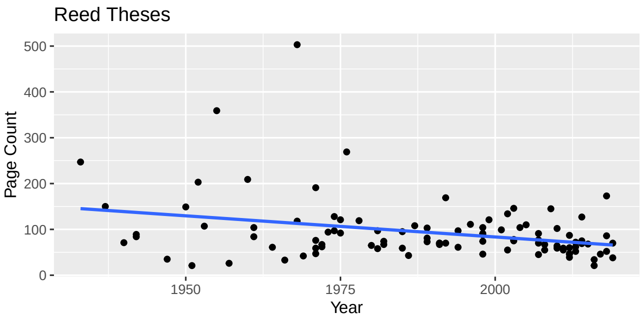
- Like shadows, certain features may accentuated or compressed compared to the genuine article
- But we can analyze how much features could change by creating model replicas and comparing the shadows of the replicas to the replica itself

Theses Page Counts

“Old Reed” Theory: Thesis page counts have decreased over time due to relaxed standards.

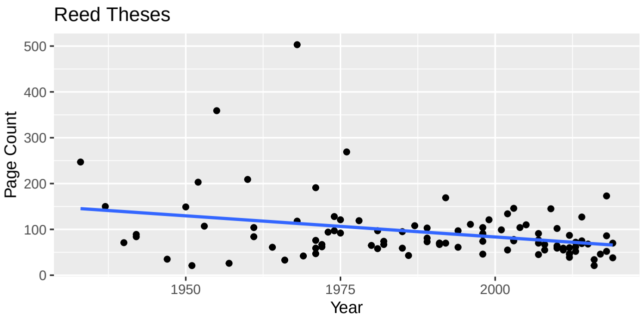
Theses Page Counts

"Old Reed" Theory: Thesis page counts have decreased over time due to relaxed standards.



Theses Page Counts

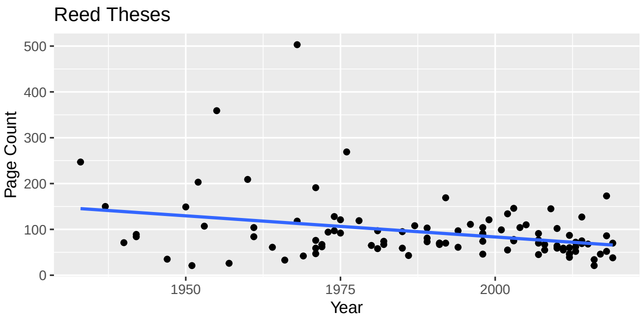
"Old Reed" Theory: Thesis page counts have decreased over time due to relaxed standards.



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

Theses Page Counts

"Old Reed" Theory: Thesis page counts have decreased over time due to relaxed standards.

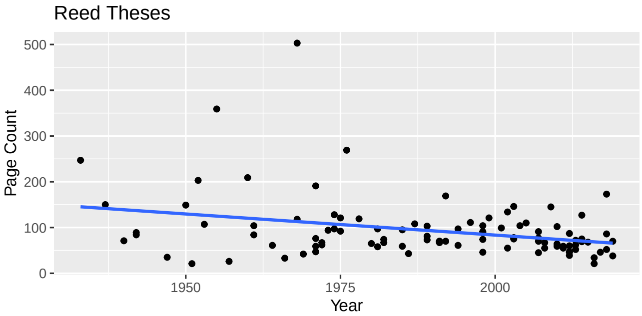


But this is just a sample of data. Would a different sample produce a different regression line?

- Almost certainly!

Theses Page Counts

"Old Reed" Theory: Thesis page counts have decreased over time due to relaxed standards.



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

- Almost certainly!
- We'll investigate how much it could change by

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 amongst 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 amongst 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 amongst 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 amongst 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 amongst 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 amongst 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 amongst 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")
```

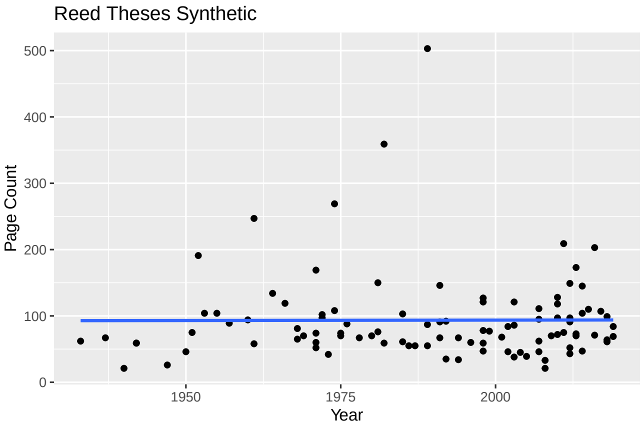
```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1     103  1985         1
## 2      46  2007         1
## 3     128  2010         1
## 4      74  1975         1
## 5      88  1976         1
## 6     127  1998         1
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1      67  1985         1
## 2     150  2007         1
## 3     269  2010         1
## 4      65  1975         1
## 5      74  1976         1
## 6      61  1998         1
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   n_pages  year replicate
##     <dbl> <dbl>     <int>
## 1     191  1985         1
## 2      59  2007         1
## 3      59  2010         1
## 4     104  1975         1
## 5      64  1976         1
## 6      55  1998         1
```

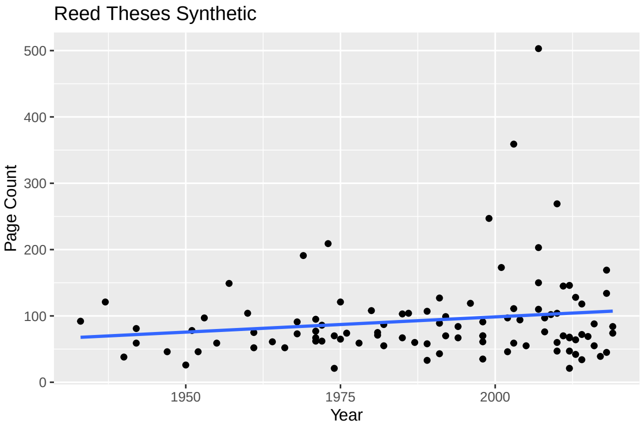
Scatterplots of Synthetic Data I

```
samp1 %>% ggplot( aes( x = year, y = n_pages)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Reed Theses Synthetic", x = "Year", y = "Page Count")
```



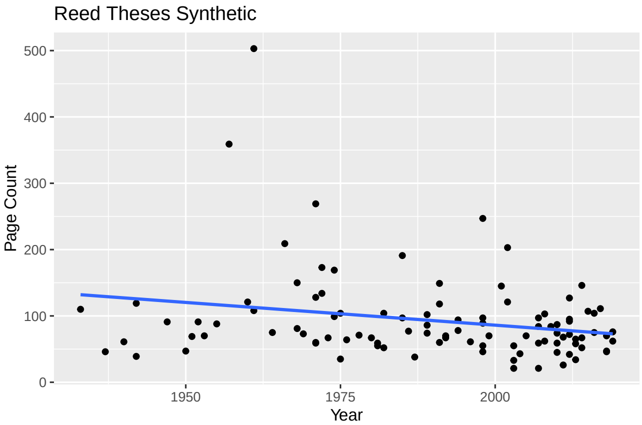
Scatterplots of Synthetic Data II

```
samp2 %>% ggplot( aes( x = year, y = n_pages)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Reed Theses Synthetic", x = "Year", y = "Page Count")
```



Scatterplots of Synthetic Data III

```
samp3 %>% ggplot( aes( x = year, y = n_pages)) +
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Reed Theses Synthetic", x = "Year", y = "Page Count")
```



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")
```

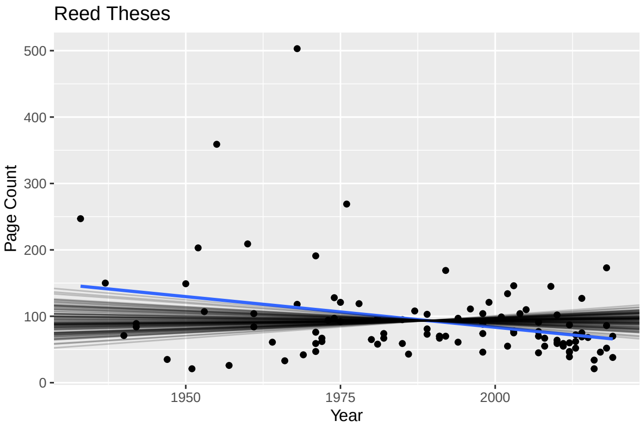
Calculate Statistics

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

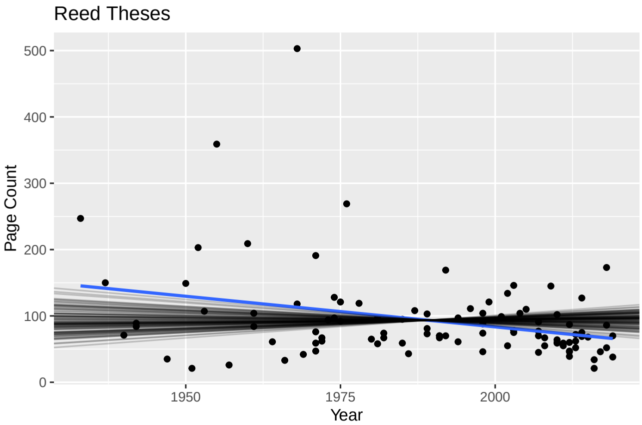
```
theses_samp %>%
  specify(n_pages~year) %>%
  hypothesize(null = "independence") %>%
  generate(1000, type = "permute") %>%
  calculate( stat = "slope")
```

```
## # A tibble: 6 x 2
##   replicate     stat
##       <int>    <dbl>
## 1         1 -0.444
## 2         2 -0.175
## 3         3 -0.405
## 4         4  0.0910
## 5         5 -0.00270
## 6         6  0.211
```

Visualizing 1000 Slopes

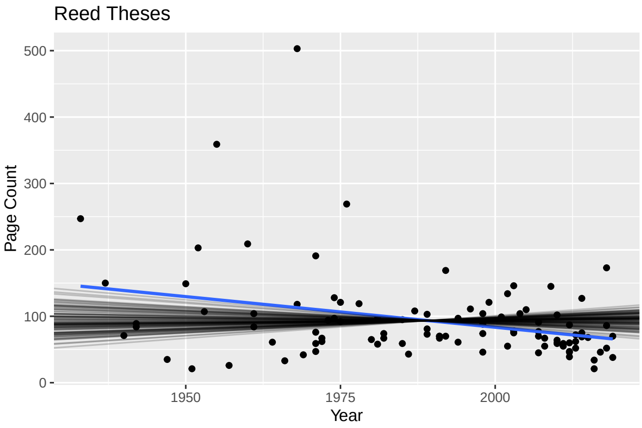


Visualizing 1000 Slopes



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

Visualizing 1000 Slopes

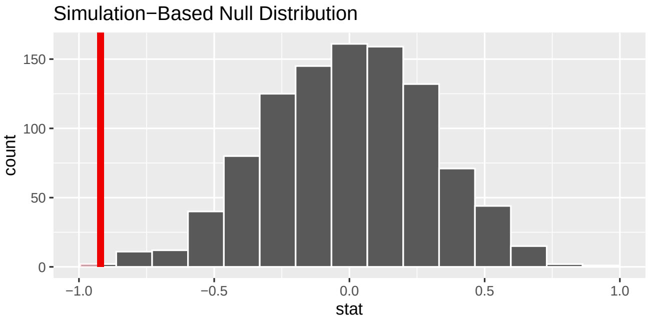


Most lines are approximately horizontal. But some have positive or negative slope.

The linear regression line for the original data is shown in blue.

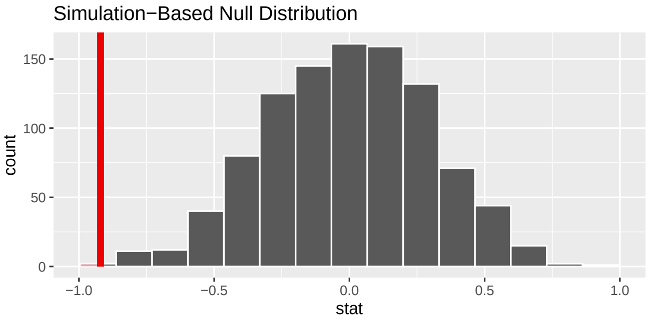
The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.92, direction = "left")
```



The Sampling Distribution of b_1

```
null_slope %>% visualize()+shade_p_value(obs_stat = -0.92, direction = "left")
```




```
null_slope %>% get_p_value(obs_stat = -0.92, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

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 slople 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 slople 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 slople 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 slople 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 slople 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 slople 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 slople 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

In order to responsibly use linear regression. . .

Conditions for Inference

In order to responsibly use linear regression. . .

1. Relationship between variables must be approximately linear. (**Linear**)
    - Check using scatterplot and/or residual plot

Conditions for Inference

In order to responsibly use linear regression...

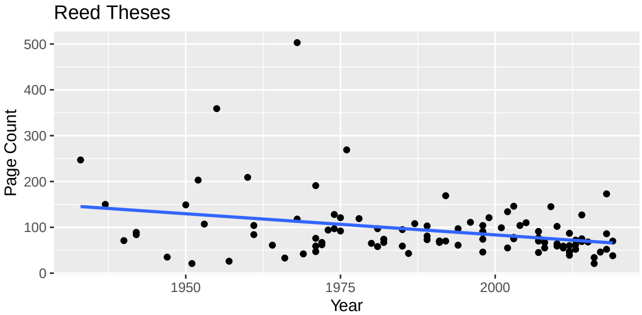
1. Relationship between variables must be approximately linear. (**Linear**)
    - Check using scatterplot and/or residual plot
2. The variability of residuals should be roughly constant across entire data set. (**Homoscedastic**)
    - Check using resdidual plot.

Conditions for Inference

In order to responsibly use linear regression. . .

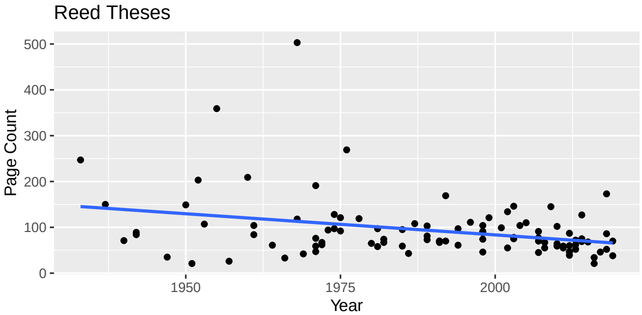
1. Relationship between variables must be approximately linear. (**Linear**)
    - Check using scatterplot and/or residual plot
2. The variability of residuals should be roughly constant across entire data set. (**Homoscedastic**)
    - Check using resdidual plot.
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
    - Check using histogram of residuals

Checking Conditions: Linear



Data is not tightly clustered around line of best fit

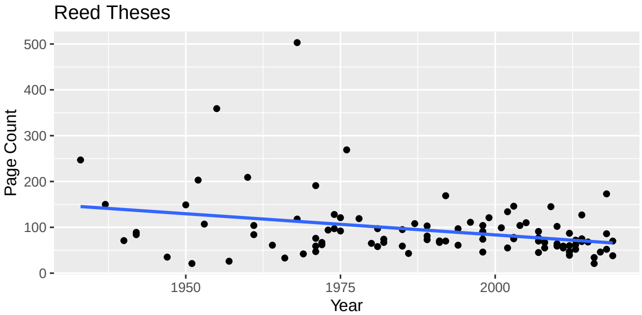
Checking Conditions: Linear



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



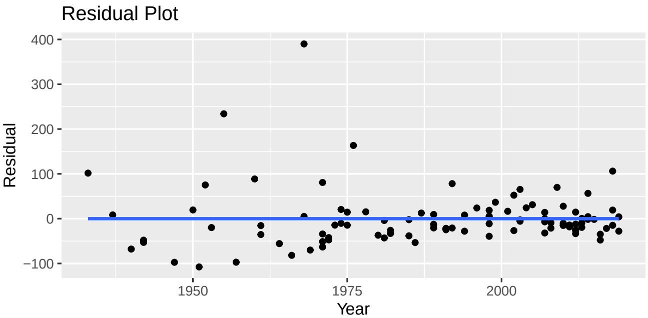
Data is not tightly clustered around line of best fit

- But this doesn't mean data is not linear. Just that residuals have high variance

```
get_correlation(data = theses_samp, n_pages ~ year)
```

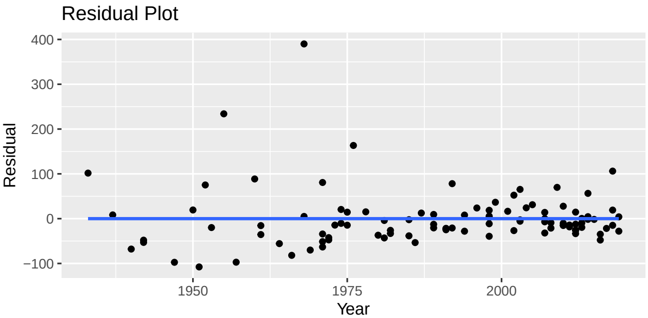
```
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.315
```

Checking Conditions: Homoscedastic



Residuals appear to have constant varaibility between 1975 and 2020

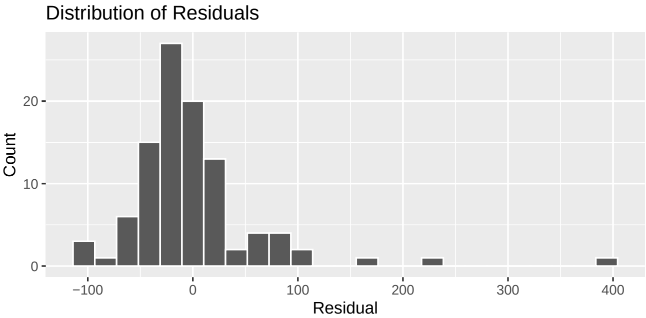
Checking Conditions: Homoscedastic



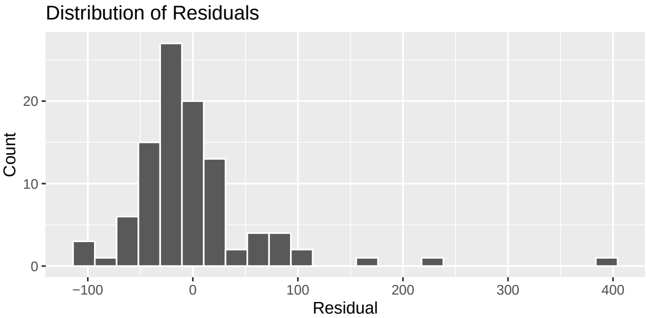
Residuals appear to have constant varaibility between 1975 and 2020

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

Checking Conditions: Normal

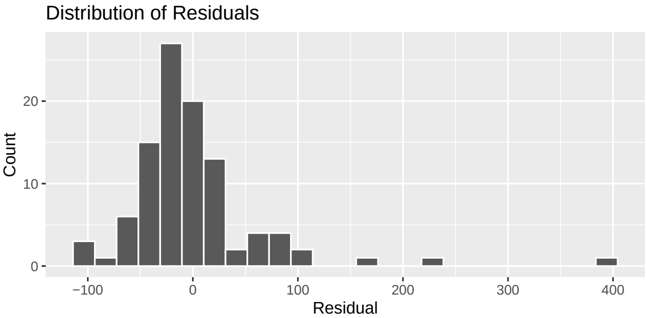


Checking Conditions: Normal



The distribution does appear to have moderate right skew, with a notable outlier

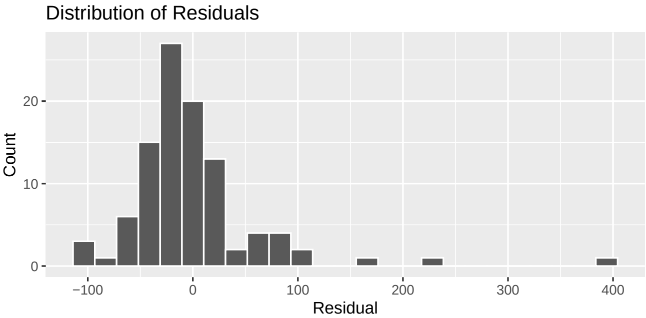
Checking Conditions: Normal



The distribution does appear to have moderate right skew, with a notable outlier

- This is relatively concerning. We should treat the results of inference with caution.

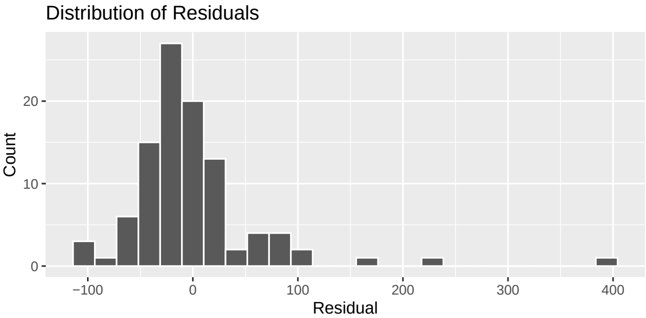
Checking Conditions: Normal



The distribution does appear to have moderate right skew, with a notable outlier

- This is relatively concerning. We should treat the results of inference with caution.
- Do we discard conclusions entirely?

Checking Conditions: Normal



The distribution does appear to have moderate right skew, with a notable outlier

- This is relatively concerning. We should treat the results of inference with caution.
- Do we discard conclusions entirely?
  - No. But this does warrant further research.

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     111  1996
## 2         1      71  1940
## 3         1      67  2008
## 4         1      97  1974
## 5         1      84  1961
## 6         1     173  2018
```

Bootstrapping for confidence intervals

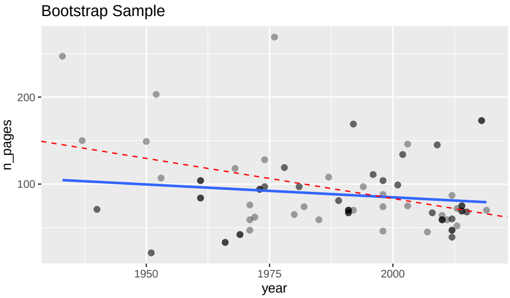
- To approximate variablity in the correlation statistic R, we create a bootstrap sample by resampling the paired data and then calculation correlation
  - This corresponds to sampling with replacement from the columns of the original sample

```
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     111  1996
## 2         1      71  1940
## 3         1      67  2008
## 4         1      97  1974
## 5         1      84  1961
## 6         1     173  2018
```

```
get_correlation(samp1, n_pages~year)
```

```
## # A tibble: 1 x 2
##   replicate    cor
##       <int>  <dbl>
## 1         1 -0.148
```



- 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")
```

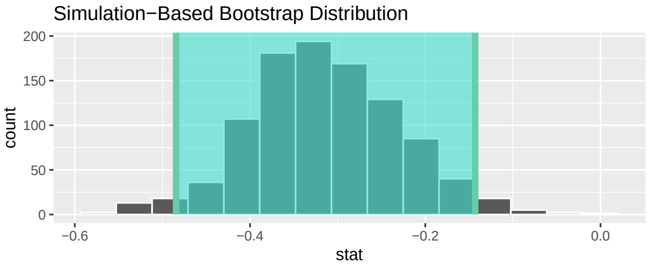
```
## # A tibble: 6 x 2
##   replicate   stat
##       <int>  <dbl>
## 1         1 -0.197
## 2         2 -0.395
## 3         3 -0.195
## 4         4 -0.379
## 5         5 -0.227
## 6         6 -0.268
```

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   -0.484   -0.143
```

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

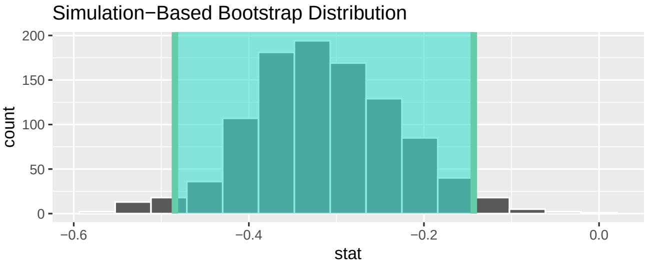


The Bootstrap Distribution for R

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

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.484   -0.143
```

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



- The original sample had correlation $R = -0.315$
  - It is possible the true relationship between page count and year has between very weak (-0.13) and moderate (-0.48) negative correlation.