Inference for Multiple Linear Regression

Nate Wells

Math 141, 4/20/22

Outline

In this lecture, we will. . .

- Review conditions for inference in simple linear models
- Create confidence intervals for parameters of linear models
- Discuss theory-based methods for regression
- Review framework for multilinear regression
- Discuss inference procedures for MLR models

Section 1

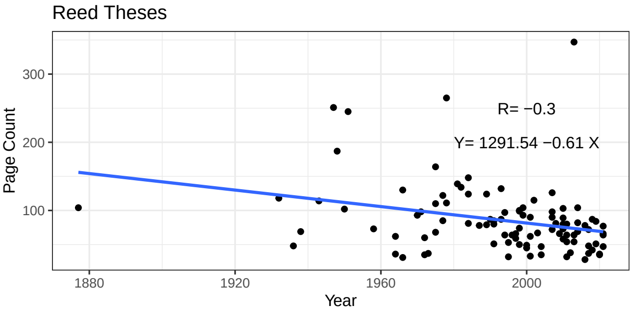
Confidence Intervals

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:



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")
```

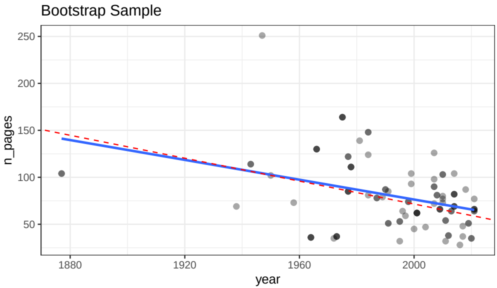
```
## # A tibble: 6 x 3
## # Groups:   replicate [1]
##   replicate n_pages  year
##       <int>   <dbl> <dbl>
## 1         1      51  1991
## 2         1      78  1987
## 3         1     103  2010
## 4         1      81  2008
## 5         1      36  1964
## 6         1      37  1973
```

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")
```

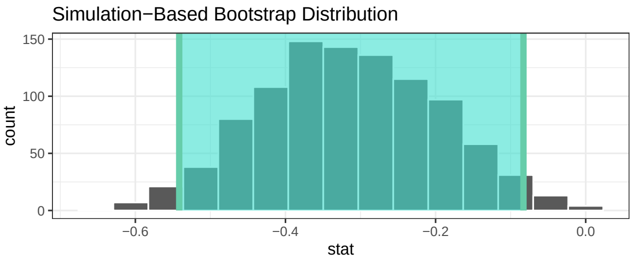
```
## Response: n_pages (numeric)
## Explanatory: year (numeric)
## # A tibble: 6 x 2
##   replicate    stat
##       <int>   <dbl>
## 1         1 -0.294
## 2         2 -0.242
## 3         3 -0.235
## 4         4 -0.0830
## 5         5 -0.268
## 6         6 -0.407
```

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)
```

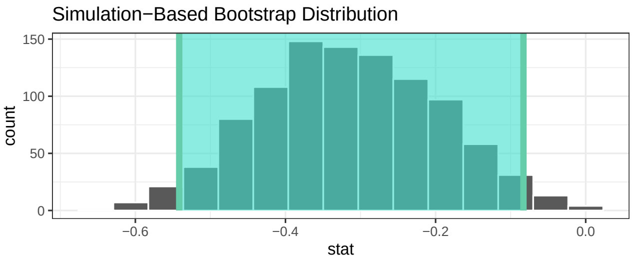


The Bootstrap Distribution for R

```
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.

Section 2

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:

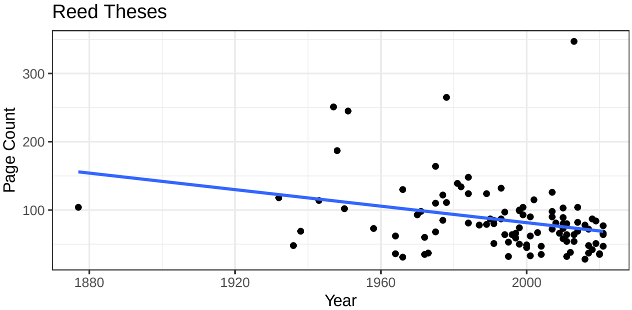
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
    - Check using scatterplot/residual plot, as well as sample design
3. The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

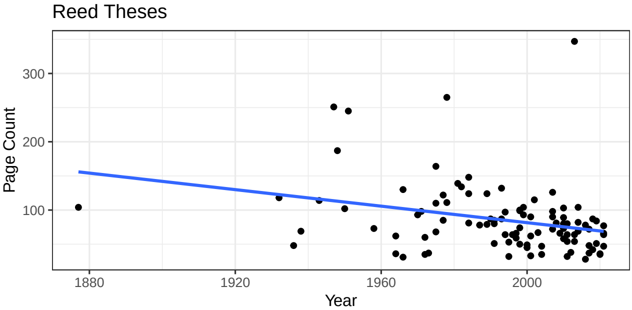
1. The relationship between explanatory and response variables must be approximately linear. (**Linear**)
    - Check using scatterplot/residual plot
2. The observations should be independent of one another. (**Independence**)
    - Check using scatterplot/residual plot, as well as sample design
3. The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

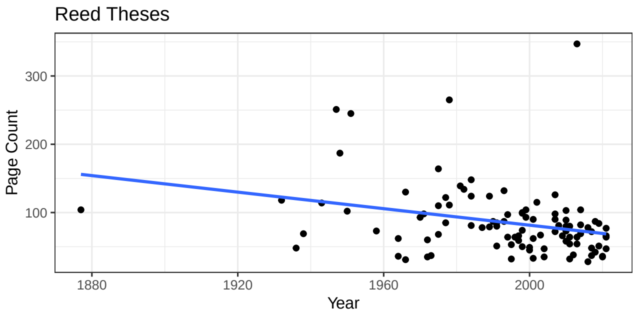
Checking Conditions: Linear



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

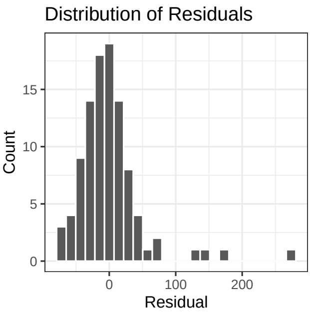
- When students were tasked with sampling theses, they were asked to consider whether their method represented an SRS. Here are some methods used:

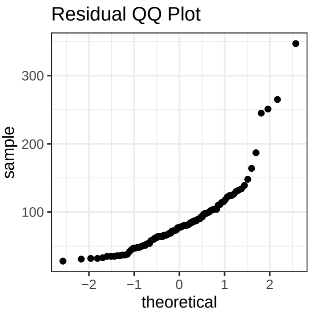
Checking Conditions: Independence

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

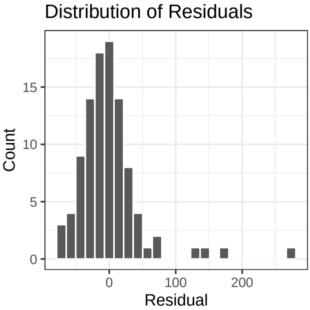
1. Sort theses in the online library catalog by year published and title. Generate 10 random numbers between 1 and 16159, and use these to select theses from catalog.
2. Use the library database with no order specified. Randomly generate a letter of the alphabet and pick the first thesis in the list 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)

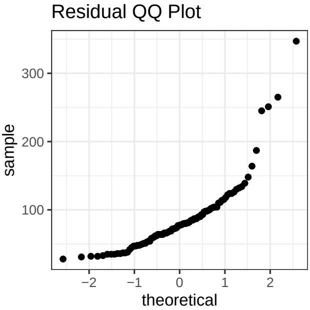
Checking Conditions: Normal





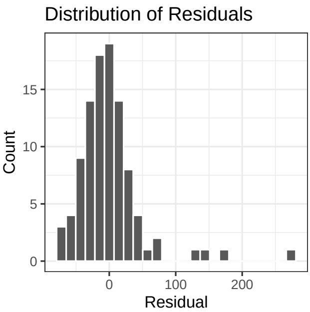
Checking Conditions: Normal

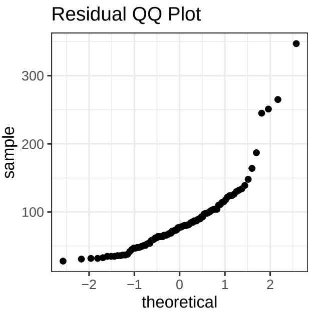




- The distribution does appears somewhat right-skewed, with a notable outliers on the right.

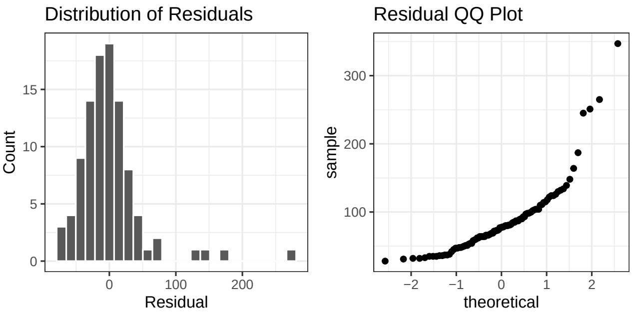
Checking Conditions: Normal





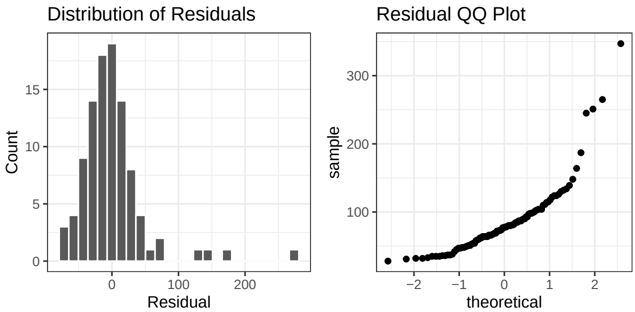
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.

Checking Conditions: Normal



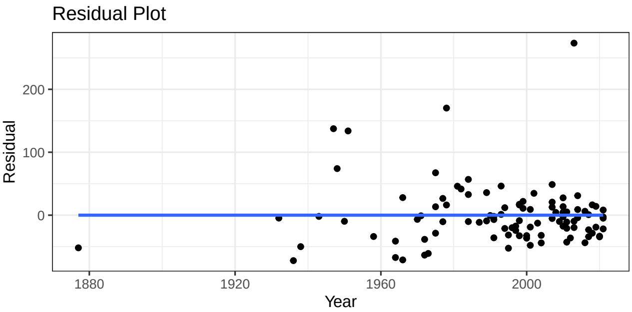
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.
- Do we discard conclusions entirely?

Checking Conditions: Normal



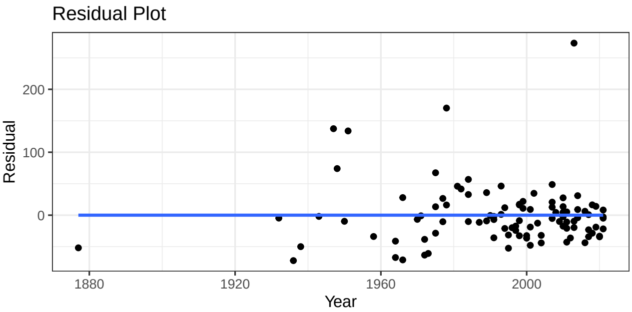
- The distribution does appears somewhat right-skewed, with a notable outliers on the right.
- This provides some evidence residuals are not Normally disributed.
- Do we discard conclusions entirely?
  - No. But this does warrant further research.

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 3

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.

```
thesis_mod <- lm(n_pages ~ year, data = theses_samp)
get_regression_table(theses_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 year        -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

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.

```
thesis_mod <- lm(n_pages ~ year, data = theses_samp)
get_regression_table(theses_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 year        -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

- 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.

```
thesis_mod <- lm(n_pages ~ year, data = theses_samp)
get_regression_table(theses_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 year        -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

- 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.

```
thesis_mod <- lm(n_pages ~ year, data = theses_samp)
get_regression_table(theses_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 year        -0.605     0.198     -3.06   0.003   -0.998   -0.212
```

- 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

Section 4

Multiple Linear Regression

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \dots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \dots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686  -13.3     19.8
## 2 X1           -1.24     0.313     -3.95   0.001   -1.89    -0.584
## 3 X2            2.68     1.94       1.38   0.182   -1.36     6.72
## 4 X3            3.20     0.397      8.06   0        2.37     4.02
```

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686   -13.3    19.8
## 2 X1           -1.24     0.313     -3.95   0.001    -1.89   -0.584
## 3 X2            2.68     1.94       1.38   0.182    -1.36    6.72
## 4 X3            3.20     0.397      8.06   0         2.37    4.02
```

- Which gives us our linear regression formula:

$$\hat{Y} = 3.26 - 1.24 \cdot X_1 + 2.68 \cdot X_2 + 3.2 \cdot X_3$$

Review: Multiple Regression Model

- In a **multiple linear regression model** (MLR), we express the response variable Y as a linear combination of k explanatory variables $X_1, X_2, \ldots, X_k$:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$$

- We use the following R code to fit and summarize a linear model:

```
mod<-lm(Y ~ X1 + X2 + X3, data = my_data)
get_regression_table(mod)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     3.26     7.94       0.41   0.686  -13.3     19.8
## 2 X1           -1.24     0.313     -3.95   0.001   -1.89   -0.584
## 3 X2            2.68     1.94       1.38   0.182   -1.36    6.72
## 4 X3            3.20     0.397      8.06   0        2.37    4.02
```

- Which gives us our linear regression formula:

$$\hat{Y} = 3.26 - 1.24 \cdot X_1 + 2.68 \cdot X_2 + 3.2 \cdot X_3$$

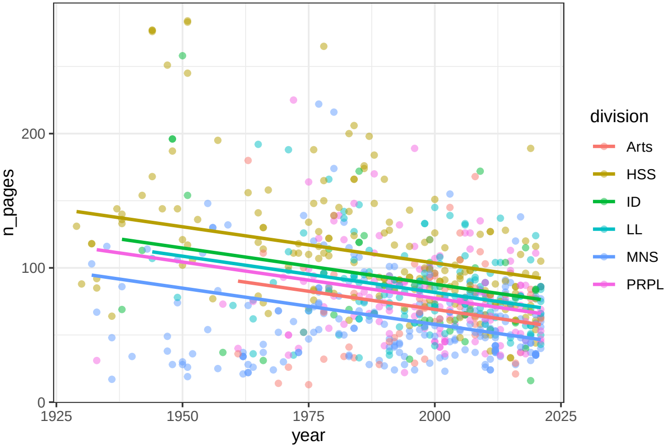
- The slope on each variable indicates the changed in the predicted value of Y per unit change in that variable, **with all other variables held constant**

Reed Theses

- How does page count vary across year of publication **and** division?

Reed Theses

- How does page count vary across year of publication **and** division?



Reed Theses

- How does page count vary across year of publication **and** division?

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0      902.    1396.
## 2 year             -0.54     0.063     -8.58   0       -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0       23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011    4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031    1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042  -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162   -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

- Which division is used as the baseline?

Reed Theses

- How does page count vary across year of publication **and** division?

```
theses_mlr <- lm(n_pages ~ year + division, data = theses)
get_regression_table(theses_mlr)
```

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0         -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0         23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011      4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031      1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042    -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162     -3.34    19.9
```

$$\text{Pages} = 1149 - 0.54 \cdot \text{Year} + 34.5 \cdot \text{HSS} + 18.7 \cdot \text{ID} + 12.6 \cdot \text{LL} - 11.2 \cdot \text{MNS} + 8.3 \cdot \text{PRPL}$$

- Which division is used as the baseline?
  - Arts (because it's first alphabetically)

Hypothesis Testing

- The regression table provides p-values for each slope in the model.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.
- The standard error, statistic, and p-values are all calculated using theory-based methods.

Hypothesis Testing

- The regression table provides p-values for each slope in the model.
  - But what hypotheses are being tested?
- In a **MLR model**, we are still interested in determining whether a slope β_i is 0.
  - But we want to investigate this slope in light of the other variables in the model.
- Each row corresponds to a hypothesis test of the form

$$H_0 : \beta_i = 0, \quad \text{given that other variables are included in the model}$$

  - I.e. The `year` row corresponds to the test of

$$H_0 : \beta_{year} = 0, \quad \text{given each division is included in the model}$$

- Reminder: The p-value is the probability of obtaining a statistic as extreme as the observed statistic, **if the null hypothesis were true**.
- The standard error, statistic, and p-values are all calculated using theory-based methods.
  - But the formula is very complicated, requiring matrices and linear algebra (If interested, take Math 392)

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     1149.      126.         9.14   0      902.    1396.
## 2 year            -0.54      0.063     -8.58   0       -0.664   -0.416
## 3 divisionHSS     34.5       5.61       6.16   0       23.5     45.5
## 4 divisionID      18.7       7.38       2.53   0.011    4.22    33.2
## 5 divisionLL      12.6       5.84       2.16   0.031    1.16    24.1
## 6 divisionMNS    -11.2       5.50      -2.03   0.042  -22.0     -0.389
## 7 divisionPRPL     8.28      5.92       1.40   0.162   -3.34    19.9
```

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term         estimate std_error statistic p_value lower_ci upper_ci
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     1149.     126.         9.14   0       902.    1396.
## 2 year            -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS     34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID      18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL      12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS    -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL     8.28     5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0       902.    1396.
## 2 year             -0.54     0.063     -8.58   0        -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0        23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011     4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031     1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042   -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162    -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
    - What does this mean in context?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0        902.    1396.
## 2 year             -0.54     0.063     -8.58   0         -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0         23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011      4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031      1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042    -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162     -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?

- For which would you fail to reject H_0?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term         estimate std_error statistic p_value lower_ci upper_ci
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     1149.     126.         9.14   0        902.    1396.
## 2 year            -0.54     0.063     -8.58   0         -0.664   -0.416
## 3 divisionHSS     34.5      5.61       6.16   0         23.5     45.5
## 4 divisionID      18.7      7.38       2.53   0.011      4.22    33.2
## 5 divisionLL      12.6      5.84       2.16   0.031      1.16    24.1
## 6 divisionMNS    -11.2      5.50      -2.03   0.042    -22.0     -0.389
## 7 divisionPRPL     8.28     5.92       1.40   0.162     -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?
- For which would you fail to reject H_0?
  - What does this mean in context?

Analysis

- Consider the regression table...

```
## # A tibble: 7 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept      1149.     126.         9.14   0      902.    1396.
## 2 year             -0.54     0.063     -8.58   0       -0.664   -0.416
## 3 divisionHSS      34.5      5.61       6.16   0       23.5     45.5
## 4 divisionID       18.7      7.38       2.53   0.011    4.22    33.2
## 5 divisionLL       12.6      5.84       2.16   0.031    1.16    24.1
## 6 divisionMNS     -11.2      5.50      -2.03   0.042  -22.0     -0.389
## 7 divisionPRPL      8.28     5.92       1.40   0.162   -3.34    19.9
```

- For which hypothesis tests would you reject H_0?
  - What does this mean in context?
- For which would you fail to reject H_0?
  - What does this mean in context?
- What is the relationship between the `estimate`, the `std_error` and `lower_ci`, `upper_ci`?

Model Assumptions for MLR

- In order to responsibly use MLR to make inference, we need...

Model Assumptions for MLR

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

Model Assumptions for MLR

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

- How do we check some of these conditions? Why can't we create a scatterplot of residuals as we did for SLR?

Model Assumptions for MLR

- In order to responsibly use MLR to make inference, we need...

1. The relationship between explanatory and response variables must be approximately multilinear linear. (**Linear**)
2. The observations should be independent of one another. (**Independence**)
3. The distribution of residuals should be bell-shaped, unimodal, symmetric, and centered at 0. (**Normal**)
4. The variability of residuals should be roughly constant across entire data set. (**Equal Variability**)

- How do we check some of these conditions? Why can't we create a scatterplot of residuals as we did for SLR?
  - Instead, we will use a scatterplot of residuals vs **predicted values**

Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr)
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

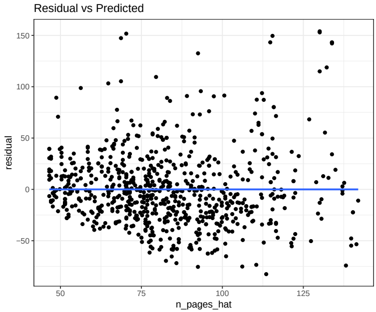
```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

Residuals vs Fitted Values

```r
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```r
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual)
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```

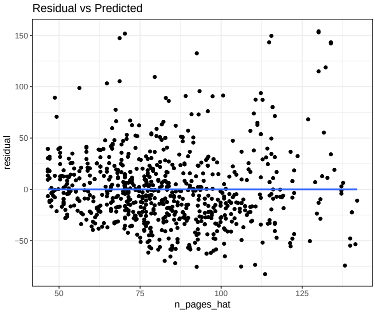


Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual)
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```



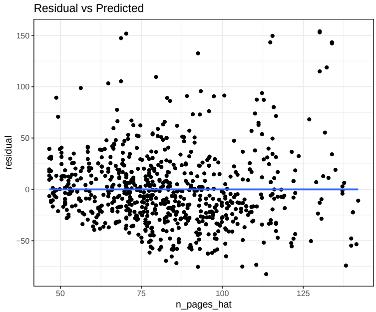
- When analyzing residual vs. predicted plots, look for. . .

Residuals vs Fitted Values

```
mod_res <- get_regression_points(theses_mlr
mod_res %>%
  select(n_pages, n_pages_hat, residual)
```

```
## # A tibble: 772 x 3
##    n_pages n_pages_hat residual
##      <dbl>       <dbl>    <dbl>
##  1      84        72.5    11.5
##  2     139        88.2    50.8
##  3      50        82.2   -32.2
##  4      79        53.6    25.4
##  5      36        47.1   -11.1
##  6      80        54.1    25.9
##  7     134        67.6    66.4
##  8      58        75.2   -17.2
##  9      69        75.7    -6.74
## 10      74        85.4   -11.4
## # ... with 762 more rows
```

```
mod_res %>% ggplot(aes(x = n_pages_hat, y = residual)
  geom_point()+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Residual vs Predicted")
```



- When analyzing residual vs. predicted plots, look for...
  - Non-linear patterns
  - Increasing variability across range of predicted values
  - Outliers with atypical predicted value or large residual

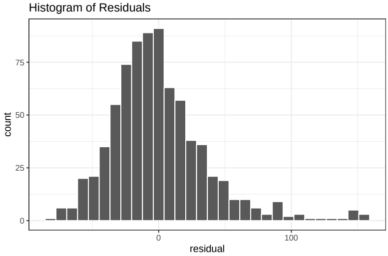
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

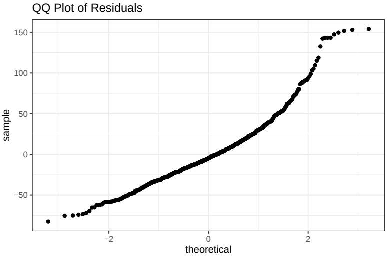
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

```
ggplot(mod_res, aes(x = residual))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Histogram of Residuals")
```



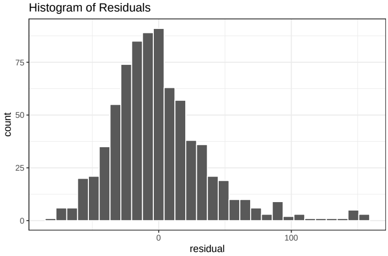

```
ggplot(mod_res, aes(sample = residual))+
  geom_point(stat = "qq")+
  labs(title = "QQ Plot of Residuals")
```



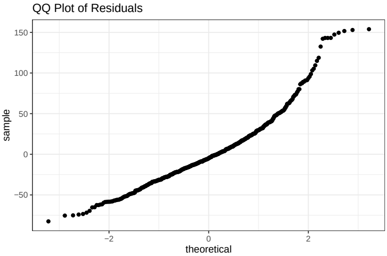
Distribution of Residuals

- We can still look at the histogram and QQ Plot of residuals, as we did for SLR:

```
ggplot(mod_res, aes(x = residual))+
  geom_histogram(bins = 30, color = "white")+
  labs(title = "Histogram of Residuals")
```




```
ggplot(mod_res, aes(sample = residual))+
  geom_point(stat = "qq")+
  labs(title = "QQ Plot of Residuals")
```



- We see some evidence that residuals are not Normally distributed (right-skew)