Inference for Multiple Linear Regression

Prof. Wells

STA 209, 5/10/23

Outline

In this lecture, we will...

- Use R to perform theory-based inference for regression models
- Review framework for multilinear regression
- Discuss inference procedures for MLR models
- Investigate tools for "Model Selection"

Section 1

Theory-Based Methods

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *standard error* and *shape* of the sampling distribution for $\hat{\beta}_1$

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *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 standard error

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *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 standard error

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \text{(DON'T WRITE / MEMORIZE!)}$$

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *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 standard error

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \text{(DON'T WRITE / MEMORIZE!)}$$

- We perform a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

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

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *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 standard error

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \text{(DON'T WRITE / MEMORIZE!)}$$

- We perform a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

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

- And we create a confidence interval for β_1 using

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

Inference for Slope

- Consider the linear model $Y = \beta_0 + \beta_1 X + \epsilon$
- Can we make inference about the slope β_1 of a linear model *without* using simulation?
  - We need to know the *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 standard error

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n-2} \frac{\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \text{(DON'T WRITE / MEMORIZE!)}$$

- We perform a hypothesis test of $H_0 : \beta_1 = 0$ using the test statistic

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

- And we create a confidence interval for β_1 using

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

- The reference distribution is the t-distribution with $n - 2$ degrees of freedom.

Calculating test statistics and confidence intervals

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

Calculating test statistics and confidence intervals

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

```
my_mod <- lm(Y ~ X, data = my_data)
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     4.86     3.17       1.53   0.137   -1.64     11.4
## 2 X             1.67     0.625      2.67   0.013    0.386     2.95
```

Calculating test statistics and confidence intervals

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

```
my_mod <- lm(Y ~ X, data = my_data)
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     4.86     3.17       1.53   0.137   -1.64     11.4
## 2 X             1.67     0.625      2.67   0.013    0.386     2.95
```

- 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 statistics and confidence intervals for β_1 *without* tedious calculation?
  - Yes! Using the `lm` function in R.

```
my_mod <- lm(Y ~ X, data = my_data)
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     4.86     3.17       1.53   0.137   -1.64     11.4
## 2 X             1.67     0.625      2.67   0.013    0.386     2.95
```

- 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 statistics and confidence intervals for β_1 *without* tedious calculation?
  - Yes! Using the `lm` function in R.

```
my_mod <- lm(Y ~ X, data = my_data)
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     4.86     3.17       1.53   0.137   -1.64     11.4
## 2 X             1.67     0.625      2.67   0.013    0.386     2.95
```

- 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)

Section 2

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 $X_1, X_2, \ldots, X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

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 $X_1, X_2, \ldots, X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- 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 $X_1, X_2, \ldots, X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- 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 $X_1, X_2, \ldots, X_k$:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k + \epsilon$$

- 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**

Newborn Birth Weights

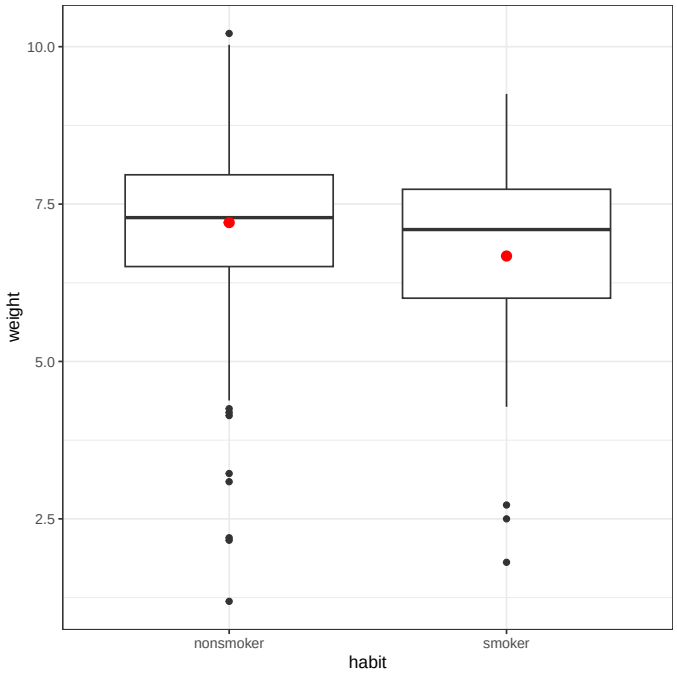
- A number of factors contribute to the birth weight of a newborn: gestational length, genetic factors, and mother's age, health, nutrition, and habits

Newborn Birth Weights

- A number of factors contribute to the birth weight of a newborn: gestational length, genetic factors, and mother's age, health, nutrition, and habits
- Researchers are interested in determining whether birth weight of babies born to mothers who smoke differs from that of babies born to mothers who do not.

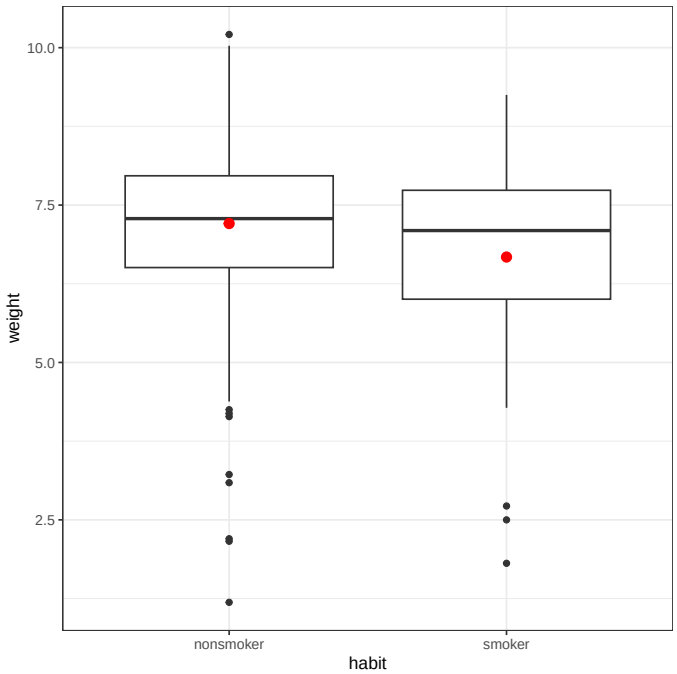
Newborn Birth Weights

- A number of factors contribute to the birth weight of a newborn: gestational length, genetic factors, and mother's age, health, nutrition, and habits
- Researchers are interested in determining whether birth weight of babies born to mothers who smoke differs from that of babies born to mothers who do not.



Newborn Birth Weights

- A number of factors contribute to the birth weight of a newborn: gestational length, genetic factors, and mother's age, health, nutrition, and habits
- Researchers are interested in determining whether birth weight of babies born to mothers who smoke differs from that of babies born to mothers who do not.




```
## # A tibble: 2 x 4
##   habit       xbar     s     n
##   <chr>      <dbl> <dbl> <int>
## 1 nonsmoker   7.21  1.20   446
## 2 smoker      6.67  1.54    54
```

- Test statistic:

$$t = \frac{7.21 - 6.67}{\sqrt{\frac{1.20^2}{446} + \frac{1.54^2}{54}}} = 2.97$$

- P-value:

```
2*(1-pt(2.97, df = 53))
```

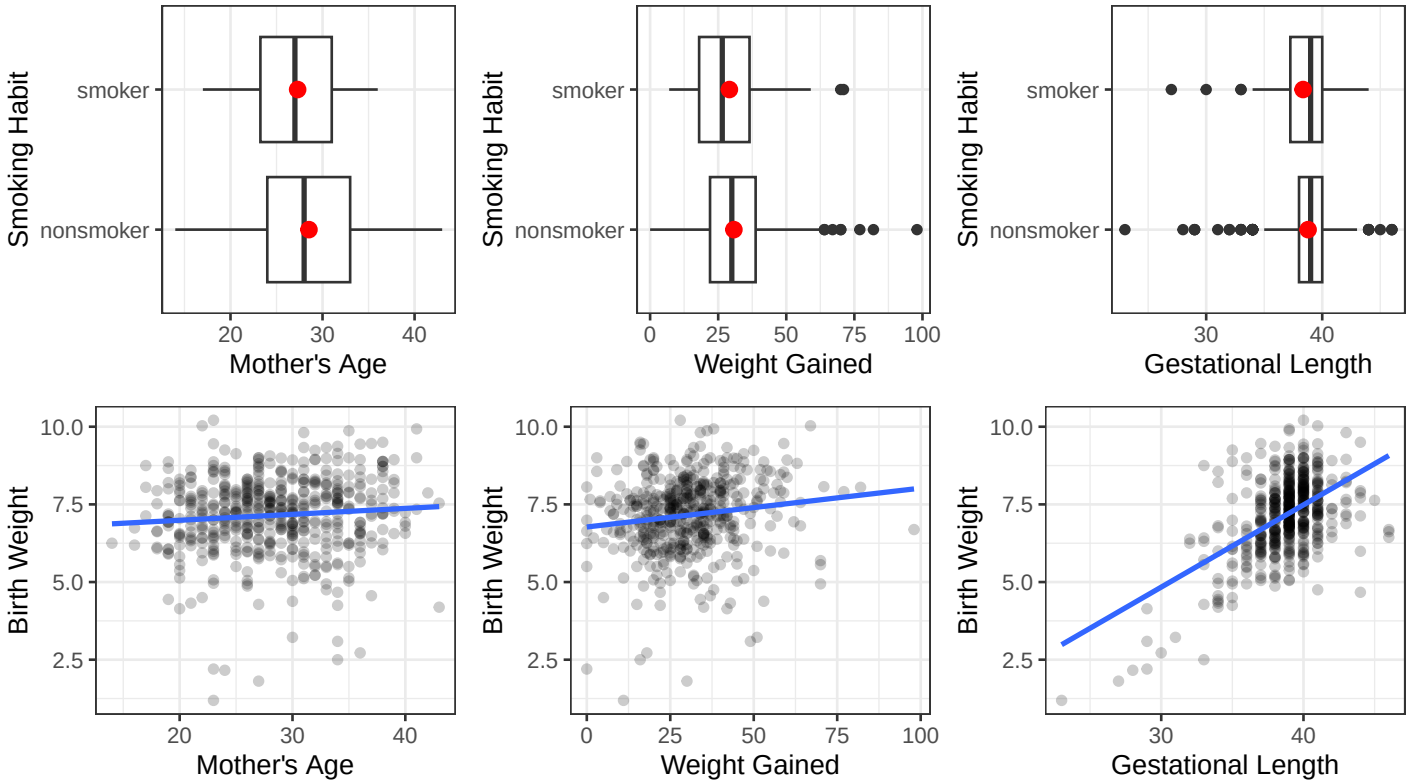
```
## [1] 0.0045
```

Confounding Factors

- However, smoking habits may be associated with other measures that also influence birth weight (mother's age and weight gained during pregnancy, gestational length)

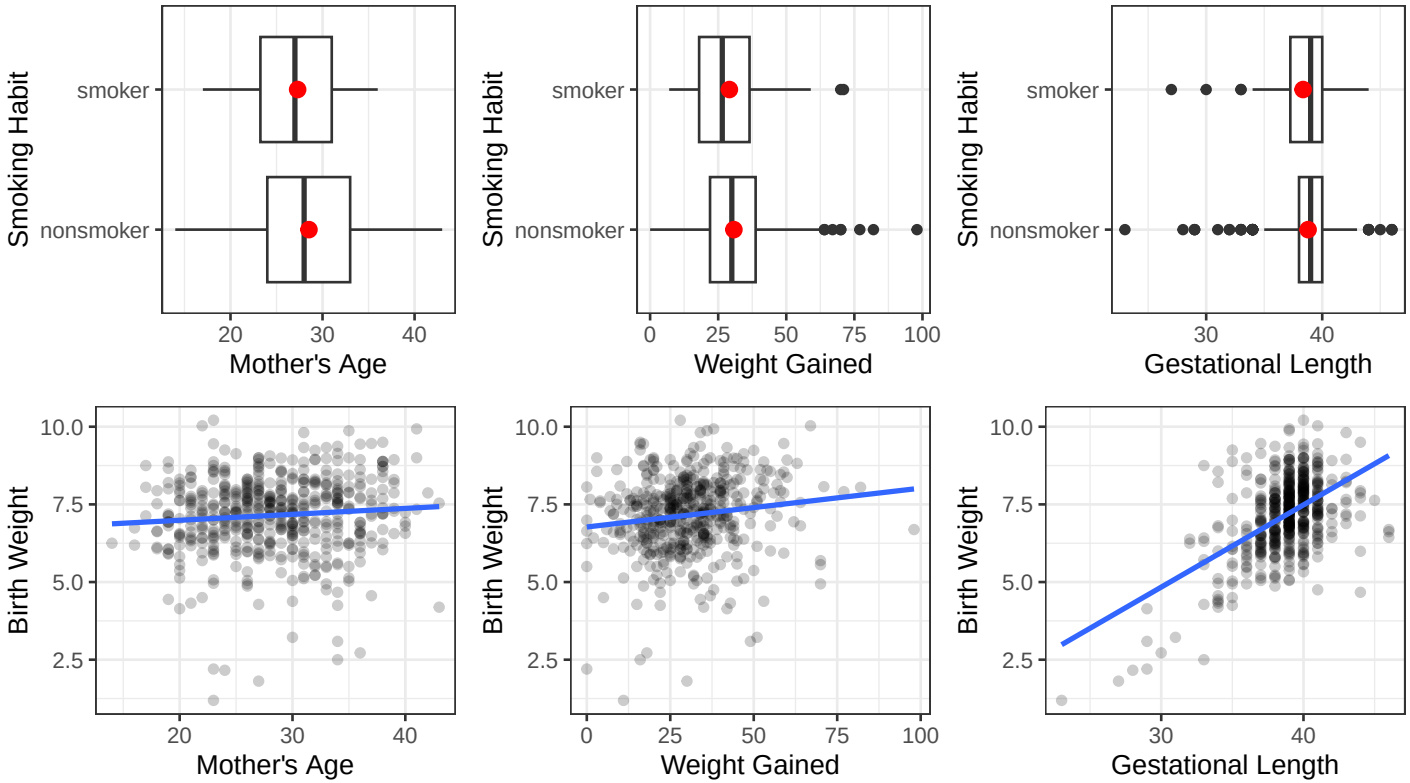
Confounding Factors

- However, smoking habits may be associated with other measures that also influence birth weight (mother's age and weight gained during pregnancy, gestational length)



Confounding Factors

- However, smoking habits may be associated with other measures that also influence birth weight (mother's age and weight gained during pregnancy, gestational length)



- We would like to isolate the effect of smoking on birth weight, while controlling these other factors.

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

```
mlr_mod <- lm(weight ~ weeks + age + gained + habit, data = births14)
get_regression_table(mlr_mod)
```

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

```
mlr_mod <- lm(weight ~ weeks + age + gained + habit, data = births14)
get_regression_table(mlr_mod)
```

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

$$\text{Weight} = -3.63 + 0.26 \cdot \text{weeks} + 0.016 \cdot \text{age} + 0.01 \cdot \text{gained} - 0.387 \cdot \text{smoker}$$

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

```
mlr_mod <- lm(weight ~ weeks + age + gained + habit, data = births14)
get_regression_table(mlr_mod)
```

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

$$\text{Weight} = -3.63 + 0.26 \cdot \text{weeks} + 0.016 \cdot \text{age} + 0.01 \cdot \text{gained} - 0.387 \cdot \text{smoker}$$

- What is the predicted birth weight of baby born at 40 weeks to a mother of 35 years who gained 20 pounds and is a non-smoker?

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

```
mlr_mod <- lm(weight ~ weeks + age + gained + habit, data = births14)
get_regression_table(mlr_mod)
```

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

$$\text{Weight} = -3.63 + 0.26 \cdot \text{weeks} + 0.016 \cdot \text{age} + 0.01 \cdot \text{gained} - 0.387 \cdot \text{smoker}$$

- What is the predicted birth weight of baby born at 40 weeks to a mother of 35 years who gained 20 pounds and is a non-smoker?
- What does the coefficient on weeks mean?

Multilinear Model

We create a multilinear model for birth weight, as a function of gestational length, mother's age, weight gained, and smoking habit:

```
mlr_mod <- lm(weight ~ weeks + age + gained + habit, data = births14)
get_regression_table(mlr_mod)
```

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

$$\text{Weight} = -3.63 + 0.26 \cdot \text{weeks} + 0.016 \cdot \text{age} + 0.01 \cdot \text{gained} - 0.387 \cdot \text{smoker}$$

- What is the predicted birth weight of baby born at 40 weeks to a mother of 35 years who gained 20 pounds and is a non-smoker?
- What does the coefficient on weeks mean?
- What does the coefficient on smoker mean?

Hypothesis Testing

- The regression table provides p-values for each variable 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 variable 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 variable 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 variable 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 `habit_smoker` row corresponds to the test of

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

Hypothesis Testing

- The regression table provides *p*-values for each variable 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 `habit_smoker` row corresponds to the test of

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

- 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 variable 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 `habit_smoker` row corresponds to the test of

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

- 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 variable 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 `habit_smoker` row corresponds to the test of

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

- 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 linear algebra (If interested, take STA 336)

Analysis

- Consider the regression table...

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

Analysis

- Consider the regression table...

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

- Should we reject $H_0 : \beta_{smoker} = 0$?

Analysis

- Consider the regression table. . .

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

- Should we reject $H_0 : \beta_{smoker} = 0$?
  - Including other variables in the model, it is unlikely we would have seen a coefficient on $\mathrm{smoking}$ as large as we did, if there were no relationship between smoking and birth weight.

Analysis

- Consider the regression table. . .

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

- Should we reject $H_0 : \beta_{smoker} = 0$?
  - Including other variables in the model, it is unlikely we would have seen a coefficient on $\mathrm{smoking}$ as large as we did, if there were no relationship between smoking and birth weight.
  - This gives relatively strong evidence that smoking has an effect on birth weight, even after taking other factors into account.

Analysis

- Consider the regression table...

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

- Should we reject $H_0 : \beta_{smoker} = 0$?
  - Including other variables in the model, it is unlikely we would have seen a coefficient on $\mathrm{smoking}$ as large as we did, if there were no relationship between smoking and birth weight.
  - This gives relatively strong evidence that smoking has an effect on birth weight, even after taking other factors into account.

- What does the p-value on age mean?

Analysis

- Consider the regression table. . .

```
## # A tibble: 5 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     -3.63      0.788     -4.61   0       -5.18    -2.08
## 2 weeks          0.26      0.019     13.5    0        0.222    0.297
## 3 age            0.016     0.008      1.95   0.051    0        0.032
## 4 gained         0.01      0.003      3.03   0.003    0.004    0.017
## 5 habitsmoker   -0.387     0.151     -2.56   0.011   -0.684   -0.091
```

- Should we reject $H_0 : \beta_{smoker} = 0$?
  - Including other variables in the model, it is unlikely we would have seen a coefficient on $\mathrm{smoking}$ as large as we did, if there were no relationship between smoking and birth weight.
  - This gives relatively strong evidence that smoking has an effect on birth weight, even after taking other factors into account.

- What does the p-value on age mean?

- How does the coefficient on `smoker` in the MLR model compare to the observed difference in our t-test?

$$\mathrm{weight}_{\mathrm{smoker}} - \mathrm{weight}_{\mathrm{non\text{-}smoker}} = 6.67 - 7.21 = -0.54$$

Section 3

Model Assumptions for MLR

Model Assumptions: LINE

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

Model Assumptions: LINE

- 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: LINE

- 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: LINE

- 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

```
mlr_res <- get_regression_points(mlr_mod)
```

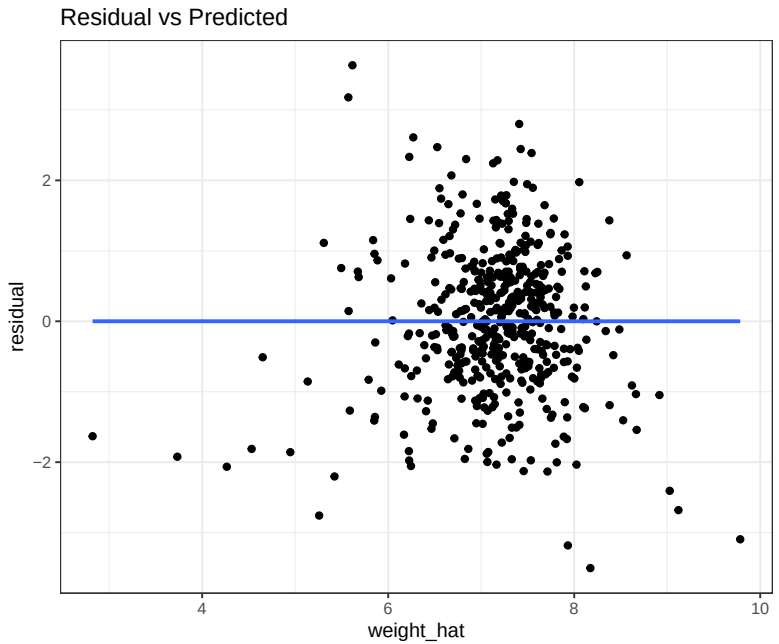
```
## # A tibble: 484 x 3
##    weight weight_hat residual
##     <dbl>      <dbl>    <dbl>
##  1   7.84       7.21    0.633
##  2   7.28       7.22    0.061
##  3   8.19       7.73    0.464
##  4   5.69       6.79   -1.10
##  5   6.26       7.27   -1.01
##  6   6.87       7.51   -0.638
##  7   7.36       7.93   -0.569
##  8   5.82       6.64   -0.823
##  9   7.25       7.47   -0.216
## 10   8.19       7.48    0.705
## # ... with 474 more rows
```

Residuals vs Fitted Values

```
mlr_res <- get_regression_points(mlr_mod)
```

```
## # A tibble: 484 x 3
##    weight weight_hat residual
##     <dbl>      <dbl>    <dbl>
##  1   7.84       7.21    0.633
##  2   7.28       7.22    0.061
##  3   8.19       7.73    0.464
##  4   5.69       6.79   -1.10
##  5   6.26       7.27   -1.01
##  6   6.87       7.51   -0.638
##  7   7.36       7.93   -0.569
##  8   5.82       6.64   -0.823
##  9   7.25       7.47   -0.216
## 10   8.19       7.48    0.705
## # ... with 474 more rows
```

```
ggplot(mlr_res, aes(x = weight_hat, y = residual))+
  geom_point()+geom_smooth(method = "lm", se = F)
```

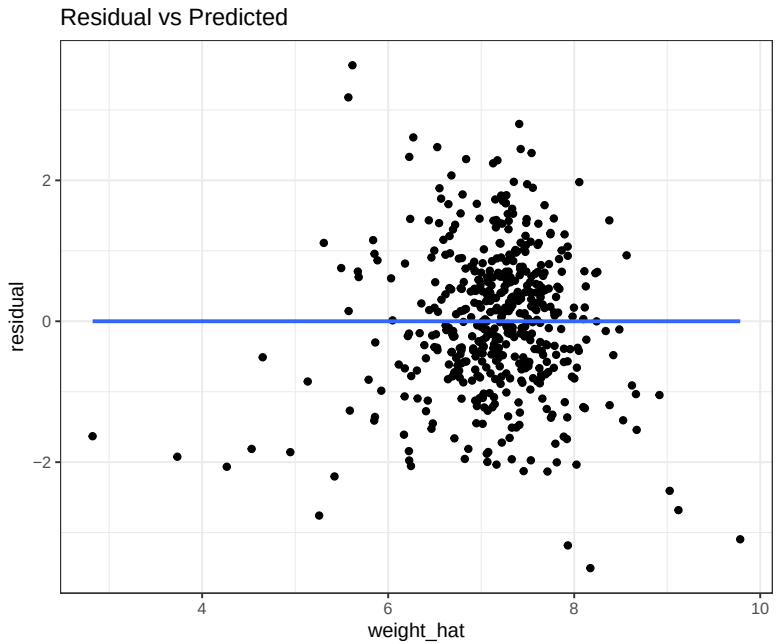


Residuals vs Fitted Values

```
mlr_res <- get_regression_points(mlr_mod)
```

```
## # A tibble: 484 x 3
##    weight weight_hat residual
##     <dbl>      <dbl>    <dbl>
##  1   7.84       7.21    0.633
##  2   7.28       7.22    0.061
##  3   8.19       7.73    0.464
##  4   5.69       6.79   -1.10
##  5   6.26       7.27   -1.01
##  6   6.87       7.51   -0.638
##  7   7.36       7.93   -0.569
##  8   5.82       6.64   -0.823
##  9   7.25       7.47   -0.216
## 10   8.19       7.48    0.705
## # ... with 474 more rows
```

```
ggplot(mlr_res, aes(x = weight_hat, y = residual))+
  geom_point()+geom_smooth(method = "lm", se = F)
```



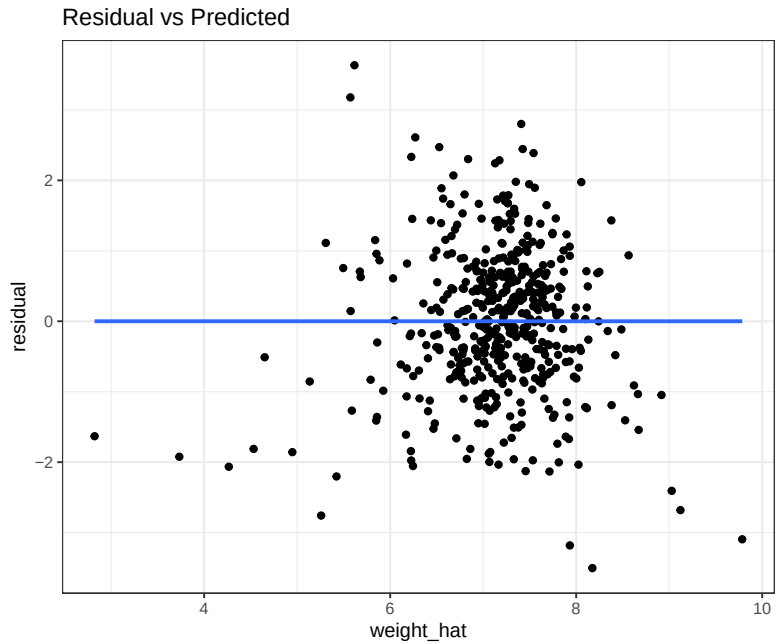
- When analyzing residual vs. predicted plots, look for...

Residuals vs Fitted Values

```
mlr_res <- get_regression_points(mlr_mod)
```

```
## # A tibble: 484 x 3
##    weight weight_hat residual
##     <dbl>      <dbl>    <dbl>
##  1   7.84       7.21    0.633
##  2   7.28       7.22    0.061
##  3   8.19       7.73    0.464
##  4   5.69       6.79   -1.10
##  5   6.26       7.27   -1.01
##  6   6.87       7.51   -0.638
##  7   7.36       7.93   -0.569
##  8   5.82       6.64   -0.823
##  9   7.25       7.47   -0.216
## 10   8.19       7.48    0.705
## # ... with 474 more rows
```

```
ggplot(mlr_res, aes(x = weight_hat, y = residual))+
  geom_point()+geom_smooth(method = "lm", se = F)
```



- 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

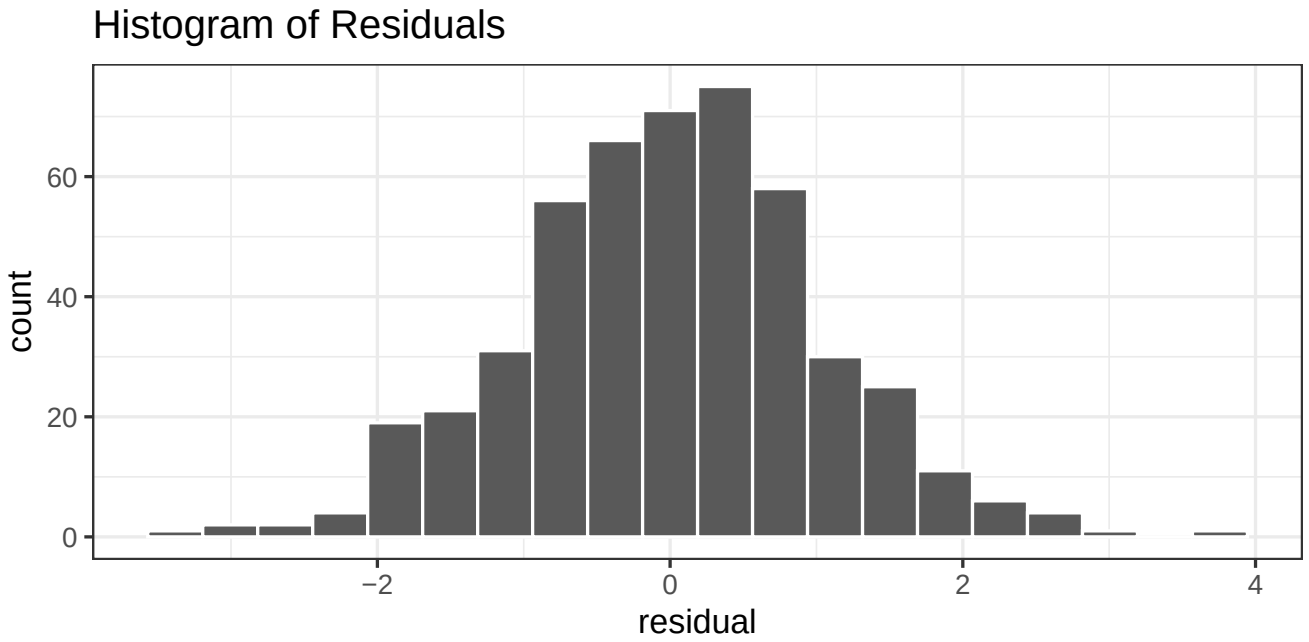
Distribution of Residuals

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

Distribution of Residuals

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

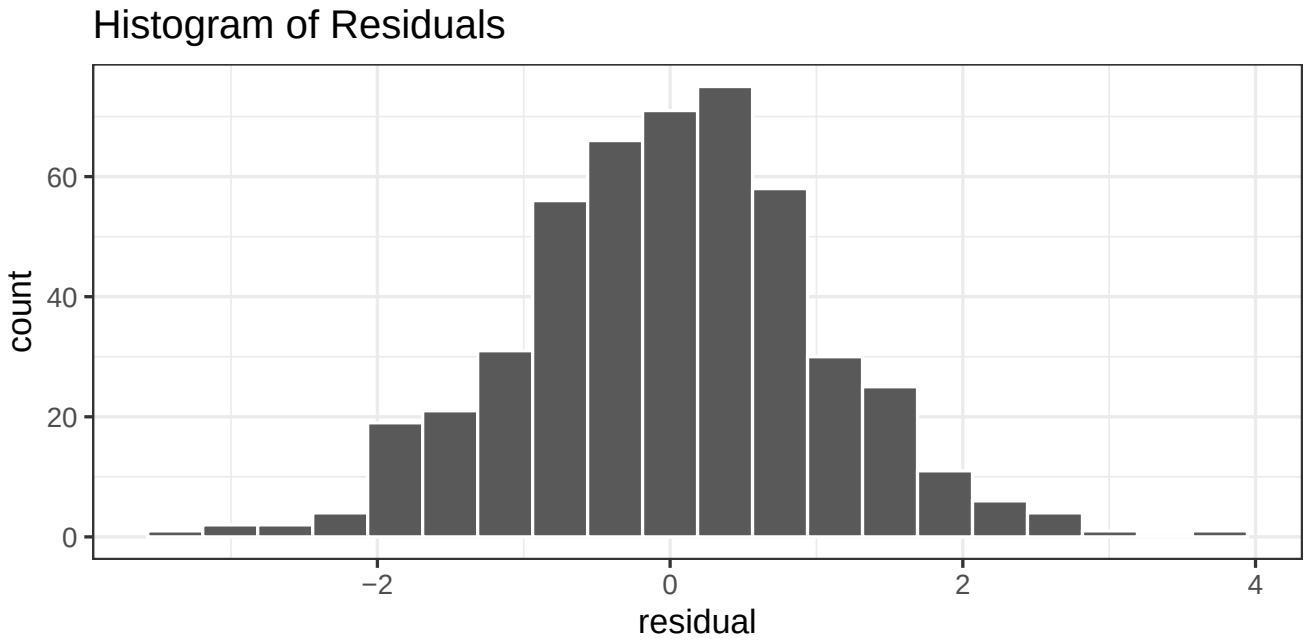
```
ggplot(mlr_res, aes(x = residual))+
  geom_histogram(bins = 20, color = "white")+ labs(title = "Histogram of Residuals")
```



Distribution of Residuals

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

```
ggplot(mlr_res, aes(x = residual))+
  geom_histogram(bins = 20, color = "white")+ labs(title = "Histogram of Residuals")
```



- Residuals do appear to be approximately Normally distributed (unimodal, bell-shaped, symmetric, centered at 0)

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.
- Therefore, the p-values and confidence intervals obtained from theory-based methods for MLR are reasonably accurate.

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.
- Therefore, the p-values and confidence intervals obtained from theory-based methods for MLR are reasonably accurate.
- We tested

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

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.
- Therefore, the p-values and confidence intervals obtained from theory-based methods for MLR are reasonably accurate.
- We tested

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

  - We obtained a p-value of 0.011, and rejected the null hypothesis in favor of the alternative, at the 0.05 level

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.
- Therefore, the p-values and confidence intervals obtained from theory-based methods for MLR are reasonably accurate.
- We tested

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

  - We obtained a p-value of 0.011, and rejected the null hypothesis in favor of the alternative, at the 0.05 level
- This data does provide evidence that, even after taking other possible confounding factors into account, smoking during pregnancy is associated with lower birth weights.

Conclusion

- Our data appears to reasonably satisfy the conditions for inference using multilinear regression.
- Therefore, the p-values and confidence intervals obtained from theory-based methods for MLR are reasonably accurate.
- We tested

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

  - We obtained a p-value of 0.011, and rejected the null hypothesis in favor of the alternative, at the 0.05 level
- This data does provide evidence that, even after taking other possible confounding factors into account, smoking during pregnancy is associated with lower birth weights.
  - Moreover, in our analysis, we also observed that gestational length and weight gained had *p*-values of approximately 0, while age had a p-value of 0.051

Section 4

Model Selection

Model Selection

- How can we use p-values to decide which of several models is best?

Model Selection

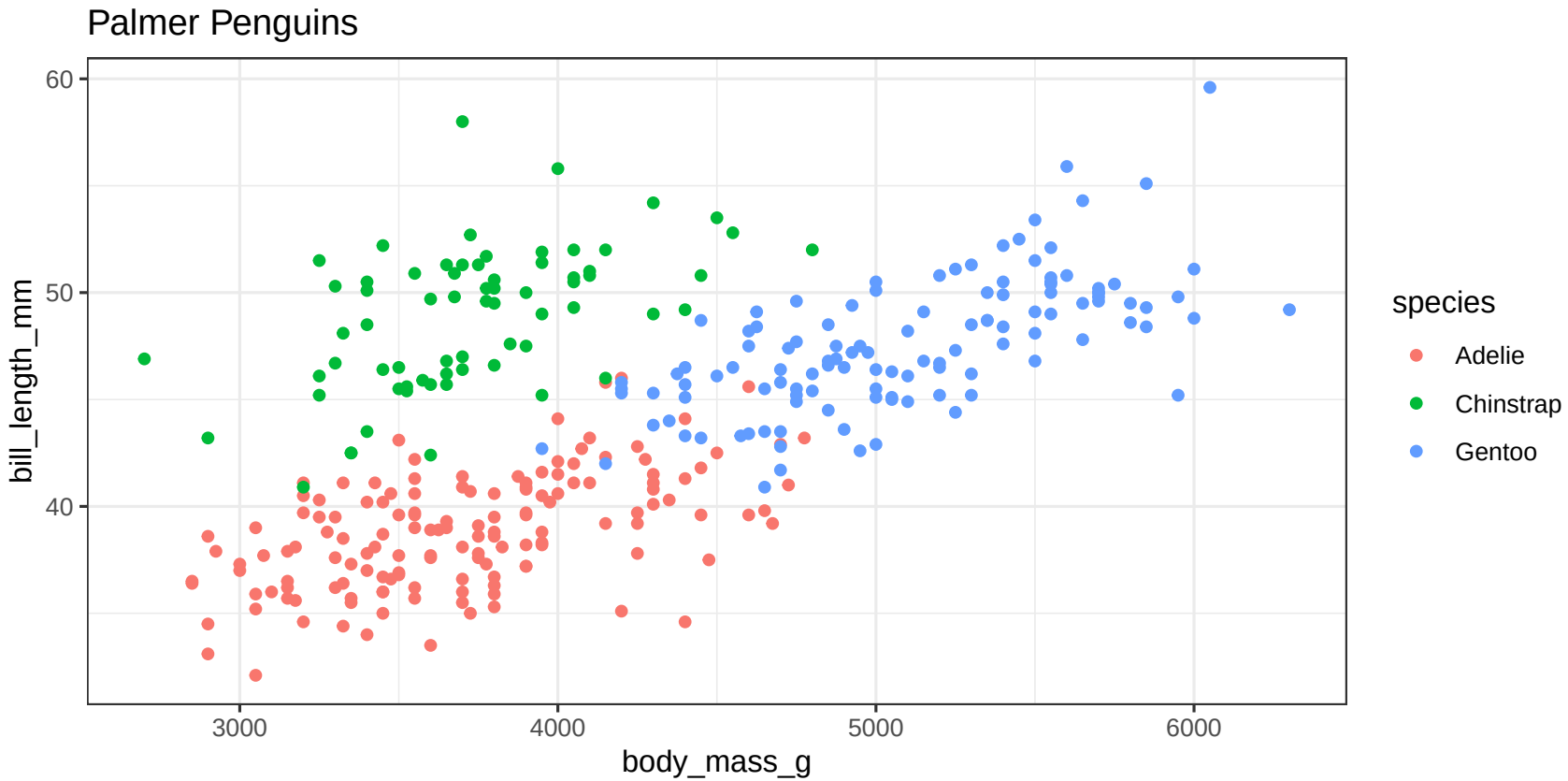
- How can we use p-values to decide which of several models is best?
- Recall that `palmerpenguins` data from earlier this term:

Model Selection

- How can we use p-values to decide which of several models is best?
- Recall that `palmerpenguins` data from earlier this term:
- We investigated the relationship between *bill length*, *body mass* and *species*

Model Selection

- How can we use p-values to decide which of several models is best?
- Recall that `palmerpenguins` data from earlier this term:
- We investigated the relationship between *bill length*, *body mass* and *species*

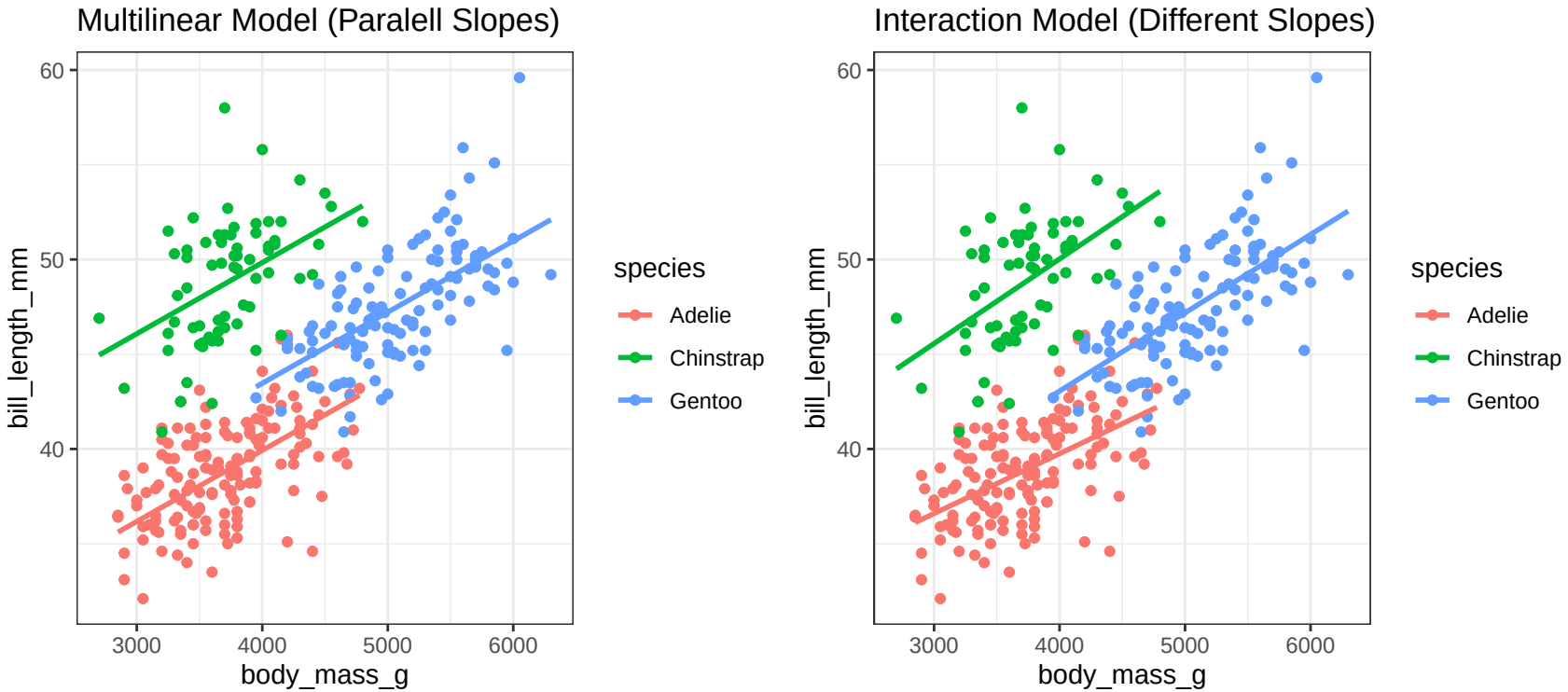


Interaction vs Multilinear Regression Model

- We had two candidates for models:

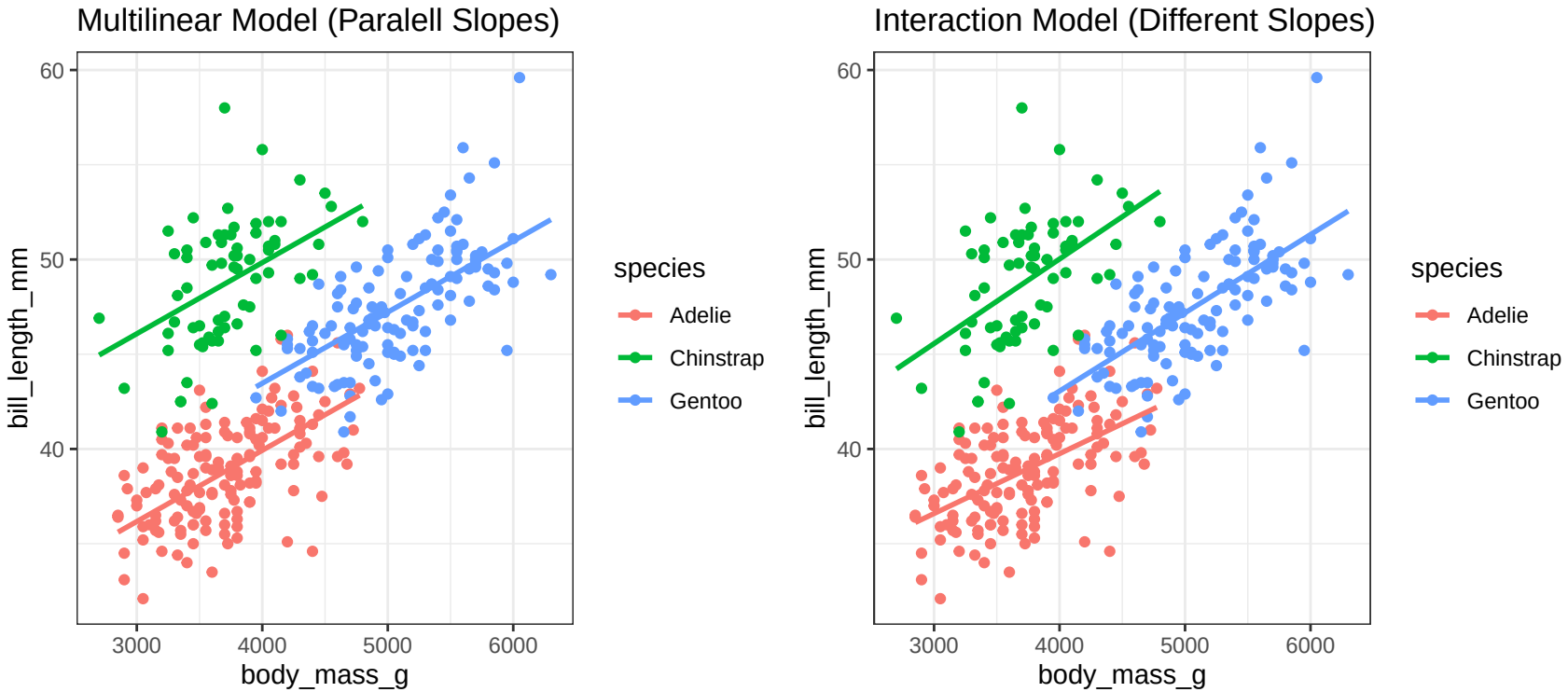
Interaction vs Multilinear Regression Model

- We had two candidates for models:



Interaction vs Multilinear Regression Model

- We had two candidates for models:



- We concluded that multilinear model was superior, since both models were relatively similar, but the multilinear model was simpler

The Multilinear Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # 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          24.9       1.09      22.9        0   22.8     27.1
## 2 body_mass_g         0.004     0         13.0        0    0.003    0.004
## 3 speciesChinstrap    9.91      0.355     27.9        0    9.21    10.6
## 4 speciesGentoo       3.54      0.5        7.08       0    2.56     4.52
```

$$\hat{\text{Bill Length}} = 24.9 + 0.004 \cdot \text{Mass} + 9.91 \cdot \text{Chinstrap} + 3.54 \cdot \text{Gentoo}$$

The Multilinear Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # 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          24.9       1.09      22.9        0   22.8     27.1
## 2 body_mass_g         0.004     0         13.0        0    0.003    0.004
## 3 speciesChinstrap    9.91      0.355     27.9        0    9.21    10.6
## 4 speciesGentoo       3.54      0.5        7.08       0    2.56     4.52
```

$$\widehat{\text{Bill Length}} = 24.9 + 0.004 \cdot \text{Mass} + 9.91 \cdot \text{Chinstrap} + 3.54 \cdot \text{Gentoo}$$

- Note the p-values for all coefficients are (very close to) 0.

The Multilinear Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # 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          24.9       1.09      22.9        0   22.8     27.1
## 2 body_mass_g         0.004     0         13.0        0    0.003    0.004
## 3 speciesChinstrap    9.91      0.355     27.9        0    9.21    10.6
## 4 speciesGentoo       3.54      0.5        7.08       0    2.56     4.52
```

$$\hat{\text{Bill Length}} = 24.9 + 0.004 \cdot \text{Mass} + 9.91 \cdot \text{Chinstrap} + 3.54 \cdot \text{Gentoo}$$

- Note the p-values for all coefficients are (very close to) 0.
  - We would reject the null hypotheses that those slope parameters are 0 in this model.

The Multilinear Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g + species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # 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          24.9       1.09      22.9        0   22.8     27.1
## 2 body_mass_g         0.004     0         13.0        0    0.003    0.004
## 3 speciesChinstrap    9.91      0.355     27.9        0    9.21    10.6
## 4 speciesGentoo       3.54      0.5        7.08       0    2.56     4.52
```

$$\hat{\text{Bill Length}} = 24.9 + 0.004 \cdot \text{Mass} + 9.91 \cdot \text{Chinstrap} + 3.54 \cdot \text{Gentoo}$$

- Note the p-values for all coefficients are (very close to) 0.
  - We would reject the null hypotheses that those slope parameters are 0 in this model.
- This suggests that together, *body mass* and *species* do a reasonable job at predicting the value of *bill length*

The Interaction Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g * species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1       1.63     16.6     0       23.9     30.3
## 2 body_mass_g                0.003     0         7.23    0        0.002    0.004
## 3 speciesChinstrap           5.06      3.31      1.53    0.127   -1.45    11.6
## 4 speciesGentoo             -0.575     2.79     -0.206   0.837   -6.07     4.92
## 5 body_mass_g:speciesChi~    0.001     0.001     1.48    0.141    0        0.003
## 6 body_mass_g:speciesGen~    0.001     0.001     1.56    0.12     0        0.002
```

$$\widehat{\text{Bill Length}} = 27.1 + 0.0032 \cdot \text{Mass} + 5.06 \cdot \text{Chinstrap} - 0.575 \cdot \text{Gentoo} + 0.0013 \cdot \text{Mass} \cdot \text{Chinstrap} + 0.001 \cdot \text{Mass} \cdot \text{Gentoo}$$

The Interaction Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g * species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1       1.63     16.6     0       23.9     30.3
## 2 body_mass_g                0.003     0         7.23    0        0.002    0.004
## 3 speciesChinstrap           5.06      3.31      1.53    0.127   -1.45    11.6
## 4 speciesGentoo             -0.575     2.79     -0.206   0.837   -6.07     4.92
## 5 body_mass_g:speciesChi~    0.001     0.001     1.48    0.141    0        0.003
## 6 body_mass_g:speciesGen~    0.001     0.001     1.56    0.12     0        0.002
```

$$\widehat{\text{Bill Length}} = 27.1 + 0.0032 \cdot \text{Mass} + 5.06 \cdot \text{Chinstrap} - 0.575 \cdot \text{Gentoo} + 0.0013 \cdot \text{Mass} \cdot \text{Chinstrap} + 0.001 \cdot \text{Mass} \cdot \text{Gentoo}$$

- Note now that many of the p-values are larger than 0.1

The Interaction Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g * species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1       1.63     16.6     0       23.9     30.3
## 2 body_mass_g                0.003     0         7.23    0        0.002    0.004
## 3 speciesChinstrap           5.06      3.31      1.53    0.127   -1.45    11.6
## 4 speciesGentoo             -0.575     2.79     -0.206   0.837   -6.07     4.92
## 5 body_mass_g:speciesChi~    0.001     0.001     1.48    0.141    0        0.003
## 6 body_mass_g:speciesGen~    0.001     0.001     1.56    0.12     0        0.002
```

$$\widehat{\text{Bill Length}} = 27.1 + 0.0032 \cdot \text{Mass} + 5.06 \cdot \text{Chinstrap} - 0.575 \cdot \text{Gentoo} + 0.0013 \cdot \text{Mass} \cdot \text{Chinstrap} + 0.001 \cdot \text{Mass} \cdot \text{Gentoo}$$

- Note now that many of the p-values are larger than 0.1
  - We would not reject the null hypotheses that those coefficients are 0 in this model

The Interaction Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g * species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1       1.63     16.6     0       23.9     30.3
## 2 body_mass_g                0.003     0         7.23    0        0.002    0.004
## 3 speciesChinstrap           5.06      3.31      1.53    0.127   -1.45    11.6
## 4 speciesGentoo             -0.575     2.79     -0.206   0.837   -6.07     4.92
## 5 body_mass_g:speciesChi~    0.001     0.001     1.48    0.141    0        0.003
## 6 body_mass_g:speciesGen~    0.001     0.001     1.56    0.12     0        0.002
```

$$\widehat{\text{Bill Length}} = 27.1 + 0.0032 \cdot \text{Mass} + 5.06 \cdot \text{Chinstrap} - 0.575 \cdot \text{Gentoo} + 0.0013 \cdot \text{Mass} \cdot \text{Chinstrap} + 0.001 \cdot \text{Mass} \cdot \text{Gentoo}$$

- Note now that many of the p-values are larger than 0.1
  - We would not reject the null hypotheses that those coefficients are 0 in this model
  - This sample does not provide sufficient evidence to suggest that each penguin species has its own slope for body mass.

The Interaction Model

```
penguins_mlr <- lm(bill_length_mm ~ body_mass_g * species, data = penguins)
get_regression_table(penguins_mlr)
```

```
## # A tibble: 6 x 7
##   term                    estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 27.1       1.63     16.6     0       23.9     30.3
## 2 body_mass_g                0.003     0         7.23    0        0.002    0.004
## 3 speciesChinstrap           5.06      3.31      1.53    0.127   -1.45    11.6
## 4 speciesGentoo             -0.575     2.79     -0.206   0.837   -6.07     4.92
## 5 body_mass_g:speciesChi~    0.001     0.001     1.48    0.141    0        0.003
## 6 body_mass_g:speciesGen~    0.001     0.001     1.56    0.12     0        0.002
```

$$\widehat{\text{Bill Length}} = 27.1 + 0.0032 \cdot \text{Mass} + 5.06 \cdot \text{Chinstrap} - 0.575 \cdot \text{Gentoo} + 0.0013 \cdot \text{Mass} \cdot \text{Chinstrap} + 0.001 \cdot \text{Mass} \cdot \text{Gentoo}$$

- Note now that many of the p-values are larger than 0.1
  - We would not reject the null hypotheses that those coefficients are 0 in this model
  - This sample does not provide sufficient evidence to suggest that each penguin species has its own slope for body mass.
  - It is still possible that the penguin species DO have different slopes on body mass, but our sample was not large enough to detect a potentially small difference