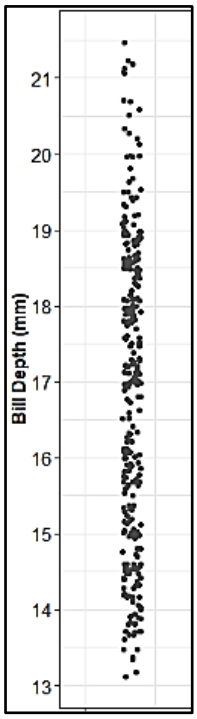
Chapter 14: Multiple Linear Modeling

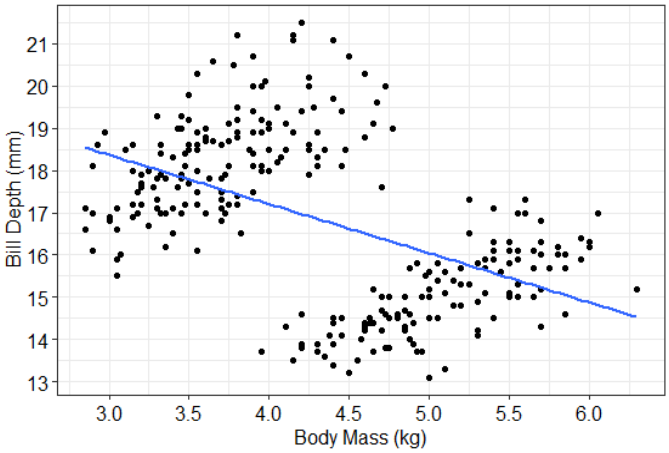
Introducing Multiple Predictors

Investigation: On Palmer Island, biologists are studying the evolutionary development of penguin populations. One variable of interest is the bill depth (beak depth) of these penguins and explaining the variation they see in this variable.

Unit of Observation: One penguin

Response variable: Bill depth





Naturally, we would expect penguins with a higher body mass to have deeper bills. Perhaps that might be a helpful predictor

Predictor: Body mass

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.0339   0.5036    43.75   <2e-16 ***
body_mass_kg  -1.1621   0.1177    -9.87   <2e-16 ***
---

Residual standard error: 1.744 on 340 degrees of freedom
Multiple R-squared:  0.2227,  Adjusted R-squared:  0.2204
F-statistic: 97.41 on 1 and 340 DF,  p-value: < 2.2e-16
```

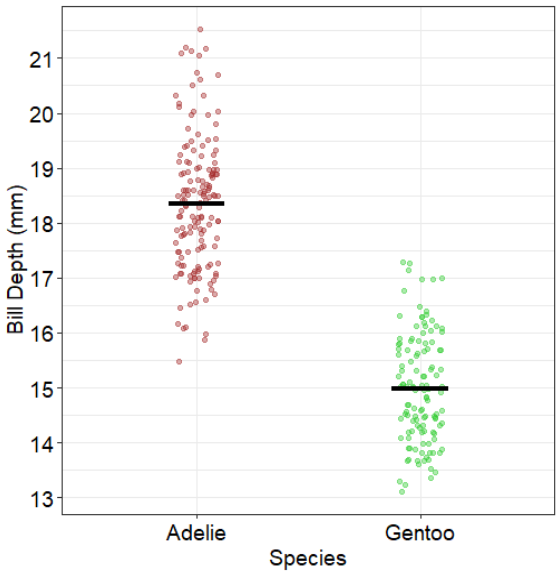
What do you notice about this relationship? Does it follow the trend you would expect? How might we explain what we see in the scatterplot?

Next, the biologists consider two different penguin species on the island: "Gentoo" and "Adelie." Perhaps the penguins' species explains differences in bill depth.

After stratifying by Species, we get the following result

Table 1. Bill Depths by Species

	Adelie	**Gentoo**
Sample Means	18.346	14.982
Sample SD	1.217	0.981



What do you notice about this relationship?

- **A Linear Model with…A binary predictor?**
  - Even without a numeric scale, we could create a linear model using only a binary predictor by treating species as a "dummy variable."
  - **Dummy Variable:** A variable whose levels have been converted to the values <u>0 and 1.</u>
  - We use the term "dummy" because the assignment of 0 and 1 to each level is arbitrary and carries no contextual meaning.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.34636    0.09092  201.79   <2e-16 ***
speciesGentoo -3.36424    0.13570  -24.79   <2e-16 ***
---

Residual standard error: 1.117 on 272 degrees of freedom
Multiple R-squared:  0.6932,  Adjusted R-squared:  0.6921
F-statistic: 614.7 on 1 and 272 DF,  p-value: < 2.2e-16
```

$\hat{y}$ (estimated bill depth) = _____*(species)

- The slope of the linear model is equivalent to…_____
- Also notice that "Gentoo" is listed in the summary output. That means that the category level "Gentoo" has been assigned to the value __.
- We expect the bill depth of a Gentoo penguin to be _____ on average than if it were an Adelie penguin.
- Additionally, the t-test for the slope is _____ to a two-sample t-test for means.

- **Multiple Linear Modeling**: Modeling with _____ predictors using linear terms.

By creating a model using both species and body mass, we can get an even more accurate understanding of the response variable, bill depth.

- Exploring an **"Additive Model"**
  - An additive model is when the effect of one predictor on the response remains constant, regardless of the value of the other predictor.
  - This means that the additive difference in bill depth between each species remains about the same, regardless of the penguin's body mass.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.9261     0.4134   31.27   <2e-16 ***
body_mass_kg     1.4647     0.1101   13.31   <2e-16 ***
speciesGentoo   -5.3787     0.1846  -29.13   <2e-16 ***
---
Residual standard error: 0.8704 on 271 degrees of freedom
Multiple R-squared:  0.8145,  Adjusted R-squared:  0.8131
F-statistic: 594.9 on 2 and 271 DF,  p-value: < 2.2e-16
```

  - Interpreting the Additive Model Coefficients
    - When fitting models with multiple predictors, the slope values represent the relationship of one predictor with the response while holding the other predictor(s) constant.

For every one kg increase in _____, we expect _____ to be _____ **higher** on average, *if comparing two penguins of the same* _____.

For penguins of _____ we expect _____to be _____ **lower** on average, *if comparing two penguins of the same* _____.

$$\hat{y} = 12.9261 + 1.4647(\text{body mass}) - 5.3787(\text{species*})$$

**Where species = 0 if "Adelie" and 1 if "Gentoo."*

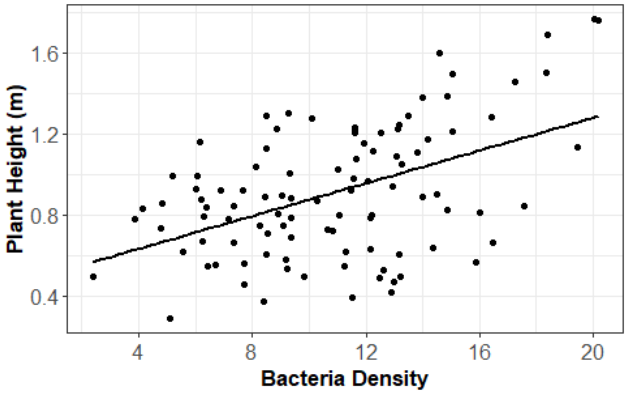
Practice: What would be the model predicted bill depth of a penguin with body mass of 3.8kg and of the species Adelie?

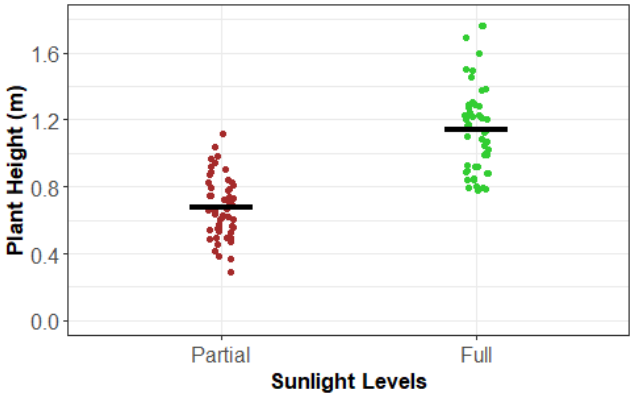
- Exploring an **"Interaction Model"**
    - An interaction model is when one predictor **moderates** the other predictor's effect on the response. The effect of one predictor on the response depends on another.

Investigation: We know that sunlight levels have a positive effect on tomato plant growth. We also know that higher soil bacteria levels tend to increase tomato plant growth as well. A biologist has a theory that more sunlight is the catalyst for bacteria's effect on tomato plant growth. Does the data support that?

Simple Models

❖ Simple models use only one predictor. In each case, we can see that bacteria density levels and sunlight levels are individually associated with larger plant heights.

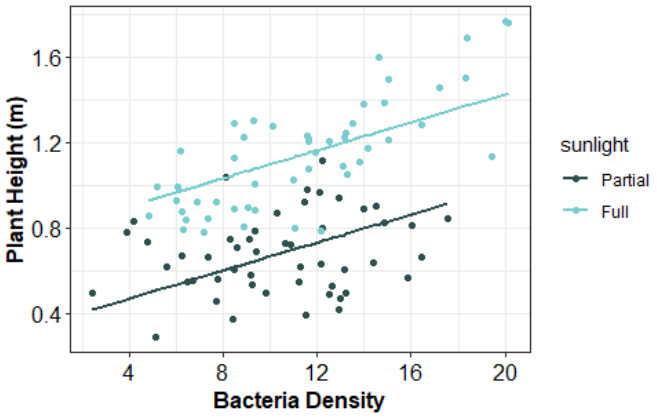




Additive Model

❖ Here, we're creating a model that assumes each predictor works _____.

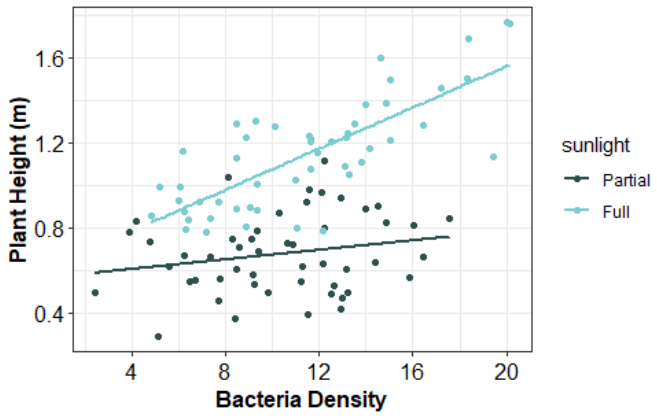
❖ The slope relating bacteria density to plant height is independent of sunlight levels, and the effect of sunlight levels on growth is independent of the bacteria density level.



Interaction Model

❖ With an interaction model, we allow the slopes to be different for each group. The predictors are **dependent.**

In context, what does the interaction model tell us?



Simple Model (for bacteria)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.470371   0.085991   5.470 3.45e-07 ***
bacteria    0.040388   0.007464   5.411 4.45e-07 ***
---
Residual standard error: 0.2856 on 98 degrees of freedom
Multiple R-squared:   0.23, Adjusted R-squared:  0.2222
F-statistic: 29.28 on 1 and 98 DF,  p-value: 4.449e-07
```

Write the Model Equation:

Additive Model (Intercept Adjustment)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.338001   0.057818   5.846 6.77e-08 ***
sunlightFull 0.431144   0.037966  11.356  < 2e-16 ***
bacteria     0.032731   0.004962   6.597 2.22e-09 ***
---
Residual standard error: 0.1881 on 97 degrees of freedom
Multiple R-squared:  0.6695,       Adjusted R-squared:  0.6627
F-statistic: 98.23 on 2 and 97 DF,  p-value: < 2.2e-16
```

Write the Model Equation:

Interaction Model (Intercept and Slope Adjustment)

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.560837   0.077834   7.206 1.31e-10 ***
sunlightFull         0.030529   0.107014   0.285 0.776041
bacteria             0.011194   0.007131   1.570 0.119779
sunlightFull:bacteria 0.037151   0.009366   3.967 0.000141 ***
---
Residual standard error: 0.1752 on 96 degrees of freedom
Multiple R-squared:  0.716, Adjusted R-squared:  0.7071
F-statistic: 80.68 on 3 and 96 DF,  p-value: < 2.2e-16
```

Write the Model Equation:

Inference for Additive and Interaction models

Investigation Revisited: Consider the biologist's original question. How would we use the scatterplots and models to address his original question? If more sunlight is truly a catalyst to bacteria's effect on plant growth, then what should we notice about the relationship between growth in bacteria in full sunlight compared to partial sunlight?

Learning from the Interaction model

Null hypothesis for interaction term:

P-value from interaction term:

Conclusion:

Learning from the Additive model: IF there were no evidence for an interaction term, we could instead judge if we should keep both predictors as independent, additive terms.

Is there evidence that sunlight contributes while controlling for bacteria levels?

Is there evidence that bacteria levels contributes while controlling for sunlight levels?

Adjusted R squared—how much variability are we explaining with this model?

- When adding predictors, multiple r^2 will only increase.
- **Adjusted r^2** is the variability explained in the response variable after adjusting for…
- In cases where the new term performs worse than random chance, adj r^2 _____!

After adjusting for expected correlation due to random chance, how much variability do we estimate is explained by including the interaction term?

Reflection Questions

14.1. What is a dummy variable in a statistical model? Why do we call it that?

14.2. What does it mean to "control for" something in a model?

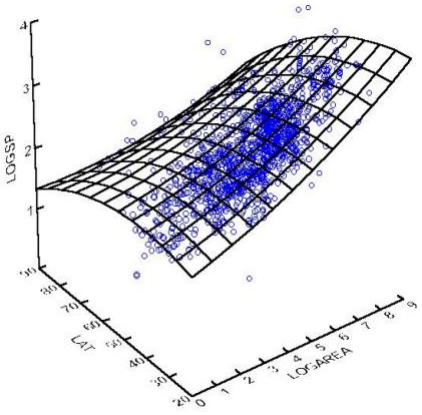
14.3. In the model involving plant growth, what did we mean when we said that the two predictors (bacteria and soil temperature) contributed in interaction to predict growth? What would it mean or look like if the two predictors contributed additively, but **not** in interaction, for modeling plant growth?

14.4. Consider a study where students are given a day to prepare for a quiz on a topic they know nothing about. Half are assigned to use "well-designed" study materials while the other half are assigned to use "poorly-designed" study materials. In addition, students can study for whatever length of time they want, and they report their total study time when they come in to take the test. What would it mean if the researchers found an interaction between material type and study time on students' quiz scores? What might that scatterplot look like?

14.5. What is the difference between adjusted r squared and regular (multiple) r squared? Why is it often more helpful to compare adjusted r squared when comparing models that may have different numbers of predictors?

Adding more Complexity!

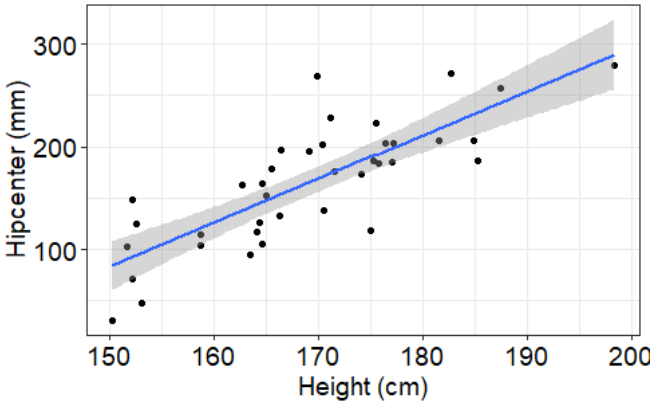
- More dimensions
  - Even though we can't easily see them, we can add more numeric dimensions to our model.
  - You can imagine this idea with 2 numeric predictors, but mathematically, we can continue adding more dimensions.

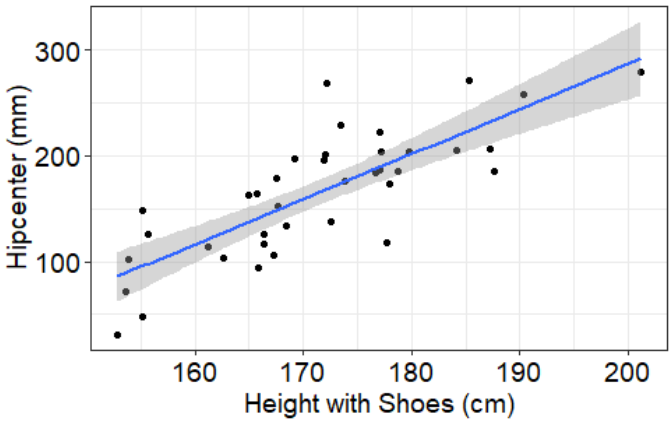


Hamzic (2016).

- Multicollinearity
  - Even though a set of predictors may have individual correlation with the response variable, there may be a multicollinearity issue.
  - **Multicollinearity:** When multiple predictor variables are, themselves, highly correlated and explain mostly the same variance in the response variable.
  - Adding predictors that don't add any additional predictive power to a model can create _____ issues.
  - It may also make it more difficult to interpret the slopes of other predictors.

Seat Distance: Consider a model to estimate someone's preferred distance away from the steering wheel while driving (*distance from wheel to hip center*) based on other physical measures. Two predictor variables we have in our data are Height, and Height with Shoes. We can see that each individually are correlated with seat distance.






```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -556.2553    90.6704    6.135 4.59e-07 ***
Height         4.2650     0.5351    7.970 1.83e-09 ***
---
Residual standard error: 36.37 on 36 DF
Multiple R-squared:  0.6383,  Adj. R-squared:  0.6282
F-stat: 63.53 on 1 and 36 DF,  p-value: 1.831e-09
```

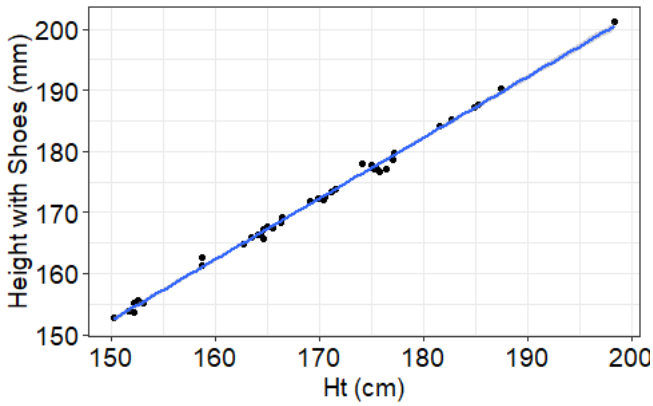
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -565.5927    92.5794    6.109 4.97e-07 ***
HeightShoes    4.2621     0.5391    7.907 2.21e-09 ***
---

Residual standard error: 36.55 on 36 DF
Multiple R-squared:  0.6346,  Adj. R-squared:  0.6244
F-stat: 62.51 on 1 and 36 DF,  p-value: 2.207e-09
```

Model with Both Predictors

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -552.569     95.755   5.771 1.55e-06 ***
Height          5.490      8.918   0.616    0.542
HeightShoes    -1.230      8.938  -0.138    0.891
---

Residual standard error: 36.87 on 35 DF
Multiple R-squared:  0.6385,  Adj. R-squared:  0.6178
F-stat: 30.91 on 2 and 35 DF,  p-value: 1.851e-08
```



What do the predictor p-values communicate to us?

How did adjusted r^2 change? What does that communicate to us?

- Variable Selection
  - Putting predictors together is like building a team—you don't necessarily want the X best all-around players on your team…you want players with different strengths.

  - We care about collinearity among predictors because a good multiple regression model should be…
  - **Parsimonious**: A model that contains as _____ predictors as possible while explaining a reasonable percentage of variance in the Response.
    - You don't want to "spend everything you have" unless it is worth it.
    - Adding redundant variables (e.g., collinear predictors) makes your model more complex and risks overfitting to the sample more than the variables in general.
    - Is the small improvement worth the cost?

  - What each component communicates
    - **P-values for your predictors** judge if each predictor makes any contribution to the model after including the other predictors/terms already present.
    - **Adj. r^2** measures the overall model's predictive power while naturally penalizing models with more predictors.
    - The **F-test p-value** judges if your entire model is performing better than random chance.
      - *In this class, we'll instead focus on predictor p-values for individual predictor contributions and adj. r squared for overall model performance.*

Practice: Let's return to the Seat Distance data again. This dataset explored the ideal seat distance for 38 drivers and captured various physical characteristics. We explored models with 4 predictors (Height, Leg length, Age, and Arm length), starting with the strongest and adding each next strongest predictor.

Model 1

	Estimate	Pr(>|t|)
Ht	4.2650	1.83e-09

Multiple R^2: 0.6383
Adjusted R^2: 0.6282

Model 2

	Estimate	Pr(>|t|)
Ht	2.565	0.0509
Leg	6.136	0.1496

Multiple R^2: 0.6594
Adjusted R^2: 0.6399

Model 3

	Estimate	Pr(>|t|)
Ht	2.3254	0.0725
Leg	6.7390	0.1099
Age	-0.5807	0.1347

Multiple R^2: 0.6814
Adjusted R^2: 0.6533

Model 4

	Estimate	Pr(>|t|)
Ht	2.0765	0.1431
Leg	6.2472	0.1552
Age	-0.7291	0.1584
Arm	1.6160	0.6548

Multiple R^2: 0.6834
Adjusted R^2: 0.6450

If my criteria is to choose the model that explains the most variability in the response after adjusting for correlation likely due to random chance, which model would we prefer?

If my criteria is to choose the model that only includes predictors with p-values below 0.10, which model would I choose?

Which of those two criteria happens to be more parsimonious in *this* situation?

Arm has a high p-value in the fullest model. Does that mean Arm length is not linearly correlated with Preferred Seat Distance?

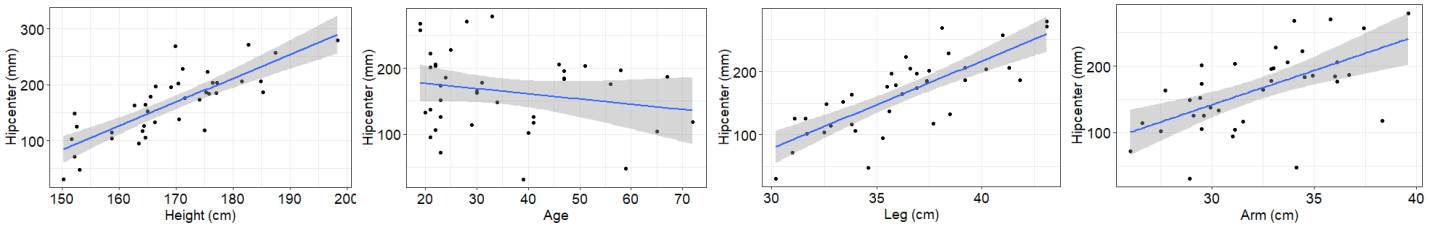
Topics Beyond on Course: *Model Selection Techniques*

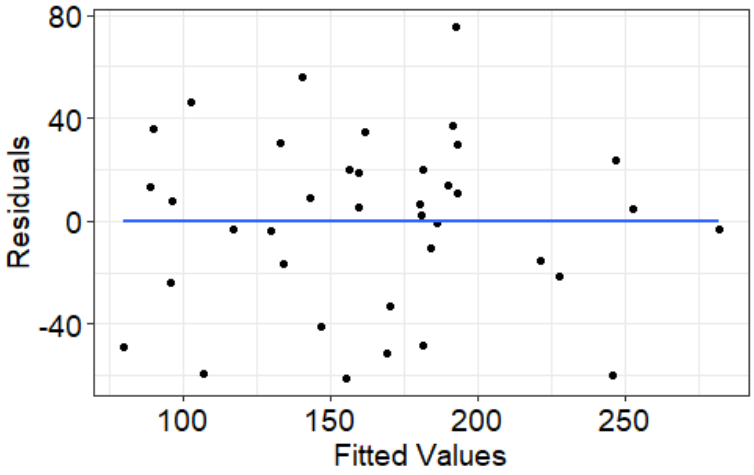
- ✓ While creating and comparing models individually is ok with few predictors, software allows for fast and systematic exploration of possible models (e.g., forward, backward, and step-wise selection methods).
- ✓ In addition to Adjusted r^2, there are several other criteria for comparing models, such as AIC, BIC, and average prediction error using training/test data splits.
- ✓ See this Applied Statistics with R textbook if you'd like to learn more!

Model Diagnostics

- When doing multiple linear regression, the LINE assumptions still apply.
  - **Linearity**
    - Linear terms make sense for a lot of predictor variables, but a linear fit is not always the right fit for every predictor.
    - It's a good idea to plot predictors individually with the response to check. If the **fit is clearly not linear**, it may make sense to complete a **"predictor transformation."**
  - **Independence** of Observations
    - No direct change from Simple Linear Regression.
    - If the **observations are dependent**, you may need a **different modeling approach**
  - **Normality of Residuals**
    - Now that we have multiple predictors, we need a residual plot to visually inspect this.
    - A **residual plot** will compare the size of a data point's residual on the y axis against the model estimated value on the x-axis.
    - We want to see a _____ distribution around the residual = 0 line.
    - If the **residuals aren't normally distributed** about the best fit line, you may need a **"response transformation."**
  - **Equal Variance (Homoscedastic)**
    - This is also best assessed with the residual plot.
    - There should be little to no pattern in the residual plot—no cone shapes or changing variability across fitted values.
    - If the **residuals are heteroscedastic**, you may need a **"response transformation."**

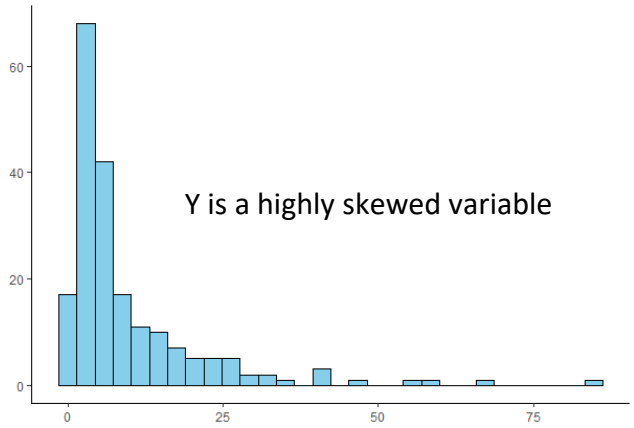
Checking the Seat Distance Model



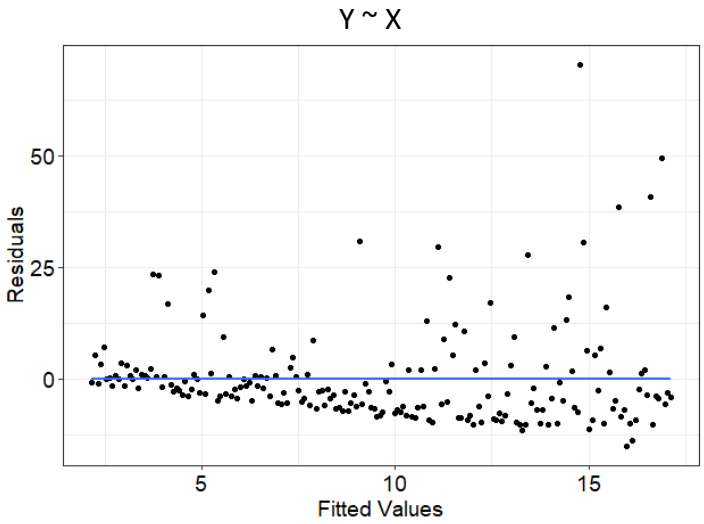


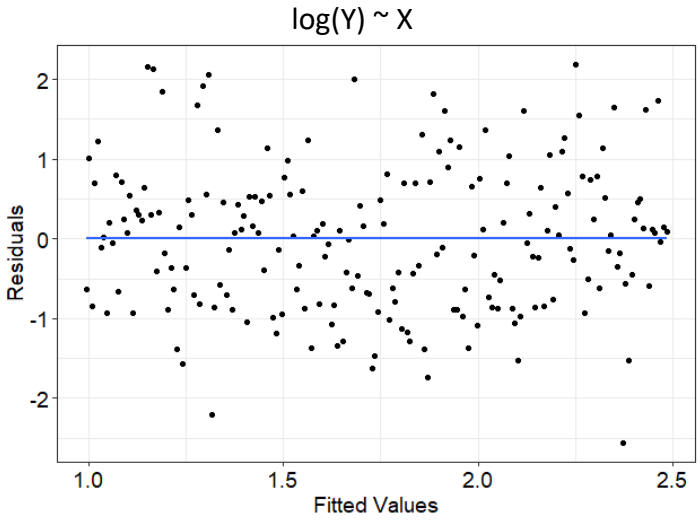
- **Handling Assumption Violations**
  - Assumption violations do ***not*** mean the regression is ruined! It simply weakens the _____ of the results.
    - Violations of normality and homoscedasticity mean that our coefficients could be slightly biased, and the SE and t-test results may be off.
  - Small violations are to be expected and are ok!
    - The _____ the sample size, the less effect violations will have on the regression.
    - But bigger violations among smaller samples can affect results more noticeably.

- **Response Transformations**
  - Transforming the variable means taking some function of it.
  - _____ response variables are sometimes difficult to model without adjustment.
  - After a log transformation on the response variable, the model diagnostics look much better.



Y is a highly skewed variable





Topics Beyond our Course: *Predictor Transformations*

- ✓ In some cases, a predictor variable may be skewed or distributed asymmetrically. Log transformations may be beneficial for a predictor variable as well!
- ✓ Polynomial transformations (e.g., squaring or square rooting a predictor) may also be appropriate when the fit doesn't appear linear.

Reflection Questions

14.6. Consider if we looked at a model summary output with two predictors. Predictor A has a very low p-value, while Predictor B's p-value is not very low at all. What does this result tell us? Would it be fair to say that Predictor B is likely not linearly correlated with the response variable?

14.7. What does it mean to value parsimony when building a model? What is a disadvantage that might arise from ignoring parsimony?

14.8. Is there one best way to select a model?

14.9. What is a residual plot? If the response variable for your model is naturally a skewed variable, how might you use a residual plot to check whether assumptions for linear regression are met?

Chapter 14 Additional Practice (Videos available in the Ch 14 module on Canvas!)

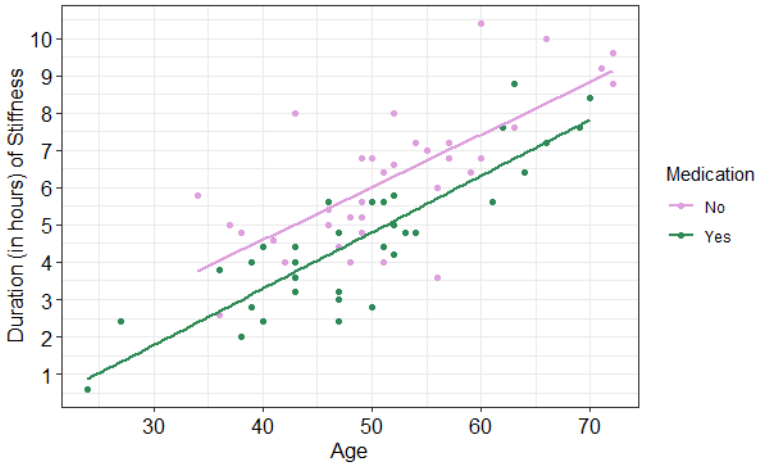
Investigation: A hospital research team is studying an experimental medication in shortening the period of stiffness (in hours) immediately after a non-invasive hand surgery. The research team already knows that the length of time for experiencing stiffness is highly dependent on patient's age. For that reason, they would like to see how much using the medication might decrease stiffness duration after controlling for age. 71 patients were randomly assigned to either medication or no medication. A simple, additive, and interaction model are presented below.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.49237    0.70095  -3.556 0.000686 ***
Age          0.15865    0.01363  11.641  < 2e-16 ***
---
Multiple R-squared:  0.6626,  Adjusted R-squared:  0.6577
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.24535    0.59918  -2.078   0.0414 *
Age            0.14759    0.01112  13.275  < 2e-16 ***
MedicationYes -1.39845    0.22560  -6.199 3.81e-08 ***
---
Multiple R-squared:  0.7844,  Adjusted R-squared:  0.7781
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.488066   0.875901  -1.699    0.094 .
Age                0.152253   0.016561   9.193 1.73e-13 ***
MedicationYes     -0.964724   1.157740  -0.833    0.408
Age:MedicationYes -0.008582   0.022462  -0.382    0.704
---
Multiple R-squared:  0.7849,  Adjusted R-squared:  0.7753
```

After controlling for patient age, is there evidence that the medication decreases stiffness duration?



On average, we would say this medication decreases stiffness duration by how much? (Is this about the same for all ages, or does it depend on age?)

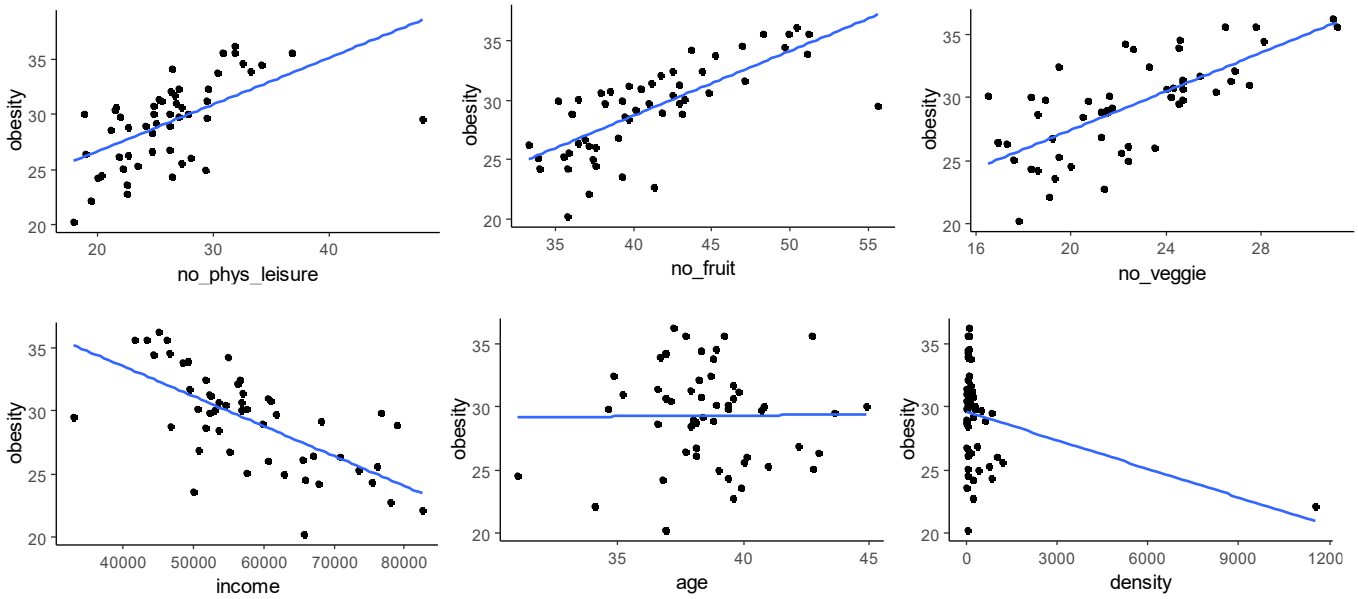
Investigation: To better understand factors that lead to high obesity rates for U.S. states/territories, researchers gathered a large sample of survey data from residents of all 50 U.S. states, DC and Puerto Rico, asking about their diet and exercise habits. The researchers collated survey data by territory and general demographic information to try to explain the obesity rates in each territory.

Unit of observation:

Response variable:

The predictor variables we will focus on here are as follows…

- **no_phys_leisure:** percentage of territory residents who report not having a regular physical activity for leisure purposes
- **no_fruit:** percentage of territory residents who report eating less than 1 piece of fruit each day on average
- **no_veggie:** percentage of territory residents who report eating less than 1 serving of vegetables each day on average
- **income:** median household income in the territory
- **age:** average age among residents in the state
- **density:** population density for the territory (avg number of people per square mile)



Which predictors appear to be linearly correlated to obesity rates?

Are there any predictors that might require a variable transformation before it is suitable to model linearly?

In context, why does obesity rate appear to have a negative correlation to income? Does that make sense contextually?

There is probably a correlation between percent of territory residents who don't eat fruits and percent who don't eat vegetables. Should we only include one of these predictors in our model, or is there evidence that both make a unique contribution? How much more variance do we likely explain with both, as compared to just the strongest one solo?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.83336    2.95982   2.309   0.0251 *
no_fruit     0.54661    0.07151   7.644 5.93e-10 ***
---
Residual SE: 2.629 on 50 degrees of freedom
Multiple R-squared:  0.5389,
Adjusted R-squared:  0.5297
F-stat: 58.43 on 1 and 50 DF,  p-value: 5.926e-10
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1145     2.4616   4.921 9.74e-06 ***
no_veggie     0.7648     0.1083   7.061 4.82e-09 ***
---
Residual SE: 2.739 on 50 degrees of freedom
Multiple R-squared:  0.4993,
Adjusted R-squared:  0.4892
F-stat: 49.85 on 1 and 50 DF,  p-value: 4.824e-09
```

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.13584    2.79421   2.196 0.032862 *
no_fruit     0.34790    0.09868   3.526 0.000927 ***
no_veggie    0.39459    0.14343   2.751 0.008302 **
---
Residual SE: 2.471 on 49 degrees of freedom
Multiple R-squared:  0.6006,
Adjusted R-squared:  0.5843
F-stat: 36.84 on 2 and 49 DF,  p-value: 1.718e-10
```

Consider this model that includes no fruit, no veggie, and income. Is there evidence that median territory income still contributes as a predictor, even after controlling for percentage of territory residents who don't typically eat fruits or vegetables on a particular day?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      17.52      6.239   2.808  0.00718 **
no_fruit       0.02055     0.1187   1.731  0.08983 .
no_veggie       0.3899     0.1391   2.803  0.00729 **
income      -9.384e-05  4.632e-05  -2.026  0.04838 *
---
Residual standard error: 2.397 on 48 degrees of freedom
Multiple R-squared:  0.632,       Adjusted R-squared:  0.609
F-statistic: 27.48 on 3 and 48 DF,  p-value: 1.714e-10
```

Now consider if we add percentage of residents who don't typically get physical exercise. What does the p-value 0.67477 communicate on that line?

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.751e+01  6.293e+00   2.783  0.00774 **
no_fruit         2.329e-01  1.361e-01   1.711  0.09373 .
no_veggie        3.977e-01  1.415e-01   2.810  0.00720 **
income          -9.614e-05  4.705e-05  -2.044  0.04661 *
no_phys_leisure -4.413e-02  1.045e-01  -0.422  0.67477
---
Residual standard error: 2.417 on 47 degrees of freedom
Multiple R-squared:  0.6334,      Adjusted R-squared:  0.6022
F-statistic:  20.3 on 4 and 47 DF,  p-value: 9.095e-10
```

1. There is little evidence that no_phys_leisure is linearly correlated with obesity rate
2. There is little evidence that no_phys_leisure makes a contribution to this model if we already have the 3 other predictors
3. There is little evidence that this model predicts obesity rate beyond what we expect by random chance