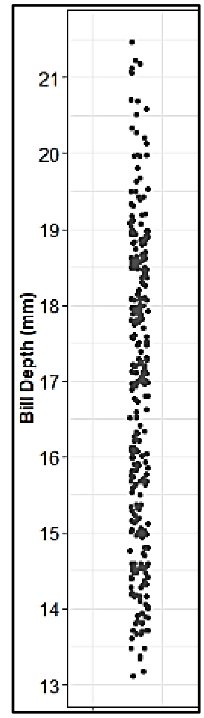
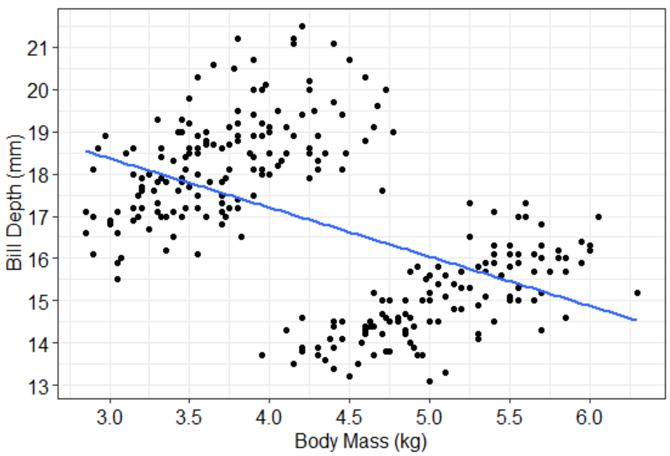
Chapter 12: Multiple Linear Modeling

Introducing Multiple Predictors



Investigation: On Palmer Island, biologists are studying and comparing 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.

Naturally, we would expect penguins with a higher body mass to have deeper bills.

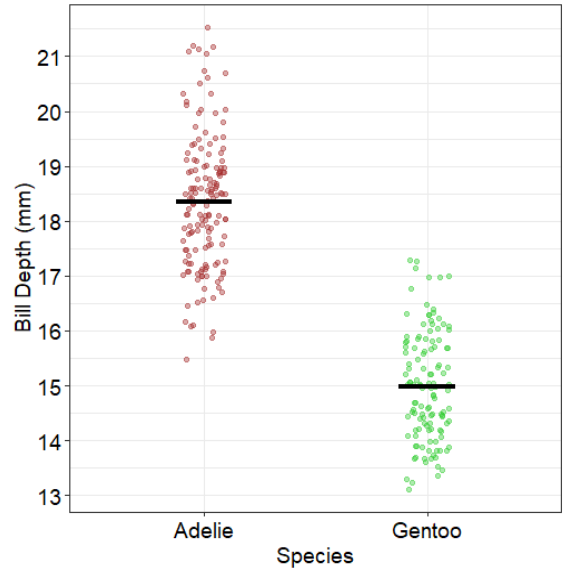



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.0339465  0.5036206   43.75   <2e-16 ***
body_mass_g -0.0011621  0.0001177   -9.87   <2e-16 ***
---

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." It might be that the penguin's species might explain differences in bill depth.



After stratifying by Species, we get the following result

	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 0 and 1.
  - We use the term "dummy" because the assignment of 0 and 1 to each level is arbitrary and carries no contextual meaning.

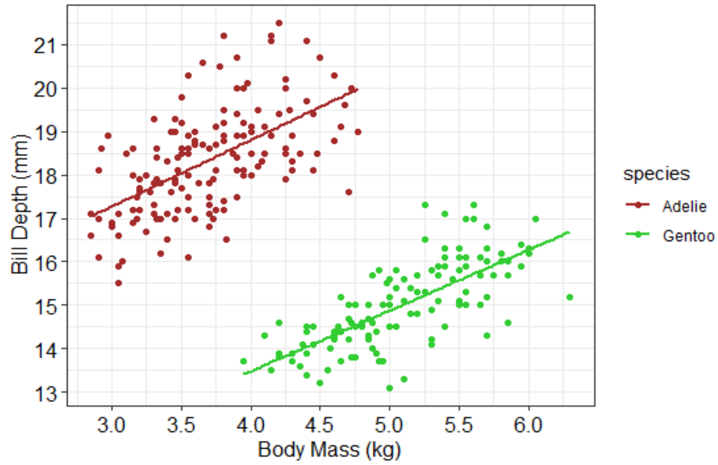
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.34636    0.09092  201.79   <2e-16 ***
speciesGentoo -3.36424    0.13570  -24.79   <2e-16 ***
---

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
```

  - The slope of the linear model is equivalent to…the sample mean difference
  - Also notice that "Gentoo" is listed in the summary output. That means that the category level "Gentoo" has been assigned to the value 1.
  - We expect the bill depth of a Gentoo penguin to be 3.364 mm shorter on average than if it were an Adelie penguin.
  - Additionally, the t-test for the slope is the same as a two-sample t-test for means.

- **Multiple Linear Modeling**: Modeling with 2 or more 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 **body mass**, we expect **bill depth** to be **1.4647 mm higher** on average, *if comparing two penguins of the same **species**.*

For penguins of **species "Gentoo"**, we expect **bill depth** to be **5.3787 mm lower** on average, *if comparing two penguins of the same **body mass.***

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

**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 the effect of one predictor on the response depends on the value of the other predictor.

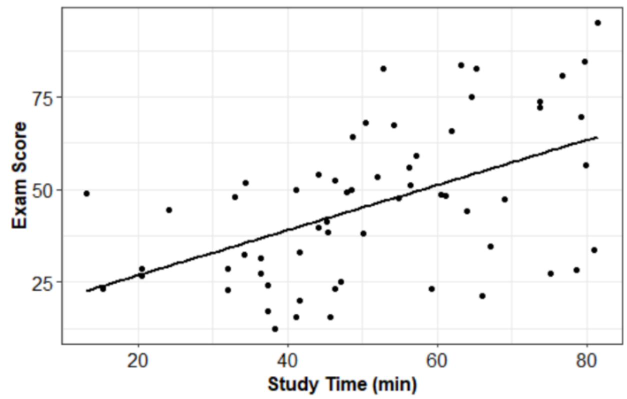
Example: Participants were asked to take a test on a topic that was unfamiliar to them. The response variable is their score on that exam. We have two variables we're going to use to predict their test score:

- ❖ How much time they studied (in minutes)
- ❖ Which study materials they were given (Clear or Unclear)

The "Clear" study materials were carefully structured to benefit students more, whereas the "Unclear" instructions were full of jargon and not very accessible for learning.

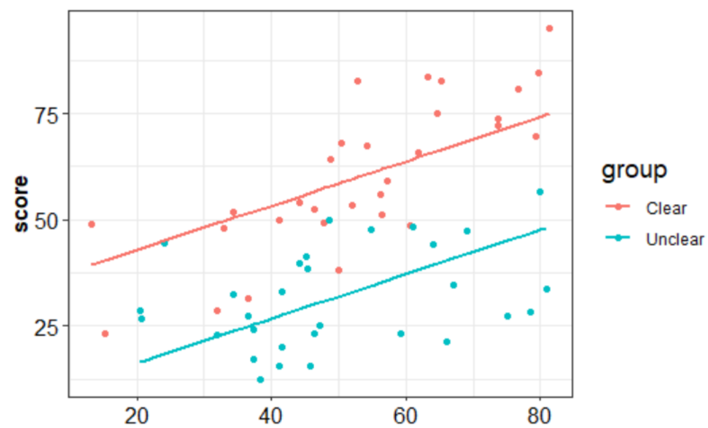
Simple Model

- ❖ This simple model uses only study time as a predictor of exam score.



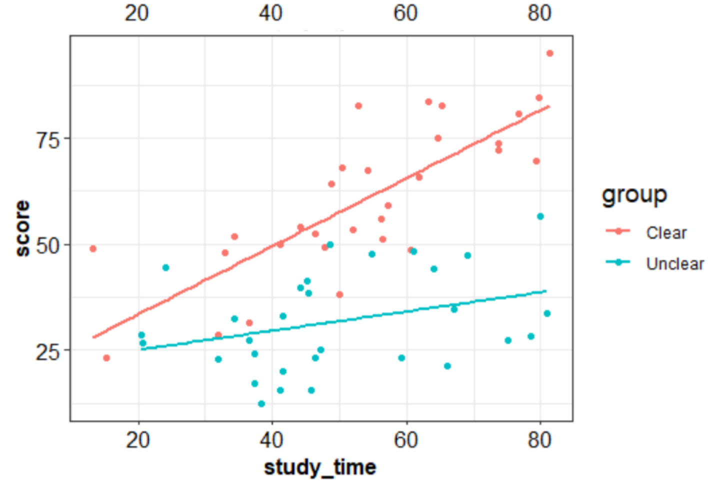
Additive Model

- ❖ Here, we're assuming that each predictor works <u>independently</u> in its effect on exam score.
- ❖ 1) There is a linear relationship between study time and exam score. 2) Students with clearer instructions did better on average than those who didn't. 3) Each predictor's effect on the response is independent of the other.



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?

<u>For unclear instructions: we see a very mild linear relationship between study time and score</u>

<u>For clear instructions: we see a much steeper linear relationship between study time and score</u>

Simple Model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.6439     7.3345   1.997   0.0506 .
study_time    0.6093     0.1352   4.507 3.24e-05 ***
---
Residual standard error: 18.05 on 58 degrees of freedom
Multiple R-squared:  0.2594,  Adjusted R-squared:  0.2466
F-statistic: 20.32 on 1 and 58 DF,  p-value: 3.24e-05
```

Write the Model Equation:

Additive Model (Intercept Adjustment)

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.45560    5.39121   6.020 1.33e-07 ***
study_time     0.52092    0.09191   5.668 5.00e-07 ***
groupUnclear -26.53222    3.16816  -8.375 1.65e-11 ***
---
Residual standard error: 12.19 on 57 degrees of freedom
Multiple R-squared:  0.668,   Adjusted R-squared:  0.6563
F-statistic: 57.33 on 2 and 57 DF,  p-value: 2.259e-14
```

Write the Model Equation:

Interaction Model (Intercept and Slope Adjustment)

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              17.4060     6.6242   2.628  0.01107 *
study_time                0.8026     0.1179   6.805 7.27e-09 ***
groupUnclear              3.1062     9.1483   0.340  0.73548
study_time:groupUnclear  -0.5766     0.1687  -3.417  0.00119 **
---
Residual standard error: 11.19 on 56 degrees of freedom
Multiple R-squared:  0.7252,  Adjusted R-squared:  0.7105
F-statistic: 49.27 on 3 and 56 DF,  p-value: 1.013e-15
```

Write the Model Equation:

- **Inference for Additive and Interaction models**

Judging Interaction Term: To determine if there is truly improvement from the interaction term (rather than just some random chance interaction), look at the p-value for the interaction term only.

Null hypothesis: Coefficient for interaction term = 0

P-value from interaction term: 0.00119

Conclusion: We have very strong evidence of at least some interaction between study time and instructions type.

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

Null hypothesis for study time: No linear relationship between study time and exam score after controlling for instructions type.

P-value and Conclusion: 5x10^-7, strong evidence of relationship, even after using instructions type

Null hypothesis for Instructions: No linear relationship between instructions type and exam score after controlling for study time.

P-value and Conclusion: 1.65x10^-11, strong evidence of relationship, even after using study time

- **Adjusted R squared—how much variability are we explaining with this model?**
    - When adding predictors, multiple r^2 will only increase. For this reason, it's more meaningful to use **Adjusted r^2**
    - **Adjusted r^2** is the variability explained in the response variable after adjusting for… correlation likely due to random chance.
    - In cases where the new term performs worse than random chance, adj r^2 drops!

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

0.7105 – 0.6563 = 0.0542 = **5.42%**

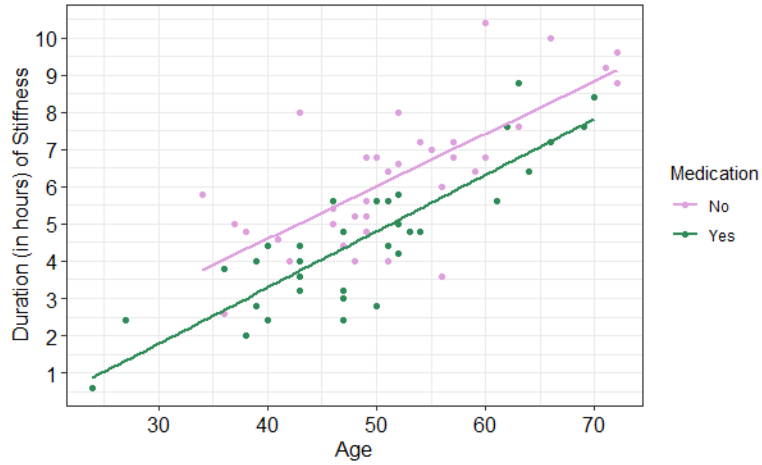
Practice: A hospital research team is studying an experimental medication in shortening the period of stiffness (in hours) immediately after a non-invasive hand surgery. The research team already knows that the length of time for experiencing stiffness is highly dependent on patient's age, so how much effectiveness does the medication have after controlling for age? 71 patients were randomly assigned to either medication or no medication. A simple, additive, and interaction model are run.

```
              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
```

Is there evidence that the medication is effective? Is there evidence that the medication's effectiveness depends on the patient's age?

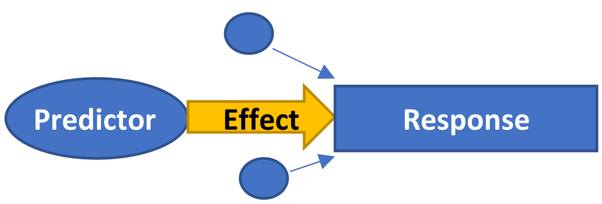


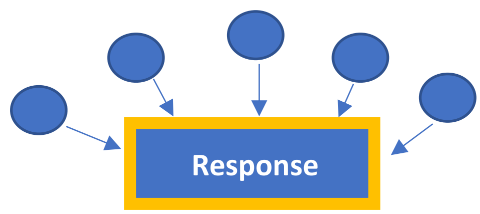
By how much should the duration of stiffness change on average if taking the experimental medication?

- Modeling for Explanation vs. Prediction
  - **Modeling for Explanation** focuses on the individual relationships.
    - If studying whether one specific predictor might reasonably cause changes in the response, we might include other potential confounding variables.
    - This allows us to stratify by other variables, and then observe if there is still a relationship between the supposed causal predictor and the response.
    - Our interest is on the slope value, the p-value for that causal predictor, and how much *improvement* we see in r^2 with its addition.

  - **Modeling for Prediction** focuses on raising r^2 without overfitting.
    - We want to include as many predictors as we have available, but filtering out any redundant or non-correlated predictors.
    - Our interest is in raising the accuracy of our predictions by finding the highest r^2 value without overfitting.

Modeling for Explanation	**Modeling for Prediction**
How effective is this medication at reducing LDL cholesterol after controlling for other known effects for high LDL cholesterol?	Can we create a model to predict LDL cholesterol accurately using easy-to-collect, non-invasive variable measures?





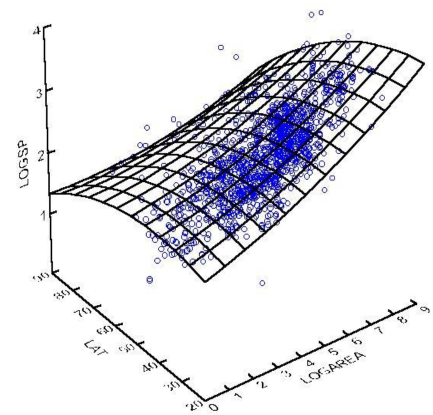
Which type of modeling approach is represented in this question: Does age affect stiffness duration? How much effect does it have after stratifying by patient's use of medication?

Modeling for explanation

Modeling for prediction might ask: "Which factors help us more accurately predict stiffness duration?"

Adding more Complexity!

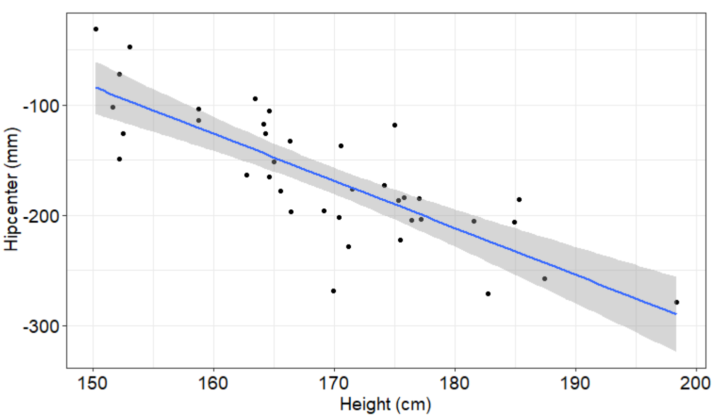
- More dimensions
    - Even though we can't easily see them, we can add more numeric dimensions to our model.
    - You can imagine this idea with 2 numeric predictors, but mathematically, we can continue adding more dimensions!

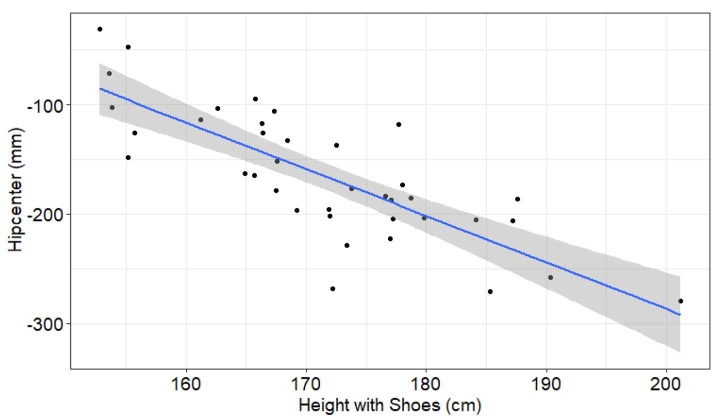


Hamzic. https://dzenanhamzic.com/2016/08/03/linear-regression-with-multiple-variables-in-matlab/

- 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
    - Multicollinearity is a ***big concern*** with ***modeling for <u>explanation</u>***—if done carelessly, the coefficient estimates will be unreliable.
    - Multicollinearity is a ***smaller concern*** when ***modeling for <u>prediction</u>***—we just don't want to overfit the model. Overfitting means being too sensitive to our sample of data and modeling noise rather than signal.

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 ***
Ht           -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
```

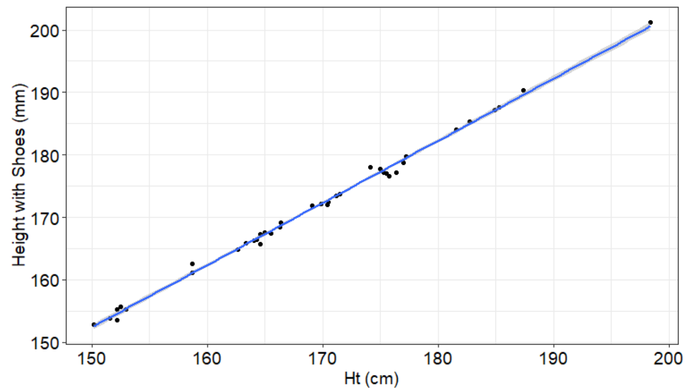
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 565.5927    92.5794   6.109 4.97e-07 ***
HtShoes      -4.2621     0.5391  -7.907 2.21e-09 ***
---

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 ***
Ht            -5.490      8.918  -0.616    0.542
HtShoes        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
```



…But is there value to including both in a model together? Contextually, what is going on with these predictors?

<u>Explaining the same variance. The two predictors are themselves highly correlated</u>

<u>Adj r^2 went down, and the individual p-value terms are not at all low. No evidence that we gain predictive power by including both.</u>

- 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 <u>few</u> 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 or difficult variables makes your model harder to use and interpret.
        - 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. Comparing adj. r^2 across models helps us measure model improvement with new terms.
        - The **F-test p-value** judges if your entire model is performing better than random chance (we will largely ignore this in our class!)

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 explore a multiple linear regression 4 predictors (Height, Leg length, Age, and Arm length), and eliminate the weakest predictor at each stage.

Model 1	Model 2	Model 3	Model 4
Estimate Pr(>|t|) Ht -4.2650 1.83e-09 --- Multiple R^2: 0.6383 Adjusted R^2: 0.6282	Estimate Pr(>|t|) Ht -2.565 0.0509 Leg -6.136 0.1496 --- Multiple R^2: 0.6594 Adjusted R^2: 0.6399	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	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

Which model seems to be explaining the most variability after adjusting for correlation likely due to random chance?

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

How confident are we that Age makes a unique contribution in Model 3 after including Leg and Height?

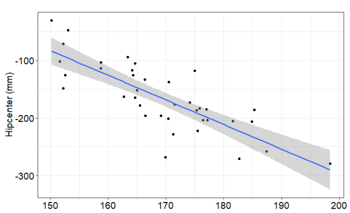
Advanced Model Selection Techniques

- ✓ 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, average prediction error, and cross-validation methods.

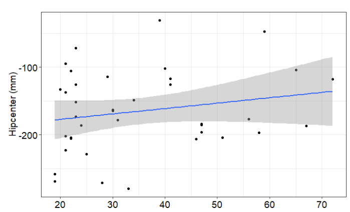
Model Diagnostics

- 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 <u>residual plot</u> to visually inspect this. We want to see a mirror-like 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."**

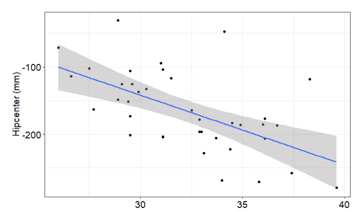
Checking the Seat Distance Model



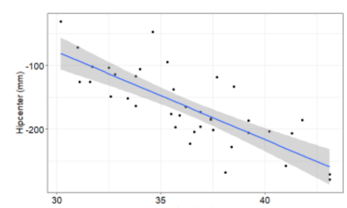
Height



Age



Arm



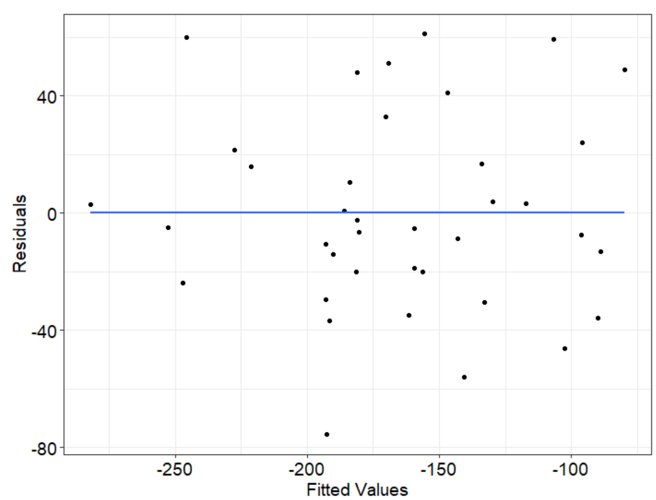
Leg

<u>No non-linear trends obvious</u>

<u>38 independent drivers</u>

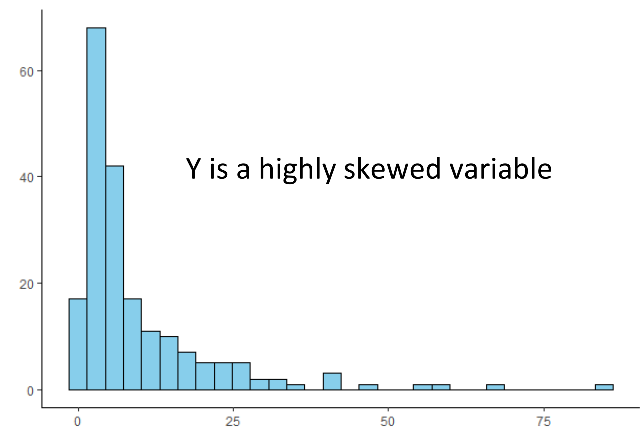
<u>No obvious skew in the residual plot</u>

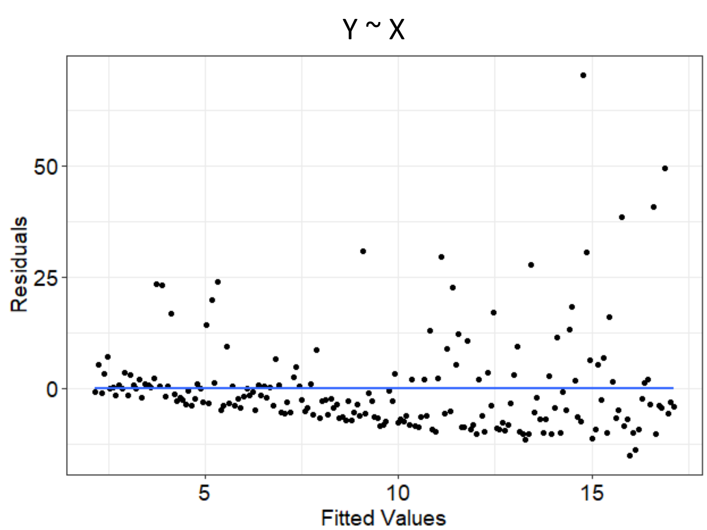
<u>Slight open cone, but may be from small sample size</u>

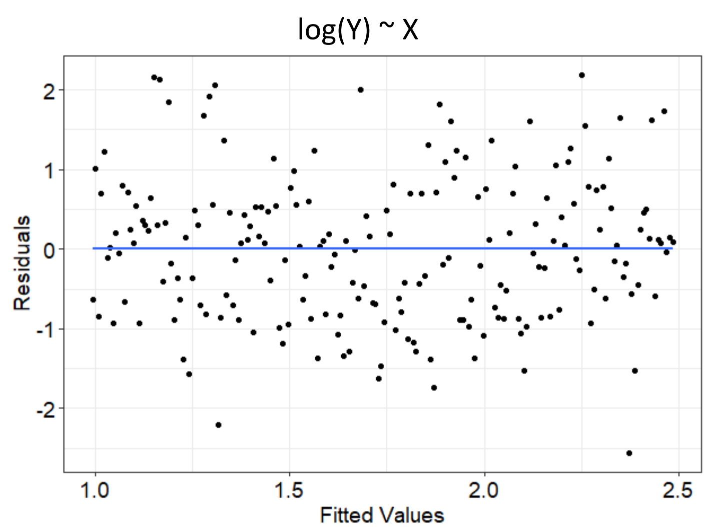


- **Handling Assumption Violations**
  - Assumption violations do ***not*** mean the regression is ruined! It simply weakens the reliability of the results.
    - Violations of normality and homoscedasticity mean that our coefficients could be slightly biased, and the SE and t-test results may be off.
  - Small violations are to be expected and are ok!
    - The larger 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.
  - Highly skewed response variables are sometimes difficult to model without adjustment.
  - In the example below, we see a residual plot showing non-normal and heteroscedastic residuals.
  - After a log transformation on the response variable, the model diagnostics look great!







Some examples of response transformations include:

- ✓ A logarithm (log) transformation
- ✓ A square root transformation
- ✓ A Power transformation ("Box-Cox" Method)

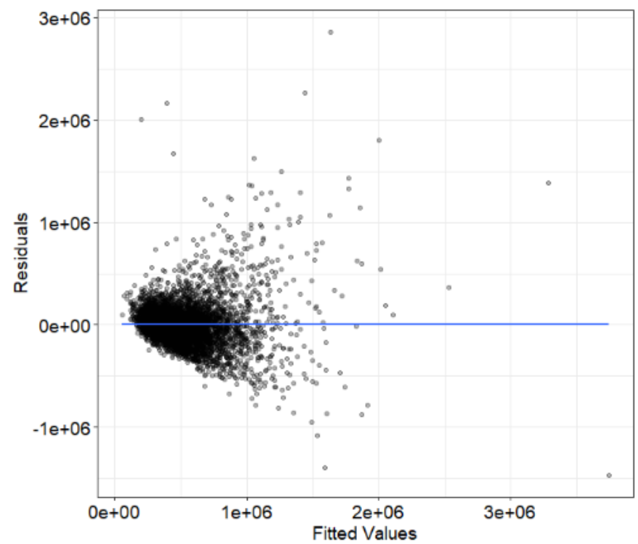
Predicting Housing Prices. Using data from 4,548 homes in the Seattle area, we are trying to better predict the prices that homes are selling for using easily-accessible variables.

The researchers start with 3 key predictors: 1) number of **bedrooms**, 2) **sqft_living** space, and 3) whether the house is in **Seattle** city limits

Below is a summary of the model, and a residual plot.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15109.213  14801.896  -1.021    0.307
bedrooms    -41933.841   4933.537  -8.500   <2e-16 ***
Seattleyes  169692.506   7748.552  21.900   <2e-16 ***
sqft_living    303.771      4.755  63.879   <2e-16 ***
---

Residual standard error: 240900 on 4544 degrees of freedom
Multiple R-squared:  0.5406,  Adjusted R-squared:  0.5403
F-statistic:  1783 on 3 and 4544 DF,  p-value: < 2.2e-16
```

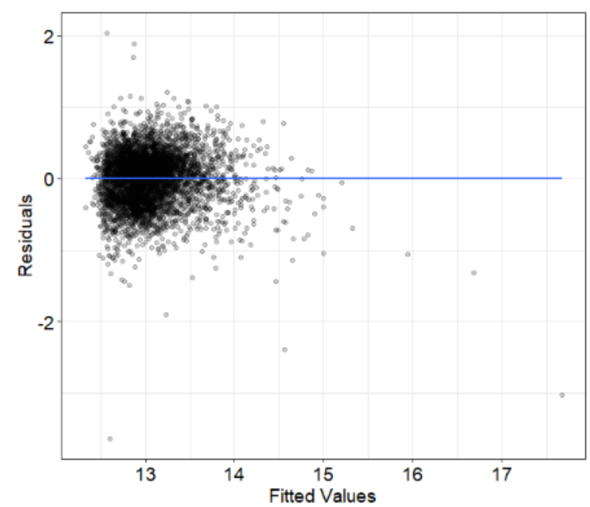


Any assumption violations we should notice with this model?

It turns out that price is a highly skewed variable, so let’s instead try modeling the log of price.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)       1214  2.260e-02 537.204  < 2e-16 ***
bedrooms      -0.03165  7.531e-03  -4.203 2.69e-05 ***
Seattleyes      0.3004  1.183e-02  25.392  < 2e-16 ***
sqft_living  4.356e-04  7.260e-06  60.011  < 2e-16 ***
---

Residual standard error: 0.3678 on 4544 degrees of freedom
Multiple R-squared:  0.5293,  Adjusted R-squared:  0.529
F-statistic:  1703 on 3 and 4544 DF,  p-value: < 2.2e-16
```



Write the equation of the log model

In general, houses with more bedrooms tend to have higher values**. Why might bedrooms have a negative coefficient in these models?**

Chapter 12 Learning Goals

After this chapter, you should be able to…

- Define Multiple Linear Modeling as a modeling technique that uses multiple predictors and uses linear terms.
- Recognize that categorical predictors may be used in linear regression by transforming them into "Dummy" 0–1 variables.
- Understand that the "slope" of a dummy variable simply represents the difference in means between the 0 and 1 categories.
- Interpret slopes in the context of multiple predictors (i.e., expected rate of change while holding other predictors constant)
- Distinguish between additive and interaction models and contextually make sense of what it means when two predictors interact in their prediction of the response
- Interpret adjusted r^2 in context and recognize its advantage over (multiple) r^2 when comparing models with several predictors.
- Use an R Model summary to draw information about a model and make judgments
    - Identify the model equation from the estimates column
    - Use the t-test information in an R output to determine the degree of evidence that a particular predictor or term is making a unique contribution
    - Identify r^2 and adjusted r^2
- Distinguish between questions that are best addressed by modeling for an explanation vs. modeling for prediction
- Define multicollinearity and recognize situations in which two predictors may be highly correlated with one another.
- Explain what it means to be parsimonious when choosing a model
- Develop awareness that model comparison is not an objective process and that there are several criteria (e.g., adjusted r^2, predictor p-values) one may use to select a model.
- Use a residual plot to visually recognize obvious violations to normality and equal variance
- Recognize skewed response variables as often prone to assumption violations and understand how response variable transformations can be used to address assumption violations.