Chapter 13: Simple Linear Modeling

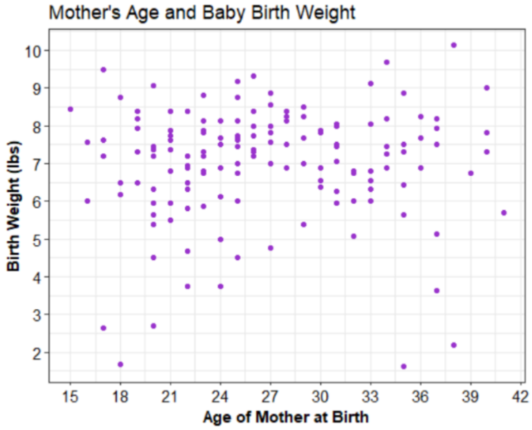
Building a Model

- What is modeling?
  - In statistics, we might often tackle a multivariate investigation by building a model
  - Modeling is the process of making predictions for a response variable based on one or more predictor variables we might have access to.
  - Models are particularly helpful when we are working with a numeric predictor variable, as we can sensibly relate our predictor and response variable together in the form of an equation.

What kinds of relationship might we notice when modeling the relationship between two numeric variables?

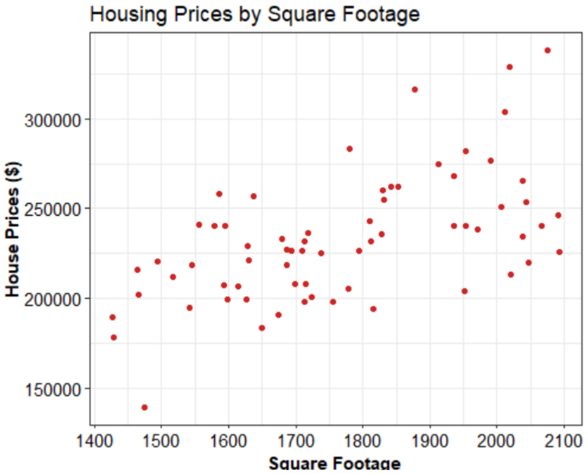
Linear Relationships

As the predictor value increases, the response value tends to change at a constant rate.



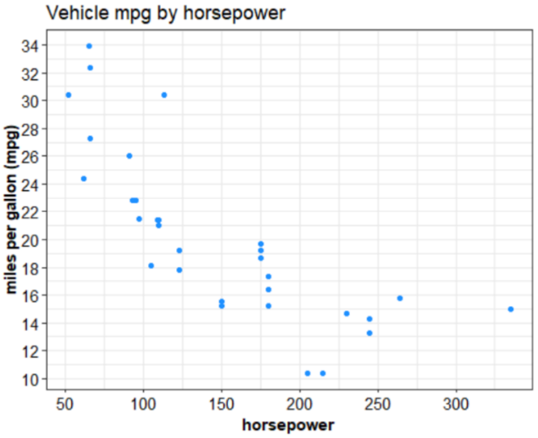
Non-Linear relationships

As the predictor value increases, the response value tends to change at a non-constant rate.



No Relationship

As the predictor value increases, the response value expresses no discernible trend



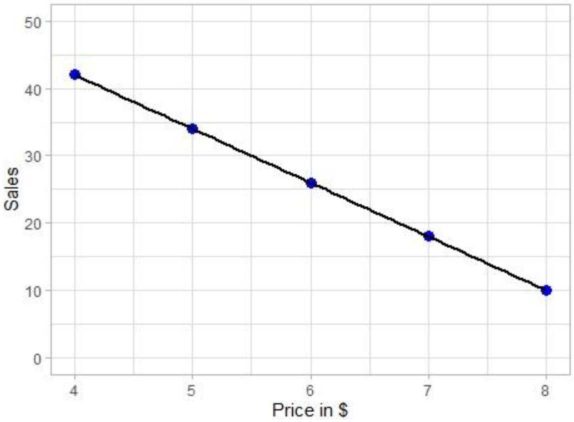
For each scatterplot, identify whether you think the relationship looks linear, non-linear, or if there is no discernible relationship.

Investigation: A vendor sells candles at maker's market. She'd like to better understand how the number of candles she sells might relate to the price she sets on each candle. She decides to collect data on this for 5 weekends in a row. Each week, she changes the price to a different value and records the number of candles she sells. The data is presented below.

Unit of observation: One weekend

Response variable: Number of candle sales (numeric)

Predictor variable: Price (numeric)



Price (X)	Sales (Y)
$4.00	42
$5.00	34
$6.00	26
$7.00	18
$8.00	10

Modeling a Linear Relationship

- In our case, this model appears to be linear—in fact, this data here provides a *perfectly* linear fit!
- But how do we represent this model with an equation?
  - **Slope** tells you the rate at which the response variable changes with respect to unit changes in the predictor.
    - For every one unit increase in (predictor), we expect (response) to be (slope) units higher / lower *on average.*

Let's fill it in and interpret the slope for this example in context:

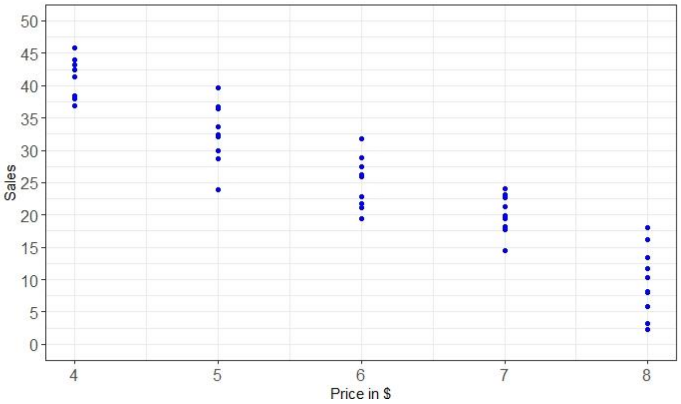
For every one unit increase in price, we expect sales to be 8 units higher / lower *on average.*

  - **Intercept** provides you a starting point/positional reference—the model's approximation for the response value when the predictor variable is at 0.

Equation of a line: $y = \underset{\boxed{\text{Intercept}}}{74} - \underset{\boxed{\text{Slope}}}{8}x$

Practice: If the price were $4.50, then according to this model, we'd expect the # of weekly sales to be what?

Investigation Revisited: Now consider if the store owner had collected data for 50 weekends. For each of the five price points she explored, she had 10 independent weekends of data at that price point.



Why would sales still vary on weeks where the price was the same?

There are other things that must affect sales besides the price!

Weather, crowd size, time of year, variability of who comes and shows interest

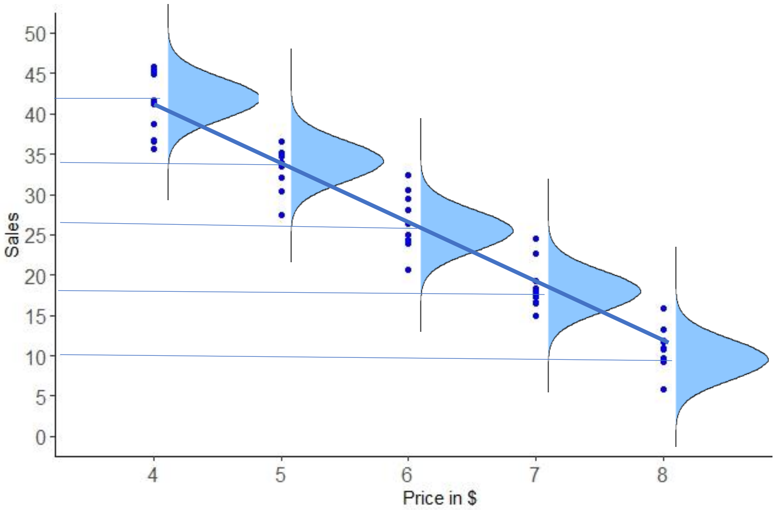
- **Grappling with Uncertainty**
  - When we had a perfect linear relationship, we could build a model that predicted Y from X with perfect accuracy. We used simple mathematics to find that equation.
  - When a relationship is not perfect, we now have uncertainty in our ability to predict Y from X. We now need to use statistics to find this equation and estimate our uncertainty!
  - One common strategy in this case is to model the mean of the Y at each value of X. This method is known as linear regression since we are regressing the relationship toward the mean!

There are two sources of uncertainty when we do this!

1) Since these two variables are not in a perfect relationship, data points will vary around the mean at each cross-section of the predictor.

2) Our best fit line can only estimate the mean at these cross-sections since we only have a sample of data.

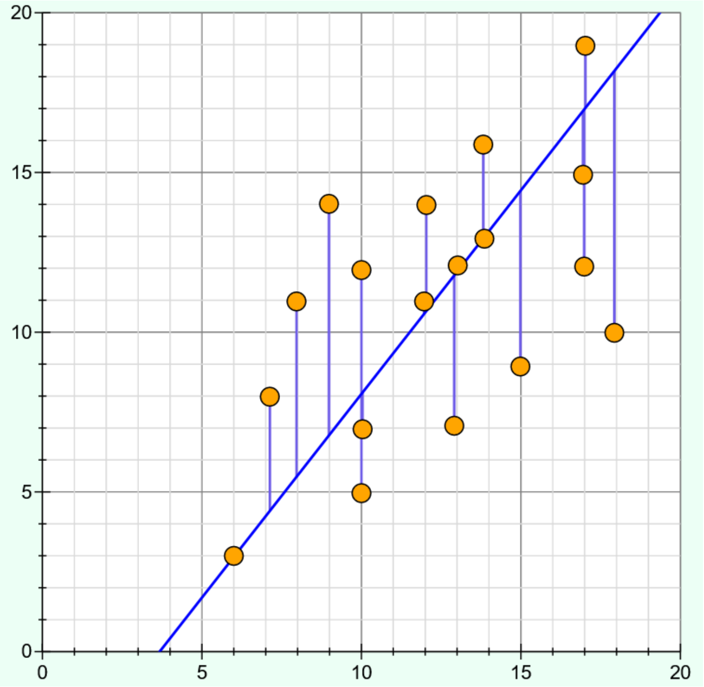
The Equation for the Best Fit Line: $\hat{y} = b_0 + b_1x$



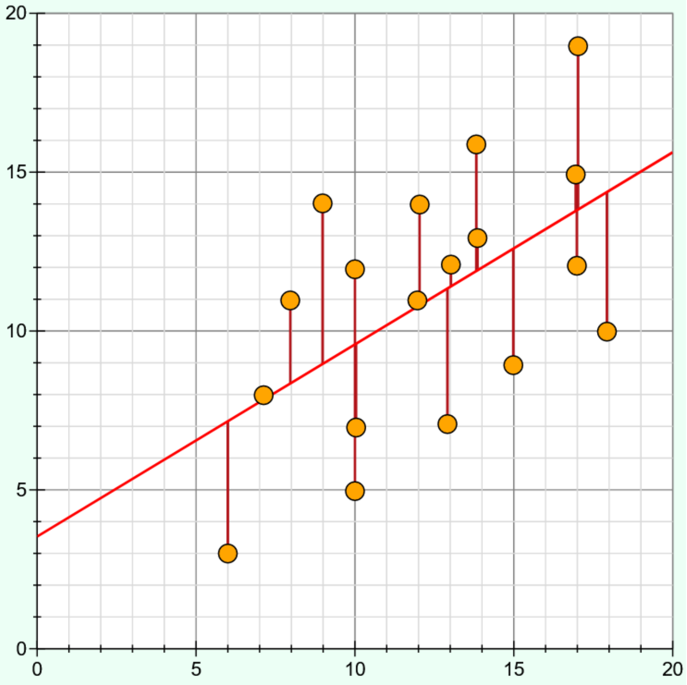
Modeling a Linear Relationship *with Uncertainty*

Mini investigation: Let's say that we had asked 18 students to take a language proficiency exam. The first part of the exam (scored out of 20 points) involves a reading/comprehension portion. The second part of the exam (also scored out of 20 points) involves an interactive speaking/listening activity with a native speaker who scores each individual using a rubric. We'd like to use this data to see how well someone's reading/comprehension score might predict their speaking/listening score.

Each plot below represents the same data, but with two different potential lines of best fit.



$\hat{y} = -4.69 + 1.28x$



$\hat{y} = 3.53 + 0.60x$

Which line do *you* think best represents the relationship? And why?

Answers will vary. Left line fits several individual data points (About 4 of them) better, but it does seem to have a lot more error for more of the data points.

The line on the right seems to have lower error on average

Potentially draw on the point that we don't want to just fit our sample data closely, but rather to project the general relationship. That's what's useful in the larger scheme of things!

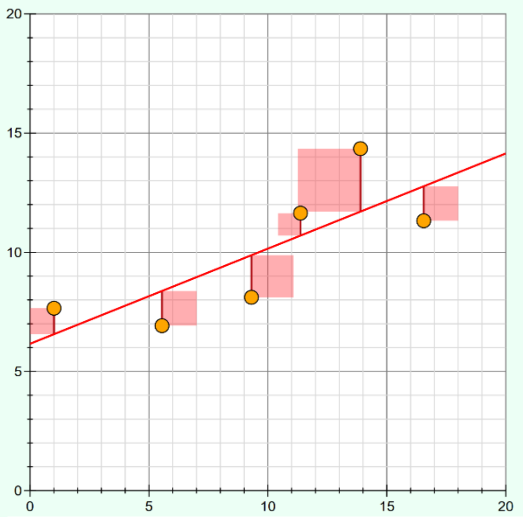
Read/Try on your own

Residuals and the Least Squares Criterion

- When we fit a model, every data point likely has at least some residual error in relation to the model.
- A **Residual** represent the distance between an actual observation of the response and a prediction of the response based on a model equation.
- Let's now introduce some symbolism to showcase this uncertainty!
  - $\mathbf{y}_i$ is an **actual** response value paired up with $\mathbf{x}_i$
    - Perhaps one student scored a 11 on reading comprehension (x)and 13 on speaking/listening (y), making their data point (11,13)
  - $\hat{\mathbf{y}}_\mathbf{i}$ represents the **model predicted** response value given that $\mathbf{x}_i$ is the observed predictor.
    - For a reading comprehension score of 14, our model predicts a speaking/listening score of 3.53 + 0.60(11) = 10.13 score
  - We calculate the **residual** for observation i as $\mathbf{y}_\mathbf{i} - \hat{\mathbf{y}}_\mathbf{i}$

Practice: What is the model's residual error in predicting the score of this student?

- How do we use residuals to choose a model equation?
  - There is not one "correct" method for choosing a model equation, but a common approach is the **"Ordinary Least Squares"** method which relies on the **least-squares criterion.**
  - The least squares criterion selects the line that minimizes the sum of the squared residuals *(this is mathematically advantageous in comparison to minimizing the sum of absolute value deviations).*



PhET Least Squares Regression: https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

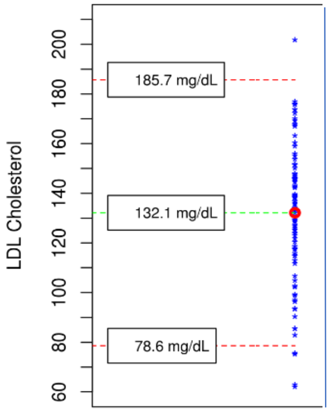
Digging Deeper

Determining the equation that minimizes squared residuals involves calculus. If you're interested to learn more, check out Chapter 7 of this book on applied statistics by my colleague Dr. Dave Dalpiaz! https://book.stat420.org/simple-linear-regression.html

Linear Regression Inference – Judging the Model's Prediction Accuracy

Investigation: Low-density lipoprotein (LDL) Cholesterol is often referred to as "bad cholesterol" that can create blood clots in your blood vessels. Doctors have noted an important association between weight levels and LDL levels. However, weight is definitely not the only thing that explains different cholesterol levels. We collected data from 92 adult males to see how well we could predict LDL from weight.

How **accurately** we can estimate one's LDL cholesterol level when using their weight as a predictor?



- Variance ***before*** building the model
  - ○ Our response variable, LDL cholesterol level, varies from person to person. Without a predictor, our best guess for someone's LDL level would just be the mean LDL level in our sample.

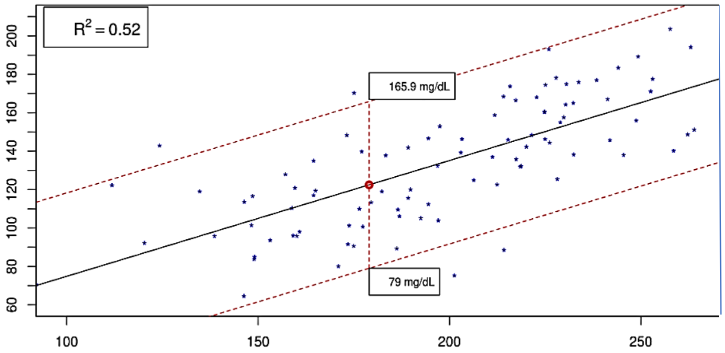
$\hat{y}$ (estimated LDL) = 132.1

So how much error should we expect using this approach?

- ○ The **standard deviation** of the response variable is $\mathbf{s_y = 27.1}$, which we hope is a reasonable estimate of the parameter σ_y
  - ▪ This represents the expected deviation of a randomly chosen individual from the mean.
- ○ More commonly, we'll use the variance of the response variable $\mathbf{s_y^2 = 734.41}$, which we hope is a reasonable estimate of the parameter σ_y^2 *(remember variance is mathematically simpler!)*

- Variance remaining ***after*** building the model
  - ○ Now, let's consider the accuracy of a model that makes a more targeted prediction of one's LDL cholesterol level based on their weight.

$\hat{y}$ (estimated LDL) = 5.25 + 0.65(weight)



- ○ Next, we need a way to measure the variation in our prediction errors using $\hat{y}$ rather than $\bar{y}$.
- ○ After fitting the model, we find the **residuals** to have a **standard deviation** of $\mathbf{s_e = 18.78}$
  - ▪ We hope is a good estimate for σ_e
- ○ We can also represent this in terms of the **residual variance** as $\mathbf{s_e^2 = 352.86}$
  - ▪ We hope this is a reasonable estimate of the parameter σ_e^2.

Practice: How might we use these values to estimate the improved accuracy in our predictions by using this model in comparison to simply using the mean LDL level?

Some may simply subtract one value from the other. Some may think about percentage of variance/SD left.

- **Coefficient of Determination: r^2**
  - The coefficient of determination (abbreviated "**r^2**") is… the proportion of total variability in the response variable that is "explained" by this predictor *(or by these predictors/this model)*
    - s_y^2 measures the **total** variance in the response variable
    - s_e^2 specifically measures the **residual** variance (the **leftover variance**) after applying our model.
  - We could calculate s_y^2 - s_e^2 measures **the variance that is explained by our model**. But this raw difference is not particularly helpful.
  - Let's go one step further and find what proportion of the total variance we have explained by this model by finding this difference as a ratio of the total. *This is an approximate formula for r^2 that doesn't account for degrees of freedom adjustments. Software can take care of that!*

$$\mathbf{r^2 \approx \frac{s_y^2 - s_e^2}{s_y^2}}$$

Let's calculate the r^2 for our LDL cholesterol model.

Which statement is correctly interpreting what we found?

1. Approximately 52% of these men have LDL cholesterol levels above the mean

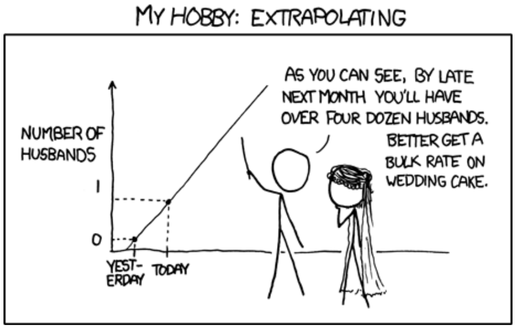
2. Mens' LDL cholesterol levels are approximately 52% of their weight

3. We can reduce the variance in our prediction of LDL cholesterol by approximately 52% when using weight as a predictor

4. The probability of observing a linear association at least this strong by random chance is approximately 52 %

Interpolation and Extrapolation

- **Interpolation:** Predicting Y based on an X value within the range of X values observed
- **Extrapolation**: Predicting Y based on an X value outside of the range of X values observed.
  - While making predictions immediately outside the range is generally safe, making predictions well out are often unreliable.


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

Consider the candle vendor from earlier. What happens to our model estimates when we plug in a price of $10?

We can't have negative sales!

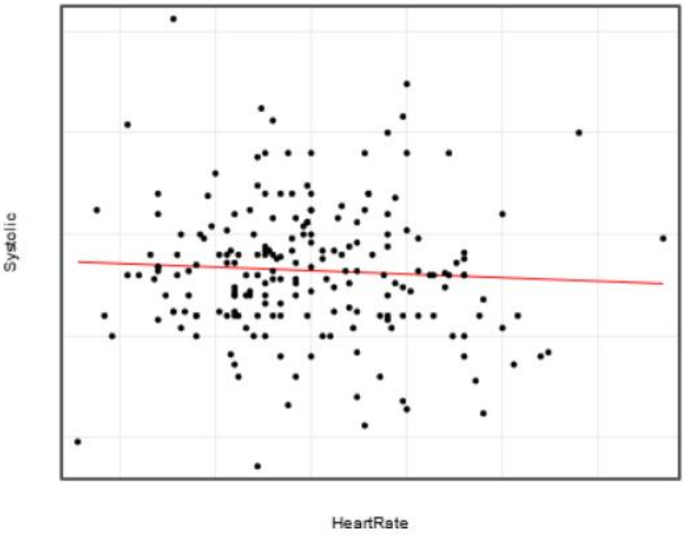
Linear Regression Inference – Judging the Model's Coefficient Accuracy

Investigation: A nurse is looking at the records of patients who were admitted to the Intensive Care Unit (ICU) at his hospital. He wants to know whether patients' heart rate (bpm) upon arrival might be correlated with their systolic blood pressure reading. He plots the data and finds a sample slope of $b_1 = -0.069$.

Unit of observation: One person admitted to ICU

Response variable: Sysolic BP

Predictor variable: Heart rate



But does that suggest heart rate is correlated with systolioc BP? Let's explore this question with a permutation test!

- **Permutation Test** for Linear Regression
  - If there were truly no relationship between one's heart rate and systolic BP, then we could redistribute the heart rate values randomly to each patient.

What is our null and alternative hypothesis? Let's phrase it in terms of β_1

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Go to the **Lock5 StatKey** site linked here:

- https://www.lock5stat.com/StatKey/randomization_2_quant/randomization_2_quant.html
- From the drop-down data menu, choose "ICU Admissions"
- Change the randomization dotplot from "Correlation" to "Slope"

Play around and explore for a bit! Why do the sample slopes tend to congregate around 0?

0, since we are permuting the X values randomly, there is no correlation. Equally likely to be positive or negative slope by chance.

Generate several thousand permutations. Then check "Two Tail." Click one of the x-axis labels to customize the range to our own sample slope of -0.069.

How often do we see sample slopes at least as unusual as ours by random chance? How might that help us answer our investigation?

About 32.6% of the time (16.3% each direction). This tells us we don't have evidence of a correlation. Could easily be random chance.

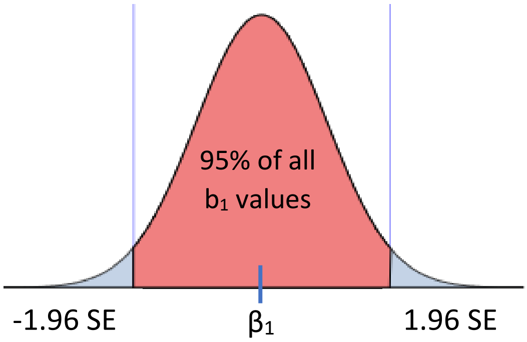
- **The Standard Error for b_1**
    - The Standard Error for b_1 (σ_{b1} or SE_{b1}) is the expected deviation of b_1 from $\underline{\beta_1}$.

$$\sigma_{b1} = \frac{\sigma_e}{\sigma_x\sqrt{n}} \approx \frac{S_e}{S_x\sqrt{n-2}}$$

        - σ_e represents the true standard deviation in the residuals.
        - σ_x represents the true standard deviation in the predictor (X) variable.
        - n represents the sample size

Note: When using s_e and s_x we lose 2 degrees of freedom rather than just 1.

    - We won't be calculating this one by hand in this class—we'll just start from the value and go from there!
        - Notice in our data investigation that the Standard error is approximated in the simulated environment.
        - It may vary depending on your simulations, but it should be around…0.071
- **Parametric Testing/Interval Options**
    - When distribution of possible b_1's is normally distributed about β_1, we can take a shortcut to simulations and simply complete a t-test!
        - Z-test is also fine in larger sample contexts, like df > 120
    - Likewise, if we simply wish to estimate a range of plausible values for β_1, we could complete a t-interval using b_1 as our point estimate.



Investigation revisited: Let's now compare our simulated p-value to what we might get using a z-test. Assume our Null Model is normally distributed with a mean of 0 and a standard deviation of 0.071.

What is the standardized position (z-score) of our sample slope in this null model?

Let's use the Normal Distribution Calculator to find the p-value: https://istats.shinyapps.io/NormalDist/

Finally, let's report a 95% confidence interval for β_1 as well.

Point Estimate:

SE:

Margin of error:

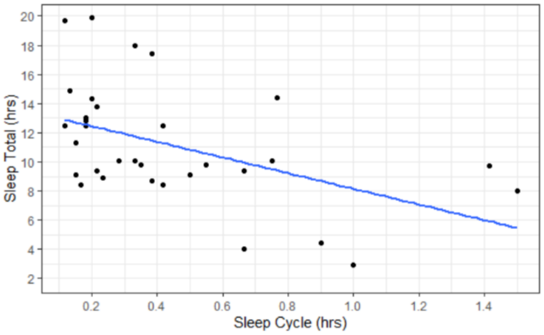
Read/Try on your own

Reading R Output

- When using R to run a linear model, you can find several important values in the model summary.
  - Use the estimate column to identify your model equation
  - You can find the standard error, t-score, and p-value of your slope coefficient by tracing down the predictor line.
  - Use "Multiple R-squared" to identify the r^2 (*We will discuss "Adjusted" R-squared later!)*

Example: The following data represents the linear model created when we use the length of a mammal's sleep cycle (predictor) to estimate the total sleep that a mammal might get on average (response)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.528      1.028  13.154 5.44e-14 ***
sleep_cycle   -5.374      1.824  -2.946  0.00617 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.643 on 30 degrees of freedom
Multiple R-squared:  0.2244,     Adjusted R-squared:  0.1986
F-statistic:  8.68 on 1 and 30 DF,  p-value: 0.006169
```



Based on this output, we can identify the model equation as:

$\hat{y}$ (estimated sleep total) = 13.528 – 5.374(sleep cycle)

For every one hour increase in sleep cycle, we expect sleep total to be 5.374 units higher / lower *on average.*

The expected error in our slope coefficient is 1.824, but we're still very confident there is a non-zero slope given that the p-value for the t-test is 0.00617.

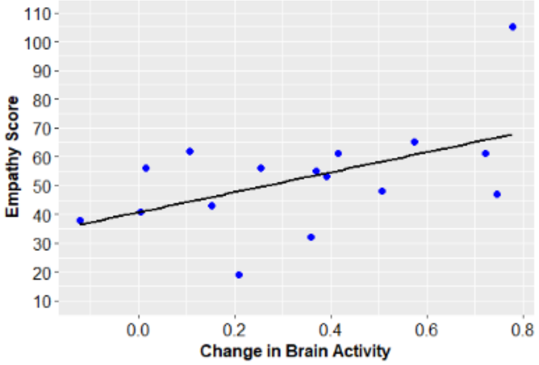
We estimate that we can reduce the variance in our prediction of a mammal's sleep total by approximately 22.44% when using their sleep cycle length as a predictor.

- **Watching for Influential Points**
  - "Outliers" are data points far removed from the consensus data.
  - In regression, we should be cautious of a special type of outlier: an "influential point."
  - An **influential point** is an outlier that can have a VERY strong effect on the best fit line, often making an otherwise "insignificant" relationship look "significant."
  - *In general, influential points will be outliers that exist near the **corners** of the graph.*
  - **What should we do with influential points?**
    - Assuming the data point was recorded correctly, *consider* running an analysis with and without that point.
    - Differentiate claims about the consensus data (general trends) from claims about all data (how variable that trend is).
    - Consider examining that special case in more detail. Why does it stand out from the rest?
  - Testing for influential points
    - At a more advanced level, data analysts might calculate the "Cook's Distance" to determine how large it's influence is on the model.

Points that differ from their peers
are often the most interesting.

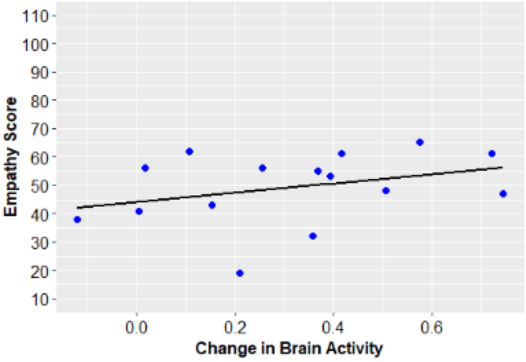
Practice: Empathy means being able to understand what others feel, but does increased brain activity signal increased empathy? 16 women watched their partner get shocked in a controlled environment, and their brain activity was measured. They also completed an empathy test. The results are shown below. Is there evidence to suggest that there is a linear relationship between brain activity and empathy score for female partners?

```
Coefficients:
              Estimate Std. Error t value P-value
(Intercept)   40.674   6.731      6.042   3.03e-05 ***
Brain (slope) 34.856   15.500     2.249   0.0412 *
---

Residual standard deviation: 16.52 on 14 df
R-squared: 0.2654, Adjusted R-squared:  0.2129
```



P-value approximately 4%, suggesting strong evidence of at least some linear relationship. R squared is 26%

What happens if we remove that one point in the top right corner?



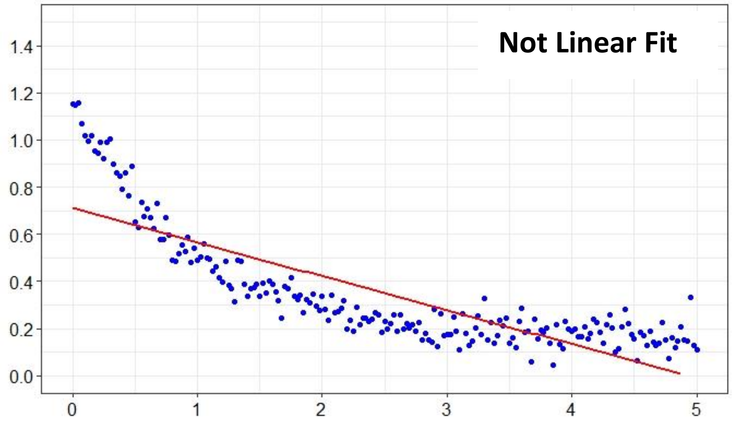
```
Coefficients:
                Estimate Std. Error t value P-value
(Intercept)     44.008   5.183      8.491   1.16e-06 ***
Brain (slope)  16.334   12.928     1.263   0.229
---

Residual standard deviation: 12.49 on 13 df
R-squared:  0.1094,    Adjusted R-squared:  0.04085
```

How might this change our conclusion about whether Brain Activity is a linear predictor of Empathy?

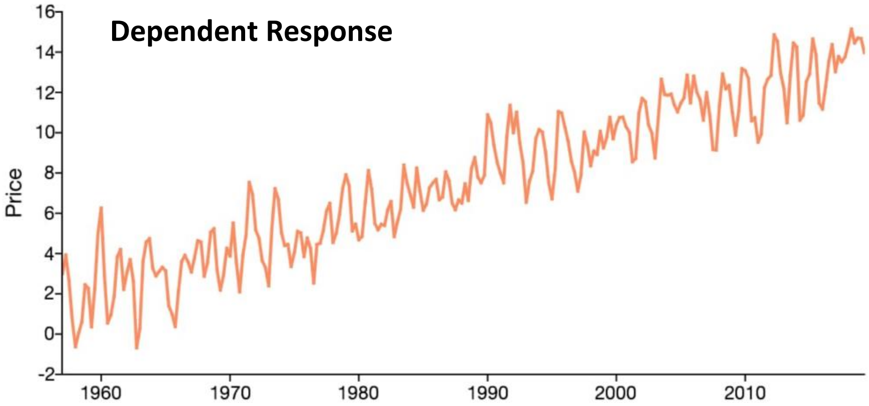
P-value approximately 23%, suggesting little evidence of a linear relationship. R squared is 11%. This one data point can change the story quite a bit! Worth exploring what is different about this one case.

Assumptions for Linear Regression Inference

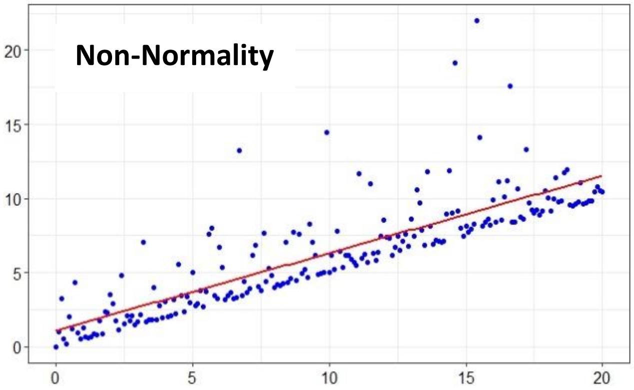
- Before doing inference for a linear relationship, there are 4 assumptions we need to check. We can remember them with the acronym **LINE**: **L**inearity, **I**ndependence of response, **N**ormality of residuals, and **E**qual variance.
  - **Linearity**
    - **Why is this important?** If the relationship is better fit by something non-linear, then doing a test on a linear term and reporting that analysis might be misleading.
    - **Never** run a regression on two variables **without looking** at the data first.
    - The picture below is an example of data that may be better fit with an exponential decay term, rather than simply a linear term. A linear term is working better than no model at all, but we could do better!



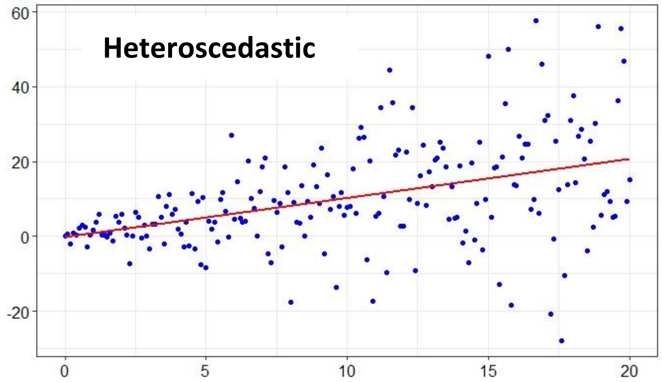
- **Independence in Response Variable Observations**
  - If the data is collected in series, where each y is dependent on the previous y, we may have a situation where Y observations are dependent on one another.
  - **Why is this important?** Linear regression is assuming our observations are independent. When the data is dependent, then we don't have a <u>random sample</u> of possible observations. This is a completely different data situation!
  - If the dependency is time-related, then there are other modeling choices like Time-Series that would fit the situation.
  - *In general, this issue is contextually recognized, rather than obvious from a graph.*



- **Normality of Residuals**
  - Ideally, we want our data points to be normally distributed about the best fit line at any cross-section (at any X value) of our plot.
  - **Why is this important?** If the data is skewed at cross-sections of X, then the distribution of possible sample slopes may not be <u>normally distributed</u>. This is an assumption we need when doing inference.
  - See picture on left: even though there is clearly some type of linear relationship, the distribution of Y at each cross-section of X is skewed.
  - Consistent with the Central Limit Theorem, this issue is minimized with larger samples.
    - ❖ Small violations should be of little concern
    - ❖ When df >100, only large violations are problematic.



- **Equal Variance (also called "Homoscedastic")**
  - Ideally, we want the variance in Y to remain fairly constant across X.
  - **Why is this important?** If the variance in Y is non-constant across values of X, then there may be more estimation error in our slope than the standard error value suggests. It can inaccurately lower the p-value for the predictor's t-test and inflate r^2.
  - If your scatterplot makes a **cone shape** (like the graph here), then your variance is **non-constant** (also called "**heteroscedastic")**.

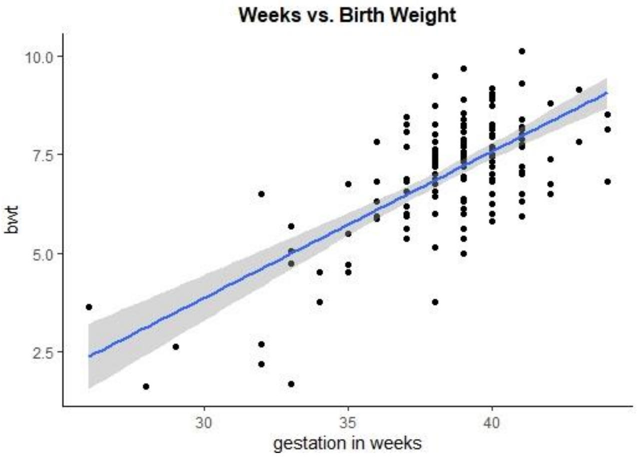


- How do statisticians deal with assumption violations?
  - **Non-linear fit?** Consider a non-linear term.
  - **Dependency?** Consider a different modeling approach that accounts for the dependency (like Time Series)
  - **Non-Normality?** Often a "Transformation" is completed on the response variable, or possibly on the predictor.
  - **Non-constant Variance?** Often a "Transformation" is completed on the response variable.

Chapter 13 Additional Practice

Investigation: Data was collected from 150 births that represent a random selection of births in one particular hospital. This dataset contains a number of variables related to the birth. Let's examine the relationship between how many weeks the mother carried the baby (weeks of gestation) and the baby's birth weight

Think through our assumptions for simple linear regression. How well is each met?



a) Is a linear fit appropriate?
   Yes.

b) Are the data points independent (no dependency in response across X)?

   These are independent observations. Example of dependency would be the same baby's weight taken every day/week.

c) Are the residuals normally distributed about the best fit line?

   Seems reasonably normal at each cross-section of X. No large violations, and sample size is quite large anyway.

d) Is the variance approximately equal across X?

   Approximately. Perhaps larger at lower gestation values, but not only a small violation

Using R, we get the following summary output from running a linear regression.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.31198    1.26305  -5.789 4.08e-08 ***
weeks        0.37248    0.03268  11.396  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.096 on 148 degrees of freedom
Multiple R-squared:  0.4674,      Adjusted R-squared:  0.4638
```

Use this information to write the equation for the line of best fit.

Predict the birth weight of a baby born at 35 weeks of gestation.

Identify r^2 and interpret this value in context (unadjusted).

Calculate a 95% confidence interval for the true slope value. Notice that the standard error value is provided in the output. *Also notice the sample size—do we need a t-interval, or is a z-interval ok?*

Are we confident that there is at least some linear relationship between gestation and birthweight? What information do we find in the output to make that determination?

Investigation revisited: The candle vendor found a sample slope of -1.183, and the SE for b_1 was calculated to be 0.2592.

Using this information, calculate a 95% confidence interval (t-interval) for β_1. Use t = 2.011

Now consider if we were testing whether or not there is a non-zero slope between price and number of sales. Based on the interval you found, would you expect the p-value from this investigation to be above or below 0.05? *Hint: What value would we use as the null hypothesized parameter?*

<u>Since the interval doesn't include 0, that suggests we'd have a < 0.05 p-value when testing 0 as the null</u>

If we had the same sample slope, but from a larger sample size, how would this most likely affect the confidence interval? *Hint: how would this affect the standard error?*

<u>Larger n would most likely reduce the standard error. This means we have more confidence in our point estimate and should have a narrower confidence interval.</u>

Chapter 13 Learning Goals

After this chapter, you should be able to…

- Differentiate when two numeric variables have a clearly linear relationship, clearly non-linear relationship, or lack of relationship
- Recognize the slope of a linear equation as the expected change in position of the response variable for a one unit increase in the predictor variable.
- Use a linear equation to identify the expected value of the response based on a predictor value.
- Acknowledge the uncertainty that exists when modeling an imperfect linear relationship
    - Uncertainty in predicting individual Y values that vary at any given X value
    - Uncertainty in selecting the coefficients of the linear equation due to a limited sample of data
- Identify a residual as the vertical distance between a response value and the model-fitted prediction for that value.
- Calculate r^2 (the coefficient of determination) based on the response variance and residual variance
- Interpret r^2 in context for a particular model
- Distinguish situations of interpolation and extrapolation, and understand why extrapolation may not lead to reliable estimates
- Conceptually make sense of a permutation test in the context of simple linear regression
- Complete a t-test/z-test for a slope under the null that the slope is 0.
    - Identify the null and alternative hypotheses
    - Identify the null model
    - Calculate the standardized position (z-score or t-score) for our sample slope
    - Generate (or simply interpret) the p-value and make an appropriate conclusion
- Complete a t-interval/z-interval for a slope (with the appropriate z-score/t-score for that particular confidence level provided)
- Identify coefficients, slope inference features, and r^2 from an R regression summary
- Recognize influential points and understand how their removal may change inferential results
- Distinguish the LINE assumptions and recognize obvious cases where these assumptions may not hold
    - Identify if there is an obviously non-linear relationship that should be explored instead
    - Identify if these are dependent observations based on the data context
    - Identify if the residuals appear non-normally distributed
    - Identify if the residuals appear to have a clearly non-constant variance across the regression.