Chapter 13: Simple Linear Modeling

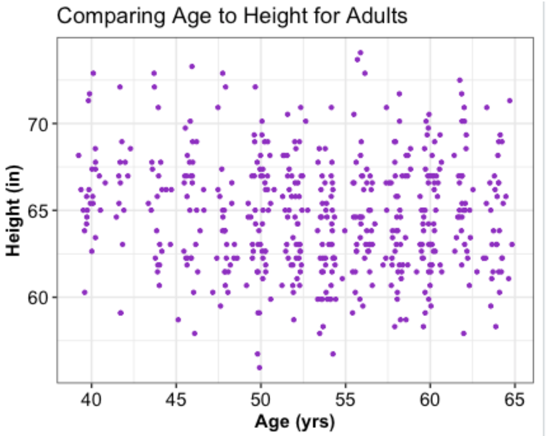
Building a Statistical Model

- What is statistical modeling?
  - In Chapter 9, we explored modeling from a more *scientific* point of view.
  - From this point of view, modeling considers how variables relate in some kind of system or process, oftentimes to consider causal and interactive mechanisms.
  - Statistical modeling allows us to examine variable using actual ________. We typically represent statistical models through the form of an _____________.

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

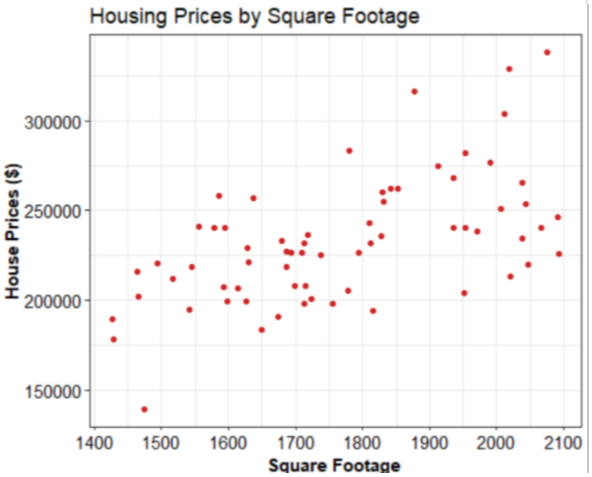
Linear Relationships

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



Non-Linear relationships

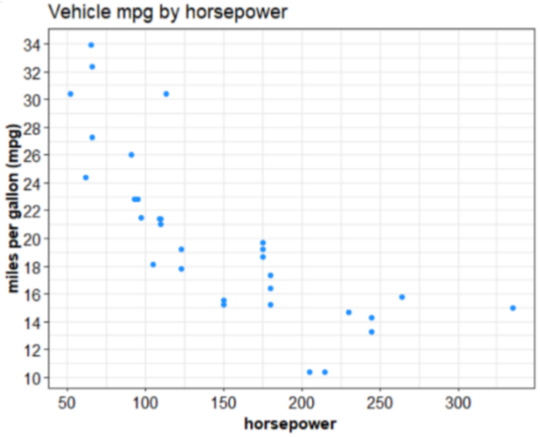
As the predictor value increases, the response value tends to change at a __________________ rate.



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 weekend, she changes the price to a different value and records the number of candles she sells. The data is presented below.

Unit of observation:

Response variable:

Predictor variable:

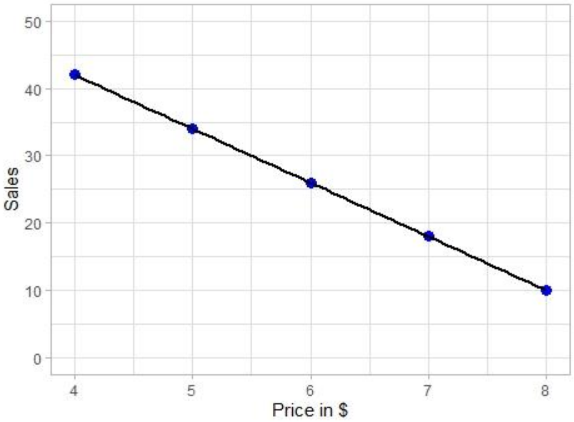


Table 1. Price and Sales Data

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. We might generally interpret slope like this:

For every one unit increase in (____________), we expect _____________) to be (__________) 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 ________, we expect _________ to be __ 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.

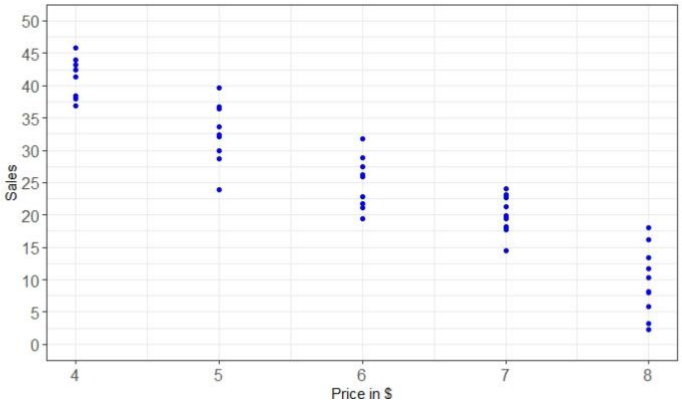
Equation of a line: $y = b_0 + b_1x$ which in this example will be sales = 74 - __*________

$\boxed{\text{Intercept}}$ $\boxed{\text{Slope}}$

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 weekends where the price was the same?



- **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 ________________ 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 _____________ to find this equation and estimate our uncertainty!
  - One common strategy in this case is to model the _______ of Y at each value of X. This method is known as linear ______________ since we are regressing the relationship toward the mean!

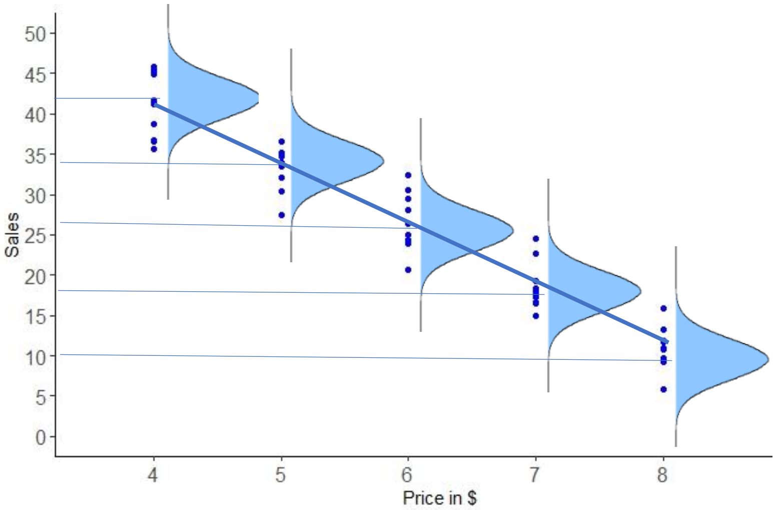
There are two sources of uncertainty when we do this!

1) Since these two variables are not in a perfect relationship, data points will ____________________________ 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 _________ of data.

Since we now have uncertainty in our prediction, we will now represent our prediction as $\hat{y}$

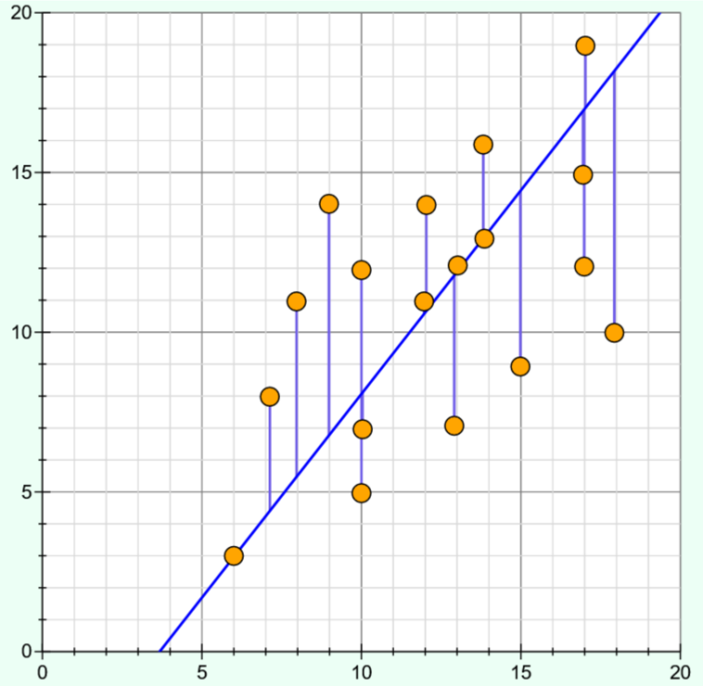
The Equation for the Best Fit Line: $\hat{y} = b_0 + b_1x$



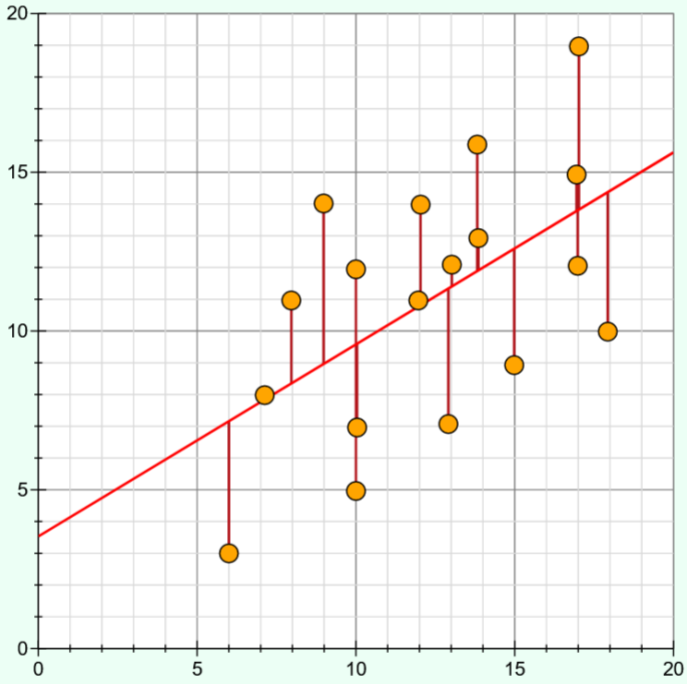
Modeling a Linear Relationship *with Uncertainty*

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

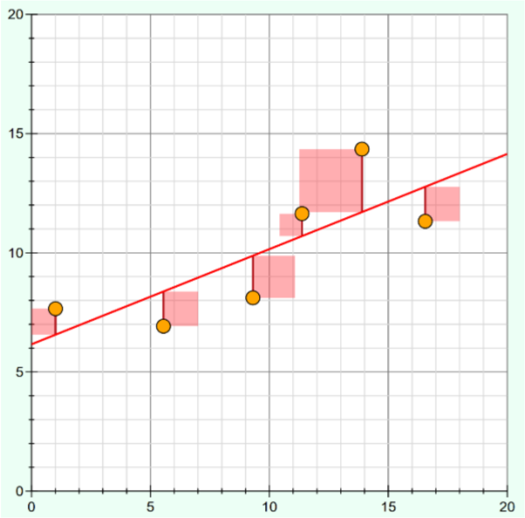
Which line do *you* think best represents the relationship? And why?

Residuals and the Least Squares Criterion

- A **Residual** represent the distance between an actual observation of the response and a prediction of the response based on a model equation.
  - $\mathbf{y}_i$ is an **actual** response value paired up with $\mathbf{x}_i$
    - Perhaps one student scored a 10 on reading comprehension (x) and 12 on speaking/listening (y), making their data point (___, ___)
  - $\hat{\mathbf{y}}_\mathbf{i}$ represents the **model predicted** response value given that $\mathbf{x}_i$ is the observed predictor.
    - For a reading comprehension score of 10, our model predicts a speaking/listening score of 3.53 + 0.60(10) = 9.53.
  - We calculate the **residual** for observation i as $\mathbf{y}_\mathrm{i} - \hat{\mathbf{y}}_\mathbf{i}$

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?
  - 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 _____________ residuals *(this is mathematically advantageous in comparison to minimizing the sum of absolute value deviations).*



PhET Least Squares Regression

Reflection Questions

13.1. Without drawing a picture, how would you describe what it means for two variables to have a linear relationship vs. a non-linear relationship? Can you think of a pair of variables to represent each case?

13.2. When building a model from a sample of data to make a prediction for some response variable, why might we have uncertainty in our prediction?

13.3. In a statistical model, what is a residual? Do we ideally want residuals to be big or small?

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. We believe there might be an association between weight levels and LDL 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?

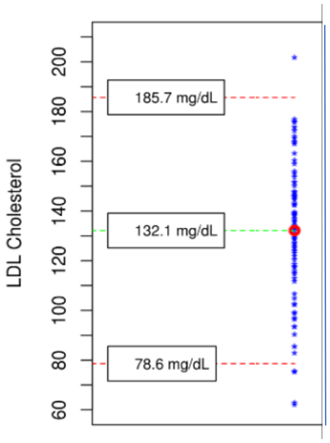
Variance ***before*** building the model

- Without a predictor, our best guess for someone's LDL level would just be the ________ 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 ________________ using our sample data will be $\mathbf{s_y}$ **= 27.1** (the subscript y identifying this is the response variable).
- This represents the expected deviation of a randomly chosen individual's LDL from the _____________.

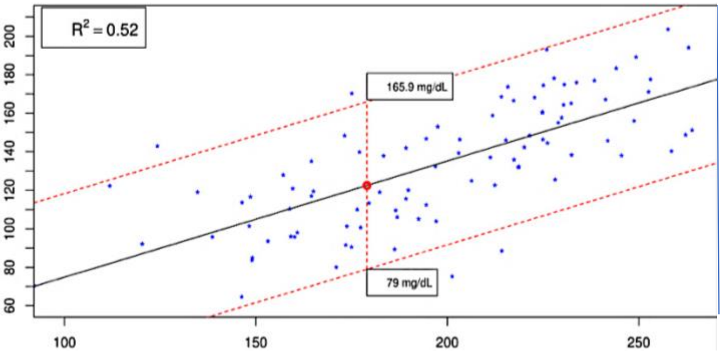


Variance remaining ***after*** building the model

- Now, we can use this model to make a more targeted prediction of one's LDL cholesterol level based on their weight.

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

- The **standard deviation** of our **model's** ________________ can be used to see how much uncertainty we have in our predictions.
- After fitting the model, we find that $\mathbf{s_e}$ = **18.78**
- This represents the expected deviation of a randomly chosen individual's LDL from our __________ ____________ for their LDL.



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?

- **Coefficient of Determination: r^2**
  - The coefficient of determination (**r^2**) is the _______________ of total variability in the ___________ variable that is "explained" by this predictor *(or by this model).*
  - While standard deviation is a more intuitive measure of uncertainty, ***variance*** is the mathematically simpler measure. So **r^2** will be based on the variances.
    - s_y^2 measures the _________ variance in the response variable. s_y^2 = _______
    - s_e^2 specifically measures the **residual** variance (the ___________ **variance**) after applying our model. s_e^2 = _______
  - We could calculate s_y^2 - s_e^2 measures **the variance that <u>is</u> explained by our model**. But it would be more helpful to standardize this measure to a common scale across contexts.
  - The ***proportion*** of the total variance we have explained by this model will be the ________ of this difference from 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!*

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

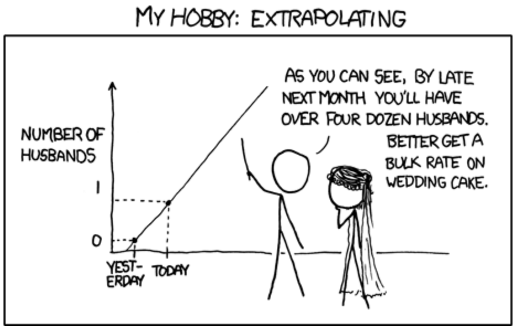
Let's calculate the r^2 for our LDL cholesterol model.

Which statement is correctly interpreting what we found?

1. Approximately 52% of these men have LDL cholesterol levels above the mean
2. Men's 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 %

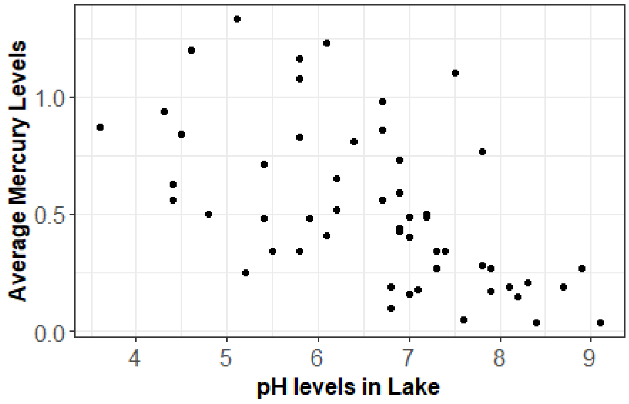
Interpolation and Extrapolation

- **Interpolation:** Predicting Y based on an X value _________ the range of X values observed
- **Extrapolation**: Predicting Y based on an X value ___________ 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?

Linear Regression Inference – Estimating and Testing β_1



Investigation: An ecologist is testing a theory that whether mercury levels might be higher in more acidic lakes. She plots the data and finds a sample slope of $b_1 = -0.152$.

Unit of observation:

Response variable:

Predictor variable:

But does that suggest pH levels and mercury levels are correlated in this way? Is this evidence that the true slope, $\boldsymbol{\beta_1}$, follows the direction we theorized?

- **Permutation Test** for Linear Regression
  - We could ___________________ the values and pair them up randomly to see what sample slope might occur due to random chance
  - This will help us determine how often we'd observe a sample slope at least as large as ours by random chance.

Let's start by writing our null and alternative hypotheses for testing the true slope.

To explore whether the slope observed in our sample could be a result of random chance, we'll use the **Lock5 StatKey** site linked here

- From the drop-down data menu, we can choose the "Florida Lakes" dataset
- Change the randomization dotplot from "Correlation" to "Slope"

When we complete permutations and plot the resulting sample slopes, what value do they center around and why?

Which "tail" checkbox should we check to get an appropriate p-value estimate? What do we find, and how does this help us answer our question?

- **The Standard Error for b_1**
  - The Standard Error for b_1 (SE_{b1}) is the expected deviation of b_1 from β_1. *We won't bother with calculating this by hand—we'll just use this result in our inferential calculations!*

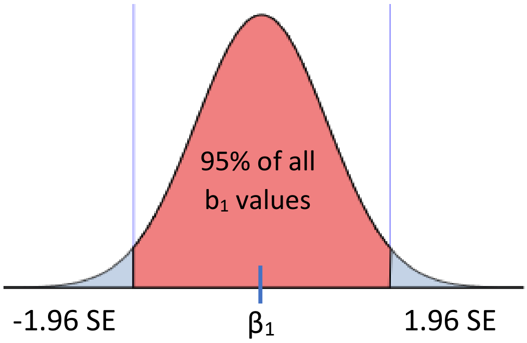
$$SE_{b1} = \frac{\sigma_e}{\sigma_x\sqrt{n}} \approx \frac{S_e}{S_x\sqrt{n-2}}$$

    - σ_e represents the true standard deviation in the ______________.
    - 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.

  - 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…__________

- **T-test and t-intervals for Slope**
  - When the 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!
  - Likewise, if we simply wish to estimate a range of plausible values for β_1, we could complete a t-interval using ___ as our ______________________.



Investigation revisited: Let's now compare our simulated p-value to what we might get using a t-test. Our null hypothesized slope is ___ and the estimated standard error for that slope is 0.037.

What is the test statistic (t-score) of our sample slope in this null model? What does it tell us?

Using the t distribution applet we've used before, we would expect to find a p-value *well* below 0.001.

Finally, let's report a 95% confidence interval for β_1 as well using a t-interval method. With 51 degrees of freedom, the t-score we'll need to complete our interval would be **2.008**.

Point Estimate:

Margin of error:

Interval Bounds:

Reflection Questions

13.4. What does the residual standard deviation measure in a model? How would you describe what the calculation $\mathbf{s_y}$ - $\mathbf{s_e}$ communicates to us about a model's performance in improving predictions?

13.5. What does r^2 communicate about a model's accuracy? What role does the denominator play in this calculation?

13.6. In your own words, can you explain how we could use a permutation test to decide whether the slope from our sample data could reasonably be a result of random chance or not?

13.7. In a **t-test** for slope, what does the t-score test statistic represent? If the t-score were around, say -1 or +1, what would that tell us about our confidence in concluding a relationship between the two variables?

13.8. In the pH and Mercury data question, our 95% confidence interval for slope was somewhere around (-0.223, -0.079). What does that communicate to us about the true slope relating lake pH levels to lake mercury levels? Without a p-value, how might we use this interval to infer a possible relationship between pH levels and mercury levels?

Read/Try on your own

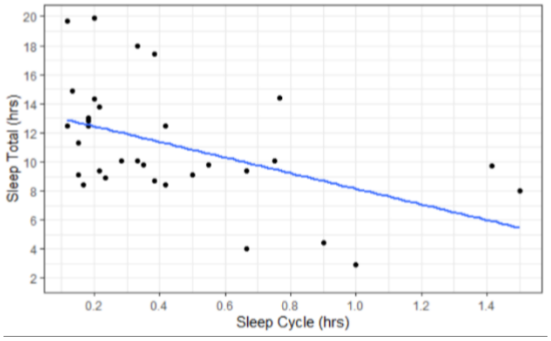
Reading R Output

- 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. This reports everything we might use to complete a **t-test** for the **slope!**
  - 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) = ________ – 5.374(sleep cycle)

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

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

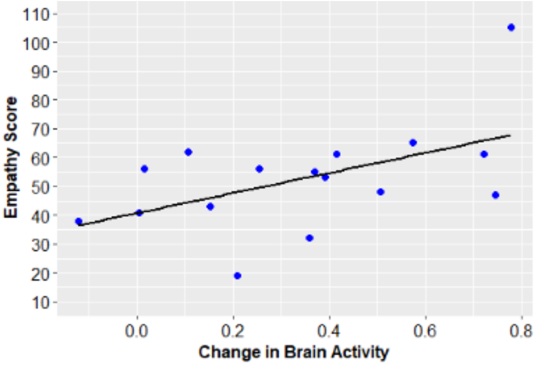
We estimate that we can reduce the variance in our prediction of a mammal's sleep total by approximately _________% when using their sleep cycle length as a predictor.

Watching for Influential Points

Practice: Does increased brain activity signal increased feelings of empathy? 16 women watched their partner get shocked in a controlled environment, and their brain activity was measured. They also completed an empathy test scored from 0 to 120.

```
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
```

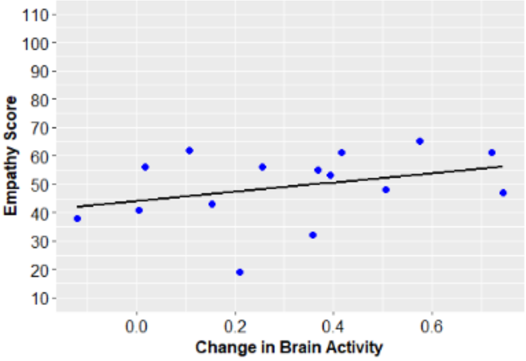


Is there evidence to suggest that there is a linear relationship between brain activity and empathy score for female partners?

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
```



Does the relationship between change in brain activity and empathy score appear to be stronger or weaker? Does our story change?

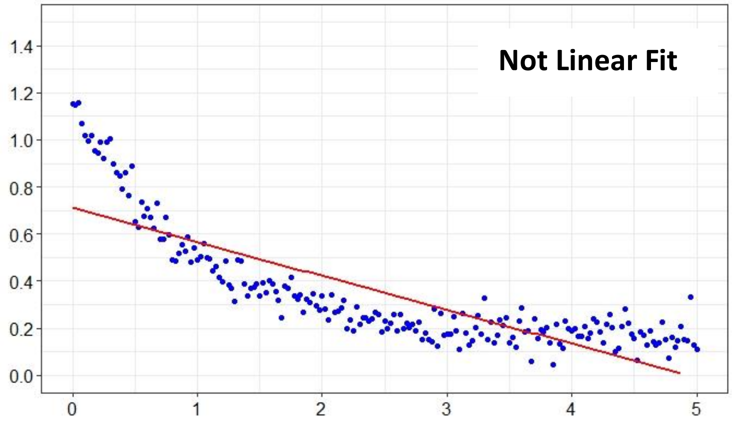
- In regression, we should be cautious of a special type of outlier: an "influential point."
    - "Outliers" are data points far removed from the consensus data.
    - An **influential point** is an outlier that can have a ________ __________ effect on the best fit line, often making an otherwise "insignificant" relationship look "significant" and vice-versa.
    - In general, influential points will be outliers that exist near the ___________ of the graph.
    - **What should we do with influential points?**
        - A Consider examining that special case in more detail. Why does it stand out from the rest?
        - If the data point represents a valid observation that was recorded correctly, it should *probably* still be included in analysis. If excluding from model creation, be transparent about why it was excluded beyond "because it was an outlier."
        - It may be valid to exclude if differentiating claims about the consensus data (general trends) from claims about all data.

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

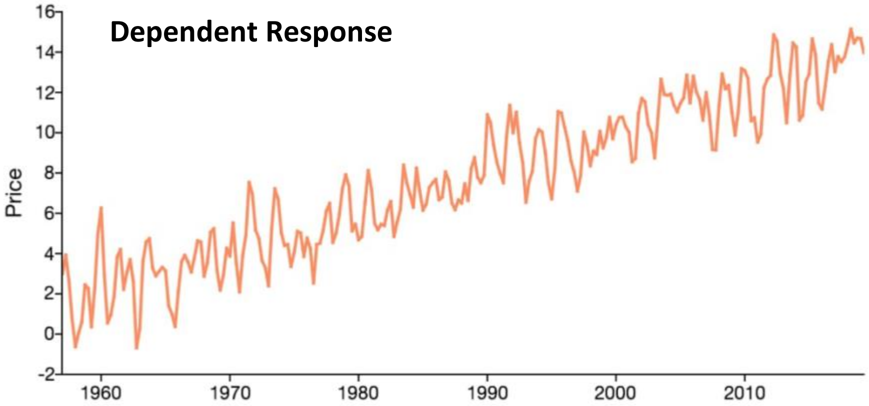
Conditions for Linear Regression Inference

Before doing inference for a linear relationship, there are 4 conditions 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.

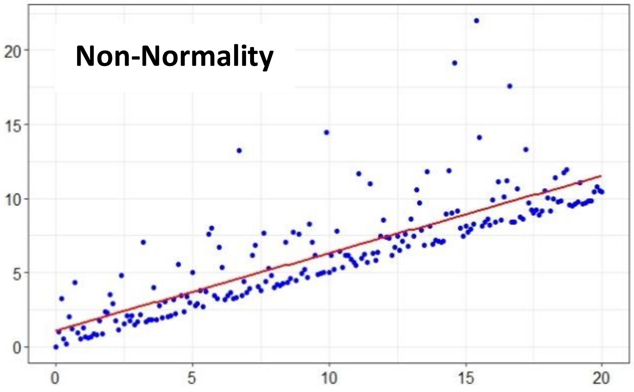
- **Linearity**
    - **Why is this important?** If the relationship is better fit by something non-linear, then doing a test on a linear term and reporting that analysis might be ______________.
    - **Never** run a regression on two variables **without looking** at the data first.
    - The picture 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!



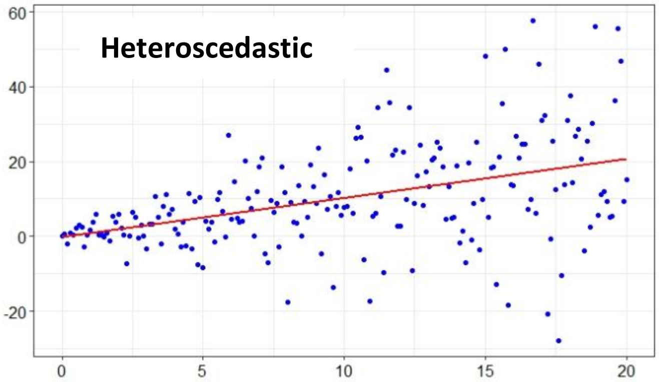
- **Independence in Response Variable Observations**
  - If the data is collected in series, where each y is dependent on the previous y, we may have a situation where Y observations are dependent on one another.
  - **Why is this important?** Linear regression is assuming our observations are independent. When the data is dependent, then we don't have a ________________ sample 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 highly skewed at cross-sections of X, then the distribution of possible sample slopes may not be ___________________. This is a condition we need when doing inference on the slope.
  - 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, this is unlikely to be an issue.



- **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 _________ the p-value for the predictor's t-test and inflate r^2.
  - If your scatterplot makes a _______ **shape** (like the graph here), then your variance is **non-constant** (also called "**heteroscedastic")**.



- What do statisticians do when conditions aren't met for linear regression inference?
  - **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.

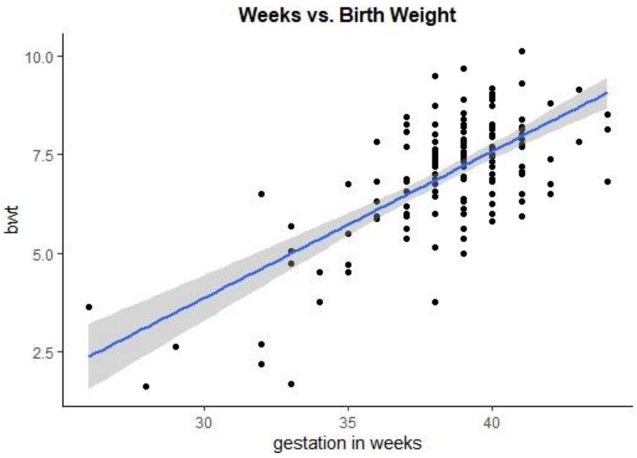
Reflection Questions

13.9. Look at an R model summary output from one of the previous examples and find the slope value. What do each of the 3 values to the right of the slope value represent? What do they communicate to us? *Hint: Walk through how we do a t-test from earlier in the notes!*

13.10. When doing regression, what makes a point an influential point? Can a data point be an outlier, but not necessarily an influential point?

13.11. What term do we use to describe when the relationship between a predictor and a response variable has non-constant variance? Why might a model fit on this type of data not be especially reliable?

Chapter 13 Additional Practice (Videos available in the Ch 13 module on Canvas!)



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 conditions for linear regression inference. How well is each met?

a) Is a linear fit appropriate?

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

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

d) Is the variance approximately equal across X?

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?*

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?*