## Chapter 13: Inference for Linear Relationships

### Simple Linear Regression Review

- **Reading R Output**
  - When using R to run a linear model, you can find several important values in the model summary. In particular, we can identify the intercept, the slope, and r squared.

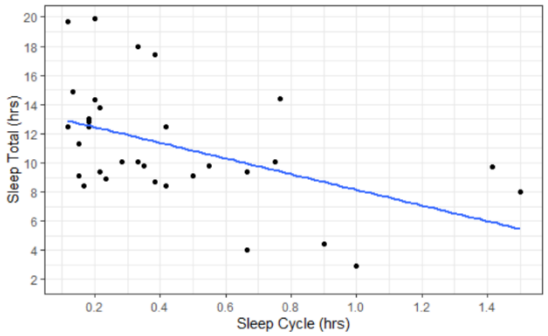
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7347     8.7594   0.769    0.445
predictor     3.0695     0.3948   7.775 8.82e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.139 on 63 degrees of freedom
Multiple R-squared:  0.4897,     Adjusted R-squared:  0.4816
F-statistic: 60.45 on 1 and 63 DF,  p-value: 8.825e-11
```

**Example:** The following data shows the relationship between how long an animal's sleep cycle is and how many hours of sleep that animal gets on average for about 32 animals.

```
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
```




1. Based on the plot, does there seem to be a negative association, a positive association, or no association between sleep cycle length and sleep total?

2. Based on this output, what would be the equation for the best fit line for this relationship? Hint, we can determine it from the "estimate" column.

3. For every one hour increase in _____, we expect _____ to increase/decrease by _____ hours on average.

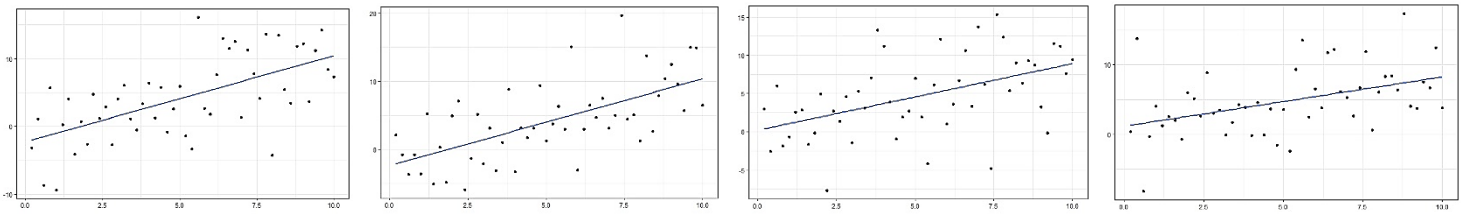
4. Approximately _____% of the variance in _____ is explained by a linear association with _____

**Inference in Simple Linear Regression**

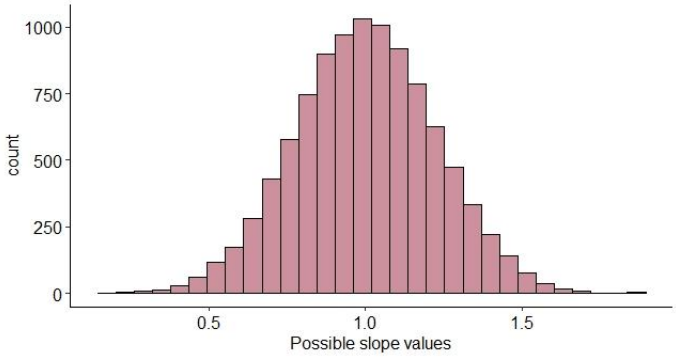
- **Linear vs. Non-linear relationships**
  - Recall that when we have two numeric variables (we can call them X and Y), they might have some type of relationship.
  - In this chapter, we will examine potential linear relationships more closely to determine if we are confident there is at least some linear relationship.
- **Sampling Distributions for the Slope**
  - Conducting inference on the slope is much like inference for a mean.
  - The true slope value for a linear relationship is symbolized ___ ("beta 1")
  - The statistic representing our slope from a sample of data is symbolized $b_1$
  - There exists a distribution of possible sample slopes ($b_1$'s) we could see when taking samples from X and Y.

**Example.** Consider two variables that have a positive linear relationship with $\beta_1 = 1$. Let's simulate some sample slopes we could see if we take 50 random observations from this population. *Note that the population correlation coefficient was set to R = 0.5, and the sample r values will also vary with each sample!*

$b_1 = 1.269$ $b_1 = 1.274$ $b_1 = 0.873$ $b_1 = 0.709$



As expected, these sample slopes vary from sample to sample (sometimes above 1, sometimes below 1). So what happens if we generate 10,000 samples of 50 to see what sample slopes we see?




- As we see here, the distribution of possible slope values is _____ and centered around the true slope value $\beta_1 = 1$.
  - *Note that the sampling distribution for the sample slope won't <u>always</u> be normally distributed—we'll talk about that more later!*

- **The Standard Error for $b_1$**
  - The Standard Error for $b_1$ ($SE_{b1}$) is the expected deviation of $b_1$ from $\beta_1$.
  - This measure is complex to derivate, but it can be shortened to the following formula:

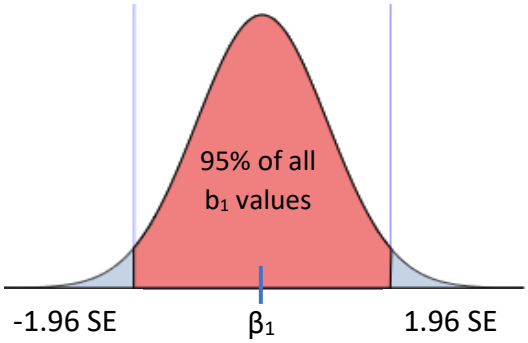
$$\mathbf{SE}_{b1} = \frac{\sigma_e}{\sigma_x\sqrt{n}}$$

    - $\sigma_e$ represents the true standard deviation in the _____
    - $\sigma_x$ represents the true standard deviation in the _____ (X) variable
    - n represents the sample size
  - When estimating this value using only *sample* data, we would estimate the standard error as:

$$\mathbf{SE}_{b1} \approx \frac{S_e}{S_x\sqrt{n-2}}$$

*We lose 2 degrees of freedom rather than just 1. We use our sample data to estimate both the slope value and intercept. For example: we need at least 3 data points before we can estimate the strength of the relationship between X and Y.*

- **Confidence Interval for $\beta_1$**
  - When distribution of possible $b_1$'s is normally distributed about $\beta_1$, we can create confidence intervals following a similar procedure as with a mean.
    - If we are estimating $\mathbf{SE}_{b1}$ with a not large sample (df ≤120), we should do a t-interval
    - When df > 120, a z-interval should be reliable.
  - Our point estimate for $\beta_1$ will be _____

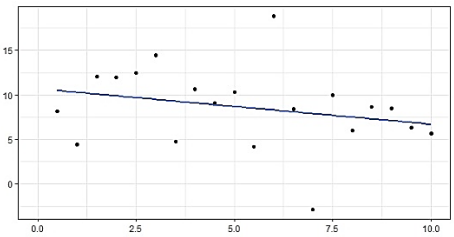



**Practice:** Using the candle data (n = 50) that we had previously reported, we found a sample slope ($b_1$) of -1.183, and the SE for $b_1$ was calculated to be 0.2592. Using this information, calculate a 95% confidence interval (t-interval) for $\beta_1$.

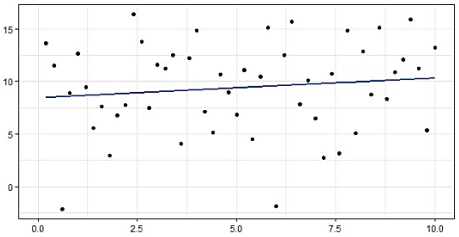
If we had the same sample slope, but from a larger sample size, how would this most likely affect the confidence interval? *Hint: how would this affect the standard error?*

- **Hypothesis testing for $\beta_1$**
  - Since sample slopes vary, it's possible that our non-zero slope could just be sampling variability.
    - This is especially a concern with smaller samples, as we could get some considerable "chance correlation" when n is small.
  - Consider three pictures below, representing samples of 20, 50, and 200. In all three cases, these **samples** were generated from an X and Y with **no relationship**.

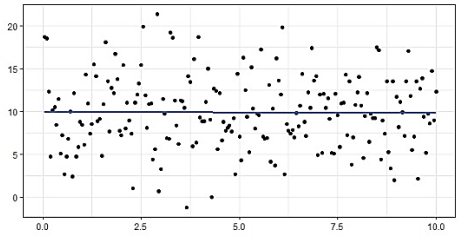
n = 20
$b_1$ = -0.398
r = -0.260



n = 50
$b_1$ = 0.189
r = 0.126



n = 200
$b_1$ = -0.015
r = -0.011



  - **Hypotheses**
    - For this class, we're specifically interested in testing to determine whether there is a linear relationship or not.
    - In this case, we would set our null hypothesis to be _____
    - The Alternative in a non-directional question would be _____
  - As with means, we need to conduct a t-test if estimating $\sigma_e$ with $s_e$, but larger samples (i.e., df > 120) could reasonably be tested with a z-test.

**Practice:** Consider the previous question where we had a sample slope of -1.183 and SE of 0.2592 from a sample of 50. Complete a test to determine how confident we are that the true slope is not equal to 0.

What would be the null and alternative hypotheses in symbols?

Calculate the test statistic.

Use a t-table to estimate the p-value. What should we conclude about the hypotheses?

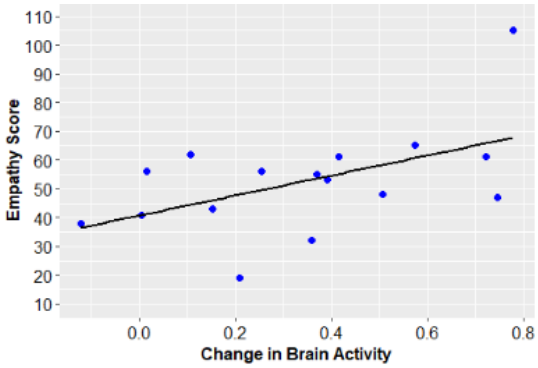
- **Watching for Influential Points**
  - "Outliers" are data points far removed from the consensus data.
  - In regression, we should be cautious of a special type of outlier: an "influential point."
  - An **influential point** is an outlier that can have a _____ effect on the best fit line, often 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?


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

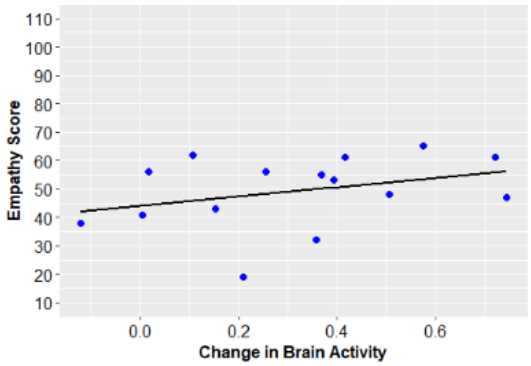
**Practice:** Empathy means being able to understand what others feel, but does increased brain activity signal increased empathy? 16 women watched their partner get shocked in a controlled environment, and their brain activity was measured. They also completed an empathy test. The results are shown 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
```



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



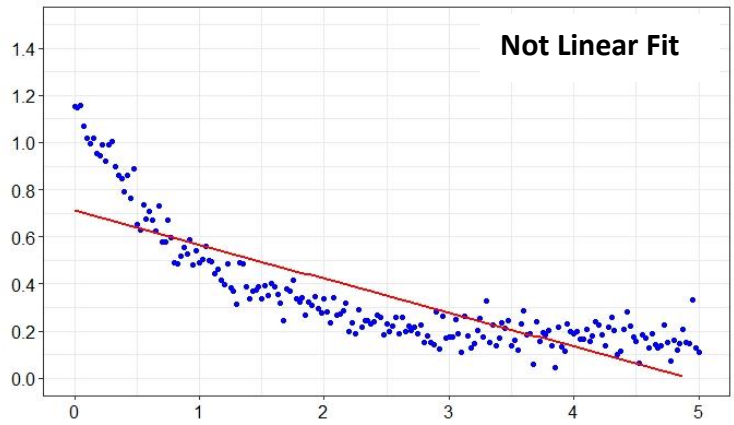

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    44.008    5.183      8.491   1.16e-06 ***
Brain (slope) 16.334    12.928     1.263    0.229
---

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

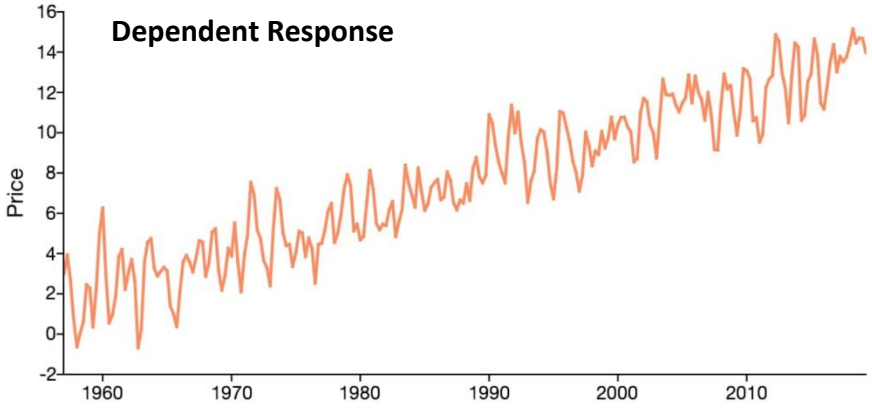
How might this change our conclusion about whether Brain Activity is a linear predictor of Empathy?

**Assumptions for Linear Regression Inference**

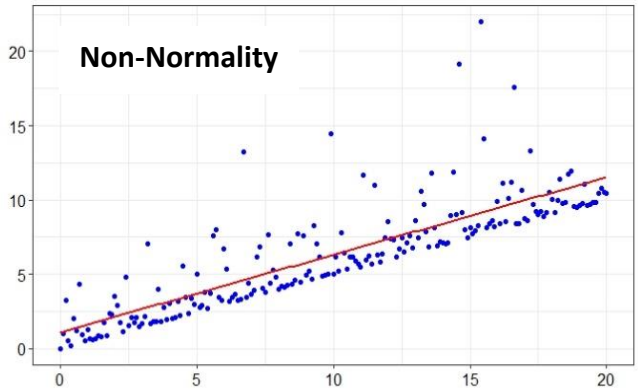
- Before doing inference for a linear relationship, there are 4 assumptions we need to check. We can remember them with the acronym **LINE**: **L**inearity, **I**ndependence of response, **N**ormality of residuals, and **E**qual variance.
    - **Linearity**
        - **Why is this important?** If the relationship is better fit by something non-linear, then doing a test on a linear term and reporting that analysis might be _____.
        - **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!



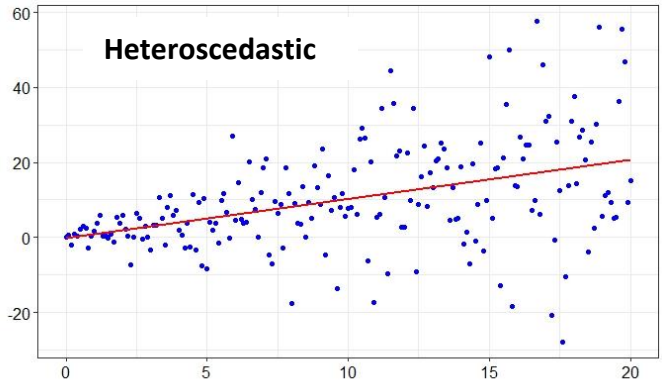
- **Independence in Response Variable Observations**
  - If the data is collected in series, where each y is dependent on the previous y, we may have a situation where Y observations are dependent on one another.
  - **Why is this important?** Linear regression is assuming our observations are independent. When the data is dependent, then we don't have a _____ 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 _____. 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 >120, only large violations are problematic.

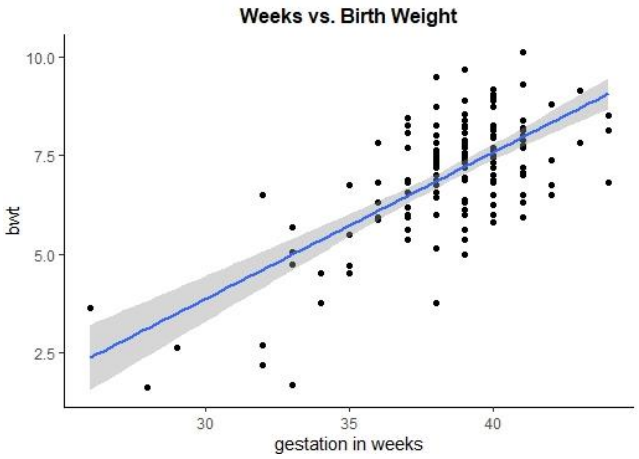


- **Equal Variance (also called "Homoscedastic")**
  - Ideally, we want the variance in Y to remain fairly constant across X.
  - **Why is this important?** If the variance in Y is non-constant across values of X, then there may be more estimation error in our slope than the standard error value suggests.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

**Practice:** 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?

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?