## Chapter 3: Modeling Relationships

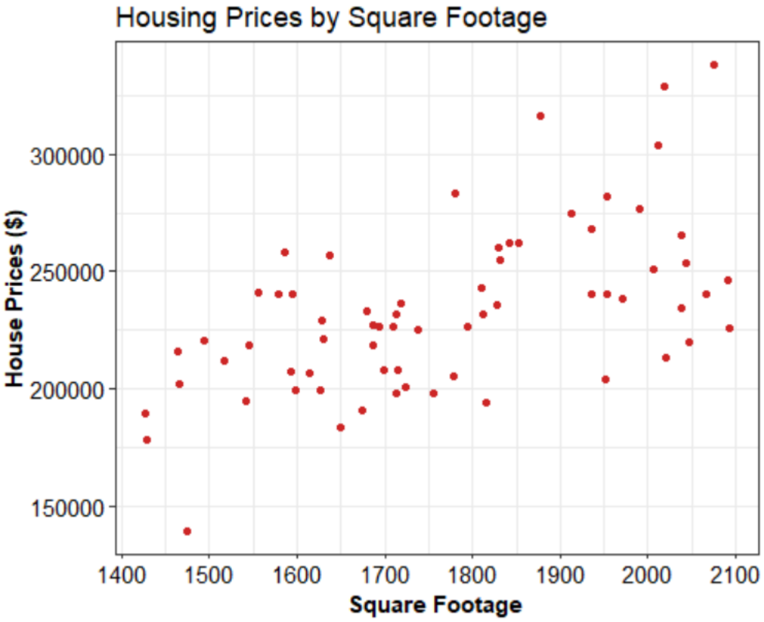
### Visualizing Relationships with Numeric Variables

- A **scatterplot** is a visualization used to represent the relationship of two numeric variables
- Each point in the plot represents one unit of observation, and its placement in the coordinate plane identifies that unit's recorded data for two different variables.
  - If one variable is considered the _______________ **variable** (the one we have interest in predicting and explaining variation in), that variable will go on the **y-axis**.
  - The _______________ **variable** goes on the **x-axis**.
- **Explaining Variability** in this context means that the distribution of the response variable is centered differently depending on the value of the predictor variable.

**Example.** Consider the following plot, representing 67 houses in a particular community. Does square footage explain variability in housing prices?

Unit of observation:

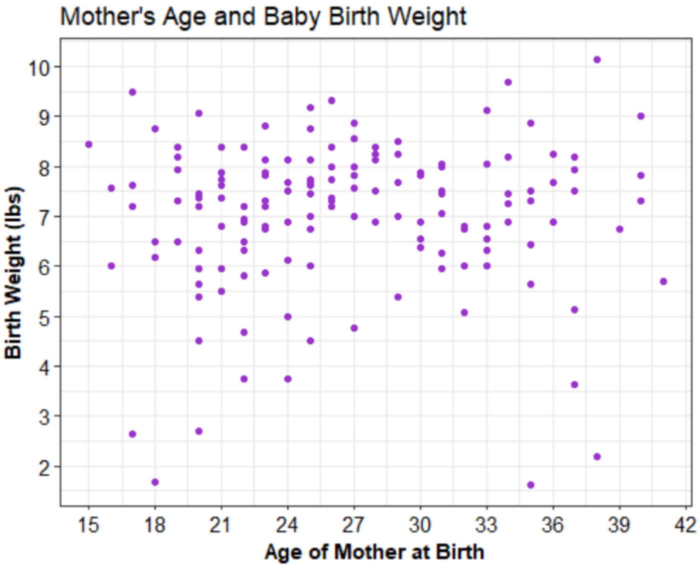
Association?



**Example.** Does knowing the age of a mother at the birth of a baby explain variability in the baby's birth weight?

Unit of observation:

Association?

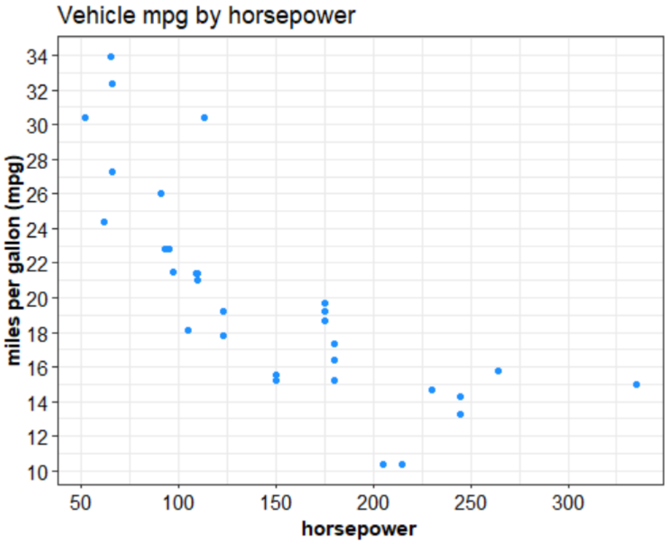


- When investigating the relationships of numeric variables, we may notice different types of patterns
    - The variables have ______ relationship *(predictor explains no variability in response)*
    - The variables have a ____________ relationship
    - The variables have a _________________ relationship

**Example.** What type of relationship does a vehicle's horsepower have with its mpg?

Unit of observation:

Association?



**Simple Linear Regression**

- In this chapter, we will focus on modeling linear relationships.

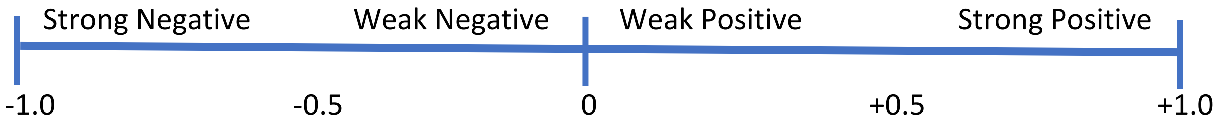
**Practice**: Explore the following simulation and use it to answer the following questions. https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

Think about two variables: "How many hours a week a college student spends outdoors" and "How many hours a week a college student spends on social media." If you collected data from these two variables with 15 college students, what do you think this data could look like? What type of relationship would they have? Plot it on the scatterplot.

Explore "r." Move your data points around and try to see what values it varies between. What do you think it represents?

Can you draw a line that best fits the data? Compare it to the "best fit line" option. How does it look different?

- **Measuring association**
  - In statistics, the **"Pearson" Correlation Coefficient** (abbreviated "**r**") is a statistic between ___ and ____ that describes the direction and strength of a **linear** association between two numeric variables.
  - Negative values imply that as one variable increases in value, the other decreases in value. *(Negative correlation).* Positive values imply that as one variable increases, the other variable increases as well *(Positive correlation).*
  - The correlation coefficient is abbreviated r (for sample statistic) or ρ (for population parameter).



**Practice:** Match the correlation coefficient with the scatterplot it represents.

r = -0.50

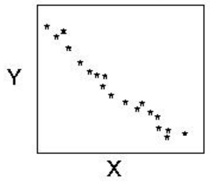
r = 0.90

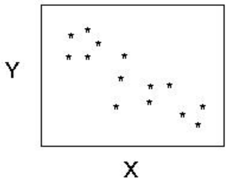
r = -0.90

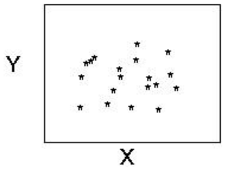
r = 0

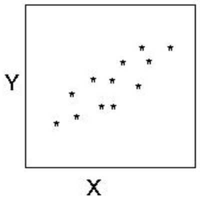
r = 1.00

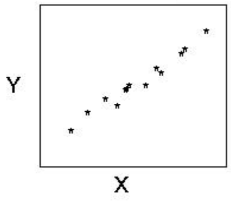
r = 0.50

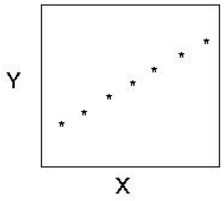








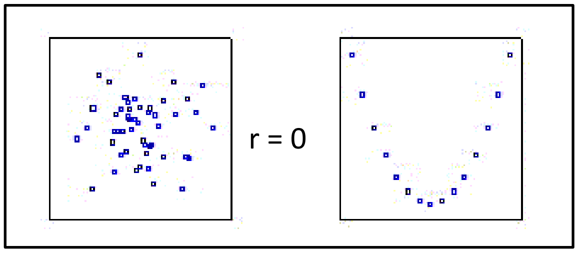




- **How would you calculate the correlation coefficient between two variables?**
  - In this class, you will **never** be asked to calculate r by hand from a set of data, but here is the formula!
    - Formula: $r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

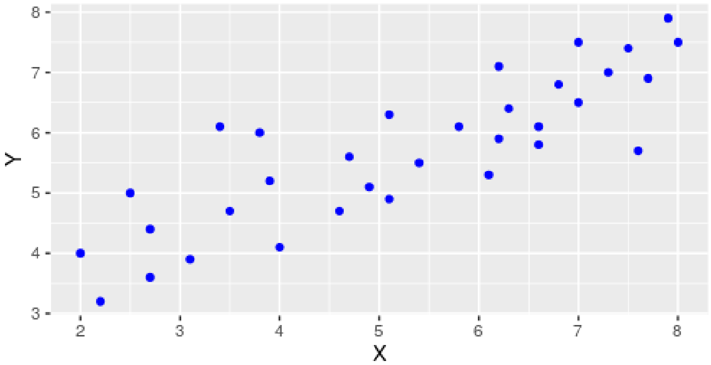
**Discussion:** Think about the numerator. What type of paired values for (x, y) would you expect to see if the relationship between the two variables was negative?

- **When is it *inappropriate* to use Pearson's correlation coefficient?**
    - As touched on earlier, not all relationships are linear in form. A low correlation coefficient signals no <u>LINEAR</u> relationship.
    - The graph on the left panel shows no association at all, but the panel on the right shows a non-linear association that would also yield a correlation coefficient of 0.



**Digging Deeper:** Let's assume two variables X and Y are positively correlated, with r = 0.5. Which situations would change r?

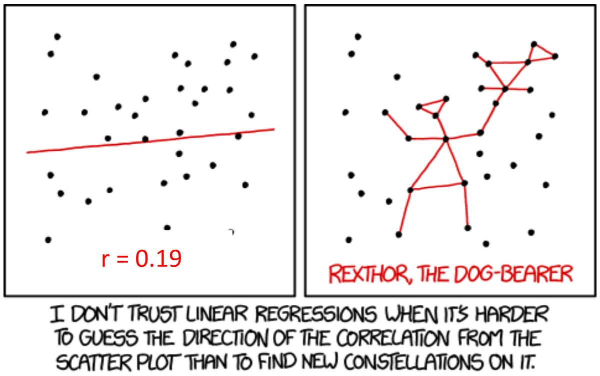
Adding 10 to every X value. Would it change r? If so, how? (if in doubt, try drawing a picture or thinking about the formula again!)



Switching the Y values linked with the 3 lowest X values with the 3 highest X values

Switching the predictor and response variable around (e.g., which is on the X axis and which is on the Y axis).

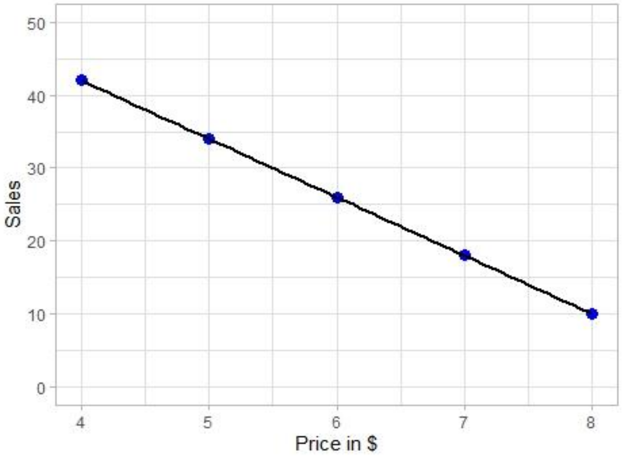
For those of you who are competitive, you can make a game of guessing correlations here: http://guessthecorrelation.com/


r = 0.19
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

- The Regression Line (**Line of Best Fit**)
  - A primary purpose of regression is to actually model the relationship with an equation. But how do we find an equation that best fits the relationship?

  - **Example:** Consider this fictional data in which a store owner is interested in seeing the relationship between the price (X) of a candle and the number of sales (Y) for a day.
  - He conducted the study over 5 days, where he randomly assigned a price ($4, $5, $6, $7, $8) to the candle each day.

| | Price (X) | Sales (Y) |
|---|---|---|
| | $4.00 | 42 |
| | $5.00 | 34 |
| | $6.00 | 26 |
| | $7.00 | 18 |
| | $8.00 | 10 |
| **r** | -1.00 | |



  - This data shows a ______________ ______________ _________________
  - As the price goes up by $1, we notice that the number of weekly sales drops by ____. This value would be the slope coefficient for the predictor variable “price.”
  - Equation of a line
    - In this situation, there is a perfect linear relationship between the two variables.
    - Equation of a line: $y = \underbrace{74}_{\boxed{\text{Intercept}}} - \underbrace{8x}_{\boxed{\text{Slope}}}$
    - **Slope** tells you the rate at which the response variable changes with respect to unit changes in the predictor.

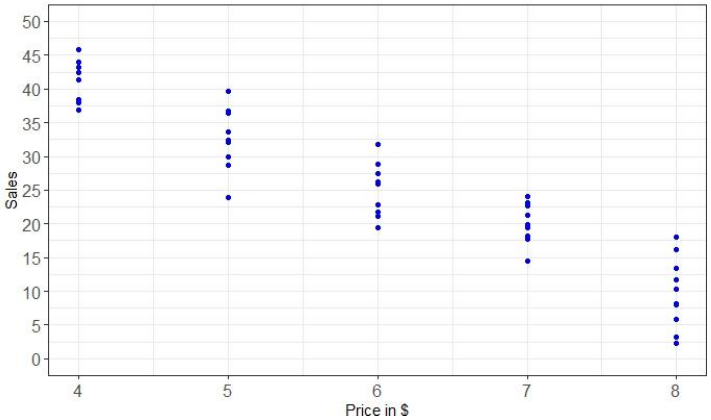
For every unit increase in _________________ we expect _________________ to increase/decrease by _____ on average.

- **Intercept** provides you a starting point/positional reference—the model’s approximation for the response value when the predictor variable is at 0 (typically not a meaningful number to interpret on its own).

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

- BUT…with real data, we likely won't find a *perfect* correlation.
  - Now consider if the store owner did his study over 50 days and collected sales data for 10 days per price point.
  - Notice that for each price point, there is a distribution of possible sales values

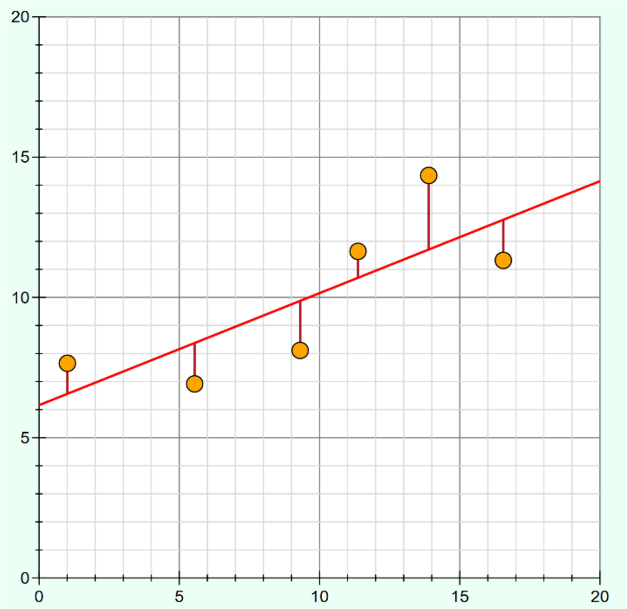
**Why?**



  - Even with variability at each price, we see a consistent, linear trend between price and sales that we can still capture with a linear equation.

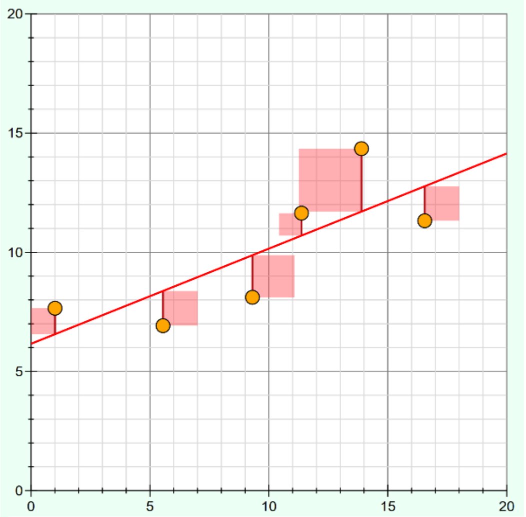
- The Equation for the Best Fit Line is $\hat{y}_i = b_0 + b_1x_i$
  - $y_i$ is the **actual** Y value paired up with $x_i$.
    - ❖ $(x_i, y_i)$ is an observed data point.
  - $\hat{y}_i$ represents the **predicted** Y value given that $x_i$ is the observed X value
    - ❖ $(x_i, \hat{y}_i)$ is a point *on the best fit line*
  - **Residual:** the vertical distance from a data point to the regression line.
  - We calculate the **residual** for observation i as $\mathbf{y_i - \hat{y}_i}$

**Practice:** The equation for the graph to the right is $\hat{y}_i = 6.16 + 0.4x$. What is the model's predicted value for Y when X is 12? What is the residual for the data point at (12,12)?



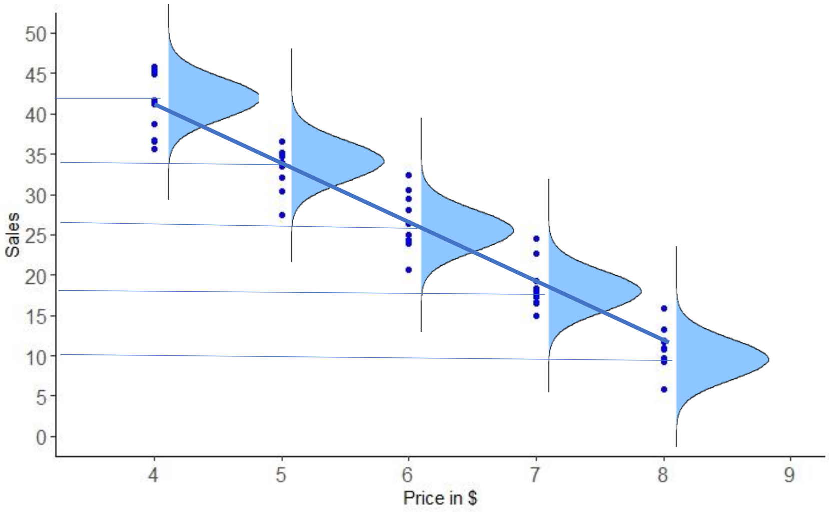
**PhET Least Squares Regression:** https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

- Constructing a Regression Line
  - One option is to find the line that minimizes the sum of absolute value residuals.
  - A mathematically cleaner option is to find the line that minimizes the ______________________ residuals!
  - This method is called the **least-squares criterion** for finding the best fit line.
  - Determining the equation of the line that accomplishes this is a bit mathematically involved (use software!), so the best fit line coefficients will always be provided in our class when working by hand.

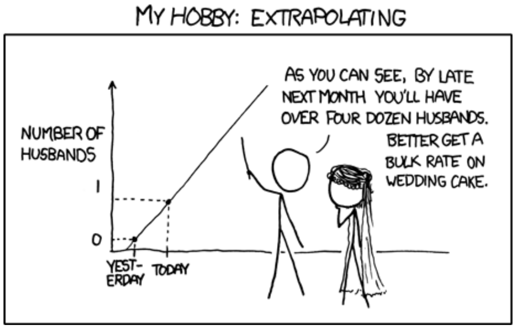


**PhET Least Squares Regression:**
https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

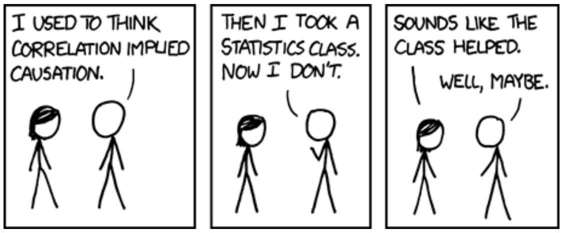
- What the Best Fit Line Represents
  - We can think of Y as having a distribution at each value of X. We call this the conditional distribution of Y given X.
  - The goal of our best fit line is to model where the mean of Y is at each X.
  - The picture below shows the true distribution of Y given X (including where the mean of Y really is in each case), along with the best fit line from the data only.



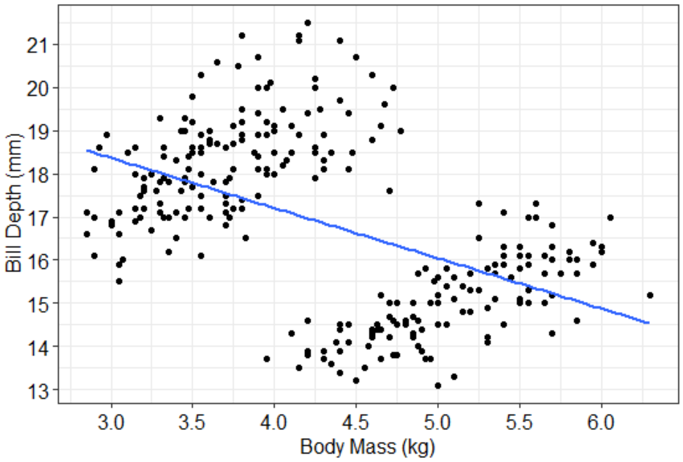
- **Cautionary Notes with Linear Regression**
  - **Interpolation and Extrapolation**
    - **Interpolation:** Predicting Y based on an X value ____________ the range of X values we have information for.
    - **Extrapolation**: Predicting Y based on an X value _____________ of the range of X values we have information for.
    - While making predictions immediately outside the range is generally safe, making predictions well out are often unreliable.
    - Consider the Store owner data; what if we wanted to predict weekly sales for a price of $10?

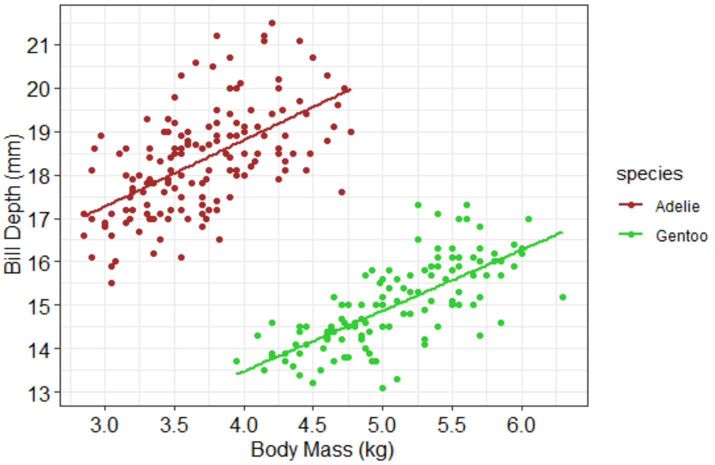

MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

  - **Association vs. Causation**
    - **Association DOES NOT imply Causation**
    - In other words, simply because we've identified a linear (or some other) relationship between variables, we don't actually know if one is the causal agent of the other.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

**Example.** Consider the first plot, showing the body mass and bill depth of penguins on Palmer Island. Penguins with a larger body mass tend to have lower bill depth on average, but something is fishy here…what happens when we reveal which of the two species of penguins there are here?





We stratified the relationship by…____________

This is an example of…______________________

## Judging the Model

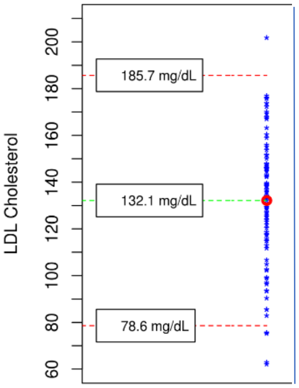
- Accounting for and explaining variation
    - The reason that we typically won't see perfect correlation between two variables is because there are often a lot of factors affecting our response variable.
    - This unexplained variation is called ___________________________. This clouds the underlying relationship between our two variables.
    - ***We are trying to model the signal (the true relationship between our two variables) amidst this noise.***

**Example:** Consider the response variable LDL Cholesterol level (the type of cholesterol that is considered more problematic). Naturally, LDL cholesterol levels vary for each person.

- The ***variance of our response variable*** is notated as $\boldsymbol{\sigma_y^2}$ (and the variance of our sample is notated as $\mathbf{s_y^2}$)

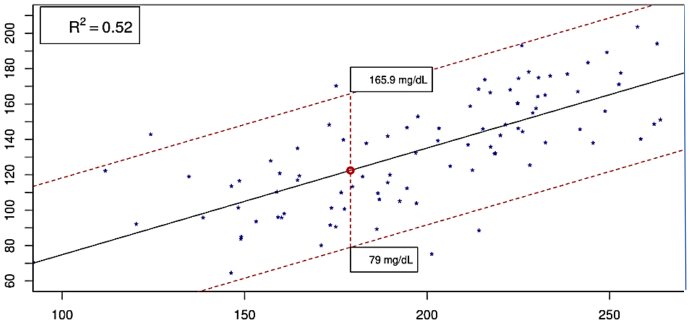
  This measures…

- As we can see on the left, if we have to predict an individual's LDL cholesterol level with **no predictor variable** to assist us, our best estimate would just be _________________ of our response variable.



- However, when we have a predictor variable like weight, we can actually make a prediction for Y given a specific value for X (perhaps Weight = 180lbs). Now, our best estimate would be the point on our best fit line at that given value for X.

- Next, we need a way to measure the variance in our prediction errors We call this the **Residual variance.**
    - It is notated as $\boldsymbol{\sigma_e^2}$ (and the variance of our sample is notated as $\mathbf{s_e^2}$)



This measures…

**Practice:** Which value would we expect to be less? $\mathbf{s_y^2}$ or $\mathbf{s_e^2}$?

- **Coefficient of Determination: $r^2$**
  - The coefficient of determination (abbreviated "**$r^2$**") is the proportion of variability in the response variable that is _______________ by the predictors in our model.
  - $s_Y^2$ measures the **total** variance present in Y (response variable).
  - $s_e^2$ specifically measures the **residual** variance (the **leftover variance**) after accounting for X (predictor variable)
  - Therefore, $s_Y^2$ - $s_e^2$ measures **the variance that IS explained by our model**.
    - Furthermore, we can divide this difference again by our starting variance of $s_Y^2$ in order to scale the value as a proportion. So this formula estimates the proportion of variance that is explained:

**Practice:** A linear model is created to explain the relationship between how many oz. of soda a person drinks a week on average and how high their blood sugar level is. The total variation within the blood sugar levels measured is 471.68. The variation among the residuals/errors after using blood sugar as a predictor of soda consumption is 386.21. Calculate $r^2$

**Interpret $r^2$:** Approximately ______% of the variability in blood sugar levels is explained by amount of soda the participants report consuming.