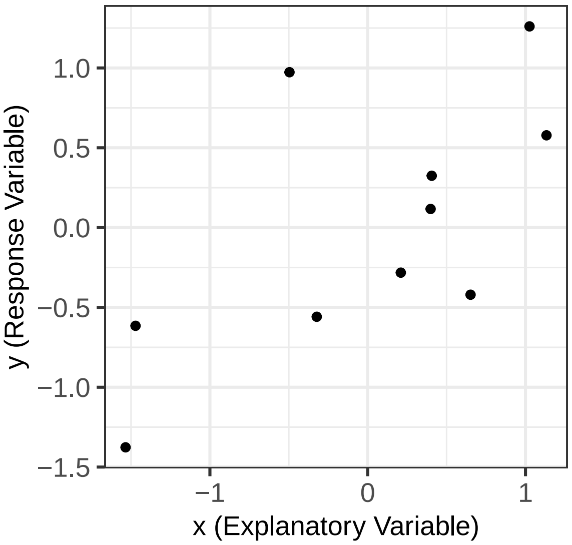
Chapter 9: Simple Linear Regression

In the previous chapters, we compared a numerical variable across two or more groups two-sample independent t-tests and paired t-tests and ANOVA. In this chapter, we will learn how to investigate the relationship between two numerical variables using a statistical analysis called simple linear regression

Overview

The principle of *simple linear regression* is to find the line (i.e., determine its equation) which passes as close as possible to the observations, that is, the set of points.

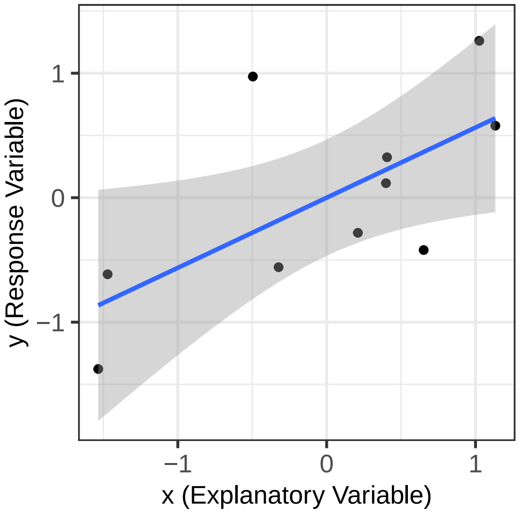
1. Draw a line through the set of points below.



2. Does your neighbor's look the same as yours?

Big Picture: Line of “Best” Fit

The principle of simple linear regression is to **find the line** (i.e., determine its equation) **which passes as close as possible to the observations**, that is, the set of points.



Equation of a line (y = mx + b):

- Estimated from sample: $\hat{y} = b_0 + b_1 \times x$
- Population true line: $y = \beta_0 + \beta_1 \times x + error$

Recall how $\bar{x}$ was our point estimate for μ and $\hat{p}$ was our point estimate for π.
Big Idea: We use least squares regression (aka math) to find the “best” line.

Example 8.1: Diving Penguins

Emperor penguins are the most accomplished divers among birds, making routine dives of 5–12 minutes, with the longest recorded dive over 27 minutes. These birds can also dive to depths of over 500 meters! Since air-breathing animals like penguins must hold their breath while submerged, the duration of any given dive depends on how much oxygen is in the bird’s body at the beginning of the dive, how quickly that oxygen gets used, and the lowest level of oxygen the bird can tolerate. The rate of oxygen depletion is primarily determined by the penguin’s heart rate. Consequently, studies of heart rates during dives can help us understand how these animals regulate their oxygen consumption in order to make such impressive dives. In this study, the researchers equipped emperor penguins with devices that record their heart rates during dives. The data set reports Dive Heart Rate (beats per minute), the Duration (minutes) of dives, and other related variables. Is there an association between dive heart rate and the duration of the dive?

In this study, the researchers equipped emperor penguins with devices that record their heart rates during dives. The data set reports Dive Heart Rate (beats per minute), the Duration (minutes) of dives, and other related variables.

Research Question Is there an association between dive heart rate and the duration of the dive?

```
diving <- read_csv("data/Diving_Penguins.csv")
head(diving)
```

```
# A tibble: 6 x 4
  Dive_HeartRate Depth Duration Bird
           <dbl> <dbl>    <dbl> <chr>
1           88.8   5       1.05 EP19
2          103.    9       1.18 EP19
3           97.4  22       1.92 EP19
4           85.3  25.5     3.47 EP19
5           60.6  30.5     7.08 EP19
6           77.6  32.5     4.77 EP19
```

1. What is the observational unit for this study?

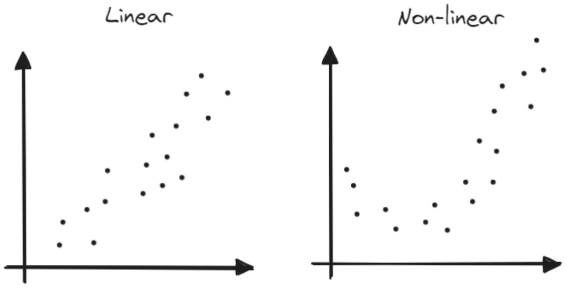
2. What are the variables assessed in this study? What are their roles (explanatory / response) and data types?

Scatter plot

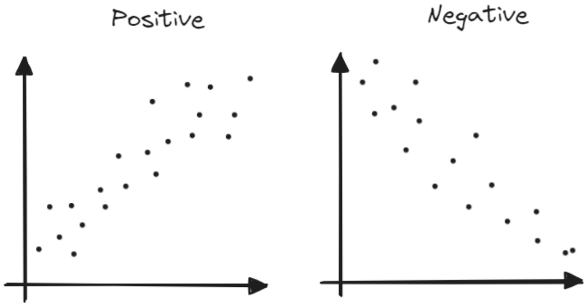
A *scatterplot* is a graph showing a dot for each observational unit, where the location of the dot indicates the values of the observational unit for both the explanatory and response variables. Typically, the explanatory variable is placed on the x-axis and the response variable is placed on the y-axis.
When describing a relationship or association between two quantitative variables as seen

through a scatterplot, we look for four aspects of association: *form, direction, strength, and outliers.*

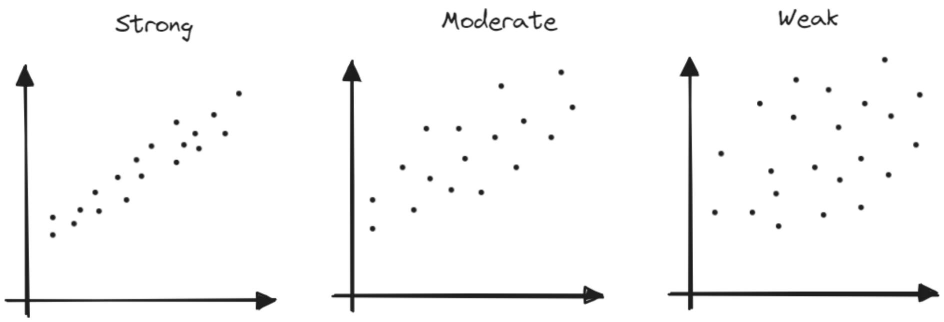
- The **form** of association between two quantitative variables is described by indicating whether a line would do a reasonable job summarizing the overall pattern in the data or if a curve would be better. It is important to note that, especially when the sample size is small, you don't want to let one or two points on the scatter plot change your interpretation of whether or not the form of association is linear. In general, assume that the form is linear unless there is compelling (strong) evidence in the scatter plot that the form is not linear.



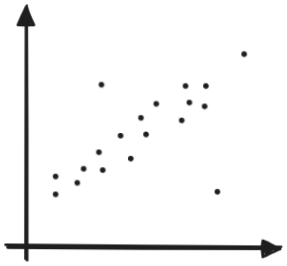
- The **direction** of association between two quantitative variables is either positive or negative, depending on whether or not the response variable (`Duration`) tends to increase (positive association) or decrease (negative association) as the explanatory variable (`Dive_HeartRate`) increases.



- In describing the **strength** of association revealed in a scatter plot, we see how closely the points follow a straight line or curve. If all the points fall pretty close to this straight line or curve, we say the association is strong. Weak associations will show little pattern in the scatter plot, and moderate associations will be somewhere in the middle.



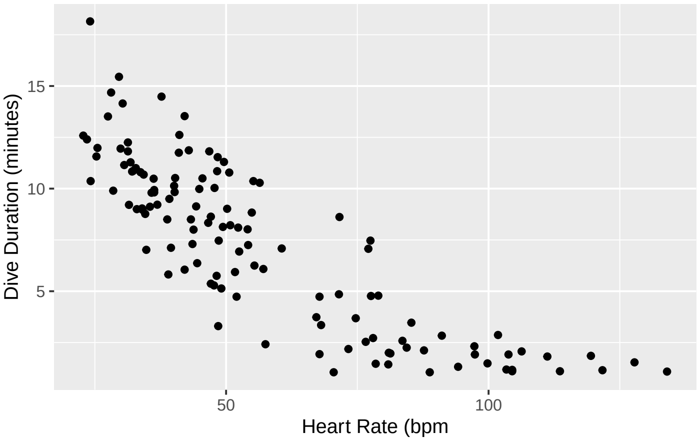
- When exploring the relationship and describing a scatter plot which visualizes two quantitative variables, we look for **unusual observations** or apparent **outliers**. Points which fall far from the trend of other points have the potential to be high leverage or high influential points.



3. Describe the association between the penguin heart rate (bpm) and dive duration (minutes) as revealed in the scatter plot below. Remember to use context of the research question.

```
ggplot(data = diving,
       mapping = aes(x = Dive_HeartRate,                         ①
                     y = Duration)                               ②
       ) +
  geom_point() +                                                 ③
  labs(x = "Heart Rate (bpm",
       y = "Dive Duration (minutes)"
       )
```

① In the `aes()` code, specify the `x = explanatory` and
② `y = response` variables.
③ Plot the `geom_points()` to create a scatter plot.



- Form:

- Direction:

- Strength:

- Unusual Observations / Outliers:

Correlation

Describing the direction, form, and strength of association based on a scatter plot, along with investigating unusual observations, is an important first step in summarizing the relationship between two quantitative variables. Another approach is to use a summary statistic. One of the statistics most commonly used for this purpose is the **correlation coefficient**. When the relationship has a roughly linear form, it's strength and direction can be quantified by the correlation.

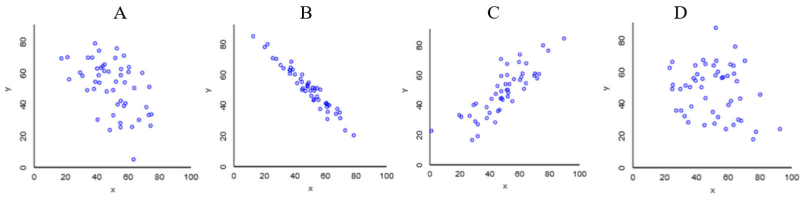
The sample correlation coefficient, denoted r, is a single number that takes a value between -1 and 1, inclusive. Negative values of r indicate a negative association, whereas positive values of r indicate a positive association. It is important to note that the correlation coefficient within a population is denoted ρ and r is an estimate of ρ.

⚠ The correlation coefficient is only applicable for data which has a linear form; non-linear data is not summarized well by the correlation coefficient. In fact, we could say that the correlation coefficient is a numerical summary of the **strength** and **direction** of a linear association between two numeric variables.

- Correlation measures the relationship between a pair of variables; the correlation is the same regardless of which one is explanatory and which is response. *(Be careful, the same is not true for regression coefficients!)*
- Correlation is a number without units. It is not a percent!
- The stronger the linear association is between the two variables, the closer the value of the correlation coefficient will be to either -1 or 1, whereas weaker linear associations will have correlation coefficient values closer to 0. Moderate linear associations will typically have correlation coefficients in the range of 0.3 to 0.7, or -0.3 to -0.7.
- Correlation can be sensitive to outliers and extreme values of either variable.

Practice Assign the following values of r, the correlation coefficient, to the data sets below.

-0.21, 0.83, -0.97, -0.5



The correlation coefficient uses a rather complex formula that is rarely computed by hand; instead, people almost always use a calculator or computer to calculate the value of the correlation coefficient. We will use the `moderndive` R package to obtain the correlation between two numerical variables. Specifically, we will use the `get_correlation()` function to obtain our observed sample correlation coefficient.

```r
library(moderndive)
get_correlation(data = diving,
                Duration ~ Dive_HeartRate)
```

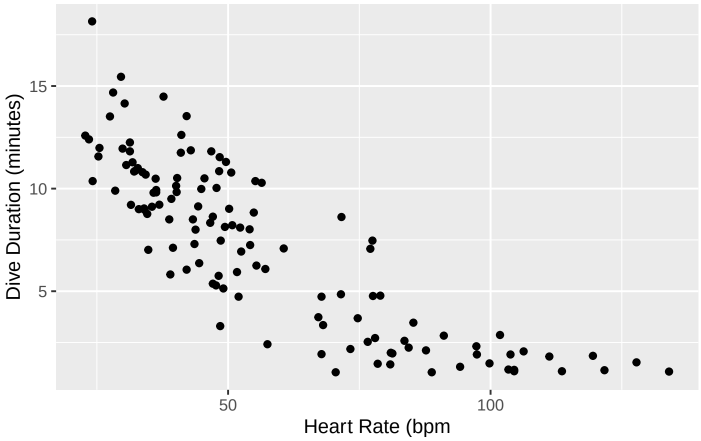
```
# A tibble: 1 x 1
     cor
   <dbl>
1 -0.846
```

4. If you knew the heart rate of a penguin, what might be a way to determine how long you would expect for them to dive for based on the data?

5. Using the scatter plot below, draw the *linear* line you believe fits the data the best. How did you decide where to draw your line? Is your line the same as your group members?

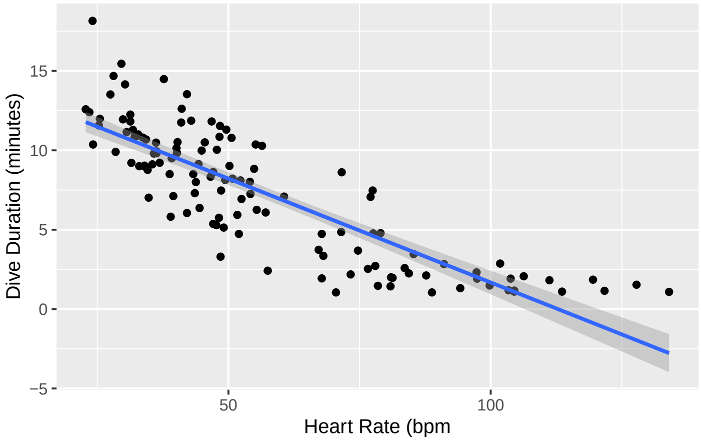
```r
ggplot(data = diving,
       mapping = aes(x = Dive_HeartRate,                       ①
                     y = Duration)                             ②
       ) +
  geom_point() +                                               ③
  labs(x = "Heart Rate (bpm",
       y = "Dive Duration (minutes)"
       )
```

① In the `aes()` code, specify the `x = explanatory` and
② `y = response` variables.
③ Plot the `geom_points()` to create a scatter plot.



```
ggplot(data = diving,
       mapping = aes(x = Dive_HeartRate,
                     y = Duration)
       ) +
  geom_point() +
  geom_smooth(method = "lm") +                                            (4)
  labs(x = "Heart Rate (bpm",
       y = "Dive Duration (minutes)"
       )
```

(4) To add a visual regression line, use `geom_smooth(method = "lm")`.



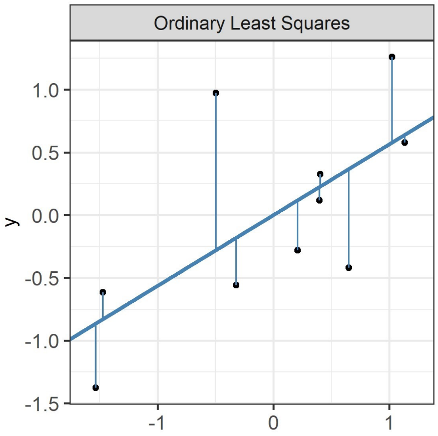
⚠ Warning

Using `geom_smooth()` is a visual representation of the linear regression and does not provide the regression coefficients as `lm()` described below does.

ℹ Least Squares Regression

How then do we decide what line is the "best"? In statistics we use a method called "least squares." The idea is that we minimize the sum of the squared distances between each point and the line. That's a mouthful! Let's visualize what this means.

These vertical distances are how far off your estimated duration is from what was actually seen in the data. These values are called ***residuals***. The least squares method finds the line that minimizes the *square* of these residuals.



6. On the scatter plot of the penguin heart rate and dive duration, draw the vertical distance between some of the points and the line you drew.

We will always use `R` to find the equation of the "best" regression line. Specifically, we will use the `lm()` function. The `lm` stands for *linear model* - the method we believe best models the relationship between our two variables.

```
diving_lm <- lm(Duration ~ Dive_HeartRate,                    ①
                data = diving)                                 ②

diving_lm |>                                                   ③
  tidy(conf.int = TRUE,                                        ④
       conf.level = 0.95)                                      ⑤
```

① Assign the model output to the object name `diving_lm <-`. Using the `lm()` function, specify the `response ~ explanatory` variables.
② Make sure to specify the `data =` set.
③ Using the object `diving_lm` you just created,
④ extract the `tidy()` table of coefficients with the `conf.int = TRUE` for confidence intervals, and
⑤ the `conf.level`.

```
# A tibble: 2 x 7
  term            estimate std.error statistic  p.value conf.low conf.high
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)       14.8      0.466       31.6 6.10e-61   13.8      15.7  
2 Dive_HeartRate    -0.131    0.00744    -17.6 2.35e-35   -0.146    -0.116
```

Warning

Make sure the order of your variables are input as `response ~ explanatory`!

7. Using the output from the `R` code above, write the equation of the regression line. Note that we've used variable names in the equation, not generic x and y. And put a carat ("hat") over the y-variable name to emphasize that the line gives **predicted** values of the y (response) variable.

$$\text{Dive } \widehat{\text{Duration}} = \underline{\hspace{3cm}} + \underline{\hspace{3cm}} \times \text{(Heart Rate)}$$

Notation

The equation of the best fit line is written as $\hat{y} = b_0 + b_1 \times \text{(x)}$ where

- b_0 is the y-intercept coefficient
- b_1 is the slope coefficient
- x is a value of the numeric explanatory variable
- $\hat{y}$ is the predicted value for the response variable

8. Is the slope positive or negative? Explain how the sign of the slope tells you about whether your data display a positive or a negative association.

9. Use the least squares regression line to predict the diving duration for penguins with a heart rate of 75 beats per minute.

   Dive $\widehat{\text{Duration}}$ = ________________ + ________________ × 75 = ________________

10. Use the least squares regression line to predict the diving duration for penguins with a heart rate of 76 beats per minute.

    Dive $\widehat{\text{Duration}}$ = ________________ + ________________ × 76 = ________________

11. By how much do your predictions in the above two questions differ? Does this number look familiar? Explain.

12. These questions above were designed to help you interpret the slope. Interpret the slope in context:

    The slope of the regression line predicting dive duration based on heart rate is ________________, meaning that for every ________________ beat per minute increase in heart rate, the predicted dive duration (increases / decreases) by ______ minutes.

Interpretation: Slope

The slope coefficient of a least squares regression model is interpreted as the predicted change in the mean response (y) variable for a one-unit change in the explanatory (x) variable.

13. Use the least squares regression line to predict the diving duration for a penguin with a heart rate of 0 beats per minute.

    Dive $\widehat{\text{Duration}}$ = ________________ + ________________ × 0 = ________________

14. Your answer to the above question should look familiar. What is this value?

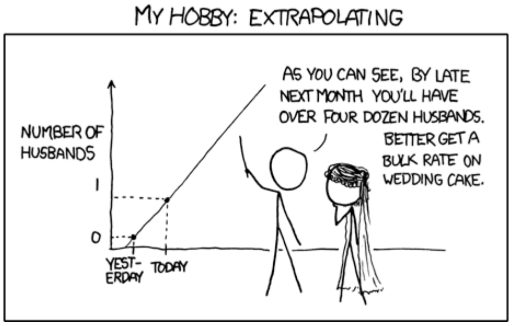
Interpretation: y-intercept

The y-intercept of a regression line is interpreted as the predicted value of the response variable when the explanatory variable has a value of zero.

Extrapolation

While we can make predictions using our least squares regression line, we should always be wary of extrapolation in interpreting the intercept or other values outside the original data range.

Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is known as *extrapolation* and can give very misleading predictions.


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.

15. What was the lowest value of heart rate observed in these data? The highest?

```
favstats(~Dive_HeartRate, data = diving)
```

```
 min   Q1 median   Q3 max   mean       sd   n missing
22.8 36.3   48.4 76.6 134 56.924 26.34303 125       0
```

16. What heart rates do you believe would be an extrapolation?

ℹ Coefficient of Determination (R^2)

A quantity related to the correlation coefficient (r) is called the coefficient of determination or R-squared (R^2). The coefficient of determination (R^2) is the percentage of total observed variation in the response variable that is accounted for by changes (variability) in the explanatory variable.

Keep in mind that R^2, like correlation, requires that the relationship between the explanatory and response variables is linear!

R^2 values are reported as proportions, but can also be thought of as a percent. Values close to 1 or 100% indicate that the explanatory variable is able to explain a large portion of the variability in the response.
We calculate R^2 by squaring the correlation (r).

17. Calculate the coefficient of determination (R^2) for the relationship between heart rate and dive duration.

18. Complete the following statement to interpret what this value means in the context of the data:

The coefficient of determination is _____%, this means that _____% of the variation in penguin's _______________is attributable to changes in their _______________.

Let's revisit the **Research Question**! Is there an *association* between dive heart rate and the duration of the dive?

19. Set up the null and alternative hypotheses.

- In words (version 1):

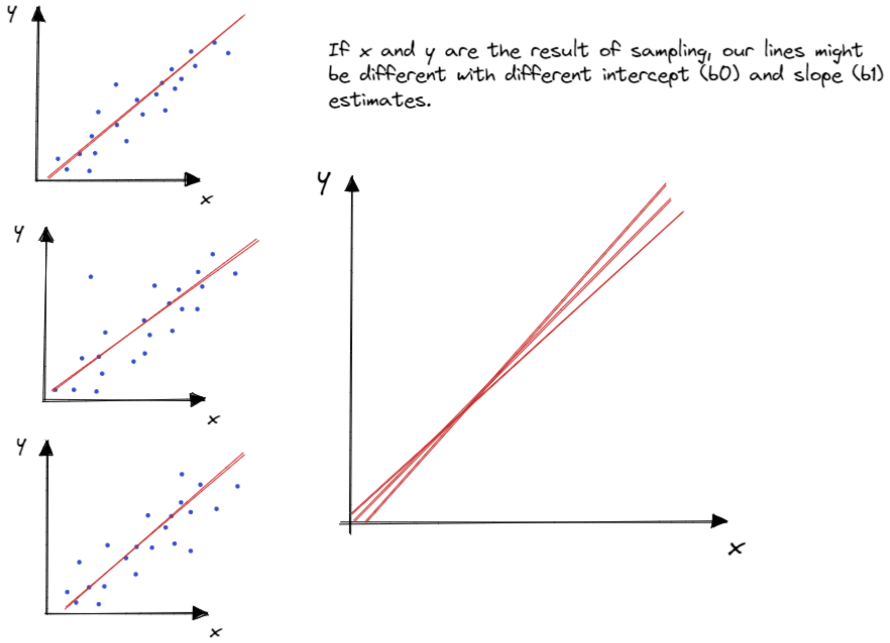
- In words (version 2):

- In symbols:

20. Based on the slope coefficient found from our regression model, do you think there is convincing evidence of an association between a penguin's heart rate and the duration of their dive?

Inference for the Slope

Recall, we want to use our data to draw inferences and make claims about the larger population. Since the slope will change from sample to sample, in order to make valid inferences about the true population slope, we must first understand the sample variability of the slope. That is, we must determine what slope coefficients are likely to happen by chance when taking random samples from populations with the relationship between heart rate and dive duration.


y
x
y
x
y
x
If x and y are the result of sampling, our lines might be different with different intercept (b0) and slope (b1) estimates.
y
x

In order to test the slope, we could conduct a simulation similar to what we did with two categorical variables (yawn experiment) and discussed when comparing one numeric variable across two groups. Recall, we:

- Step 1: Write the ____________________ and ____________________ on __________ cards.

- Step 2: Simulate what could have happened if the null was true and ___________________________.

- Step 3: Generate a new data set by ____.

- Step 4: Calculate the ____ for the new simulated data set and add it to the dot plot.

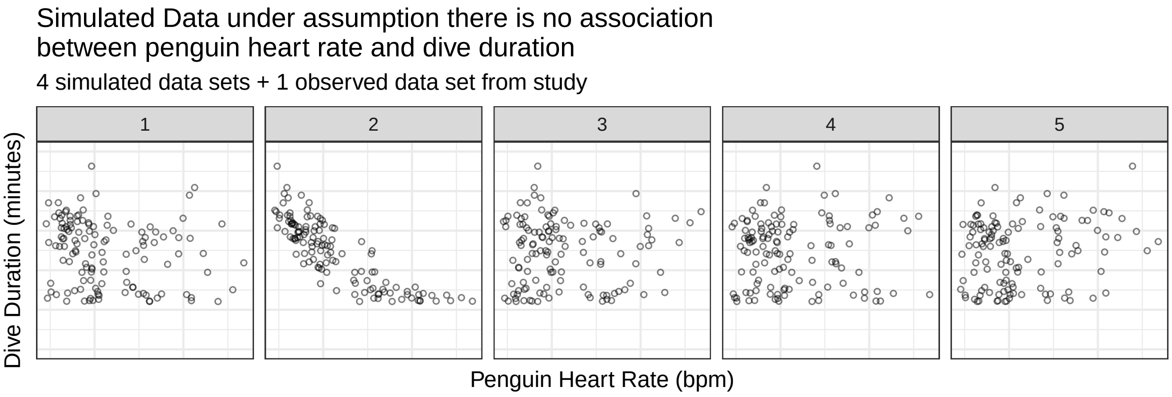
We would then repeat this process 100 or 1000 times to get an idea of what the sampling distribution of the *slope* looks like.

21. Assuming there is *no relationship* between the penguin heart rate (bpm) and dive duration (minutes), where do you expect the sampling distribution of the slope to be centered? Explain.

22. Suppose we wanted to complete the simulation using correlation as the summary measure, instead of slope. What would you calculate in Step 4 instead?

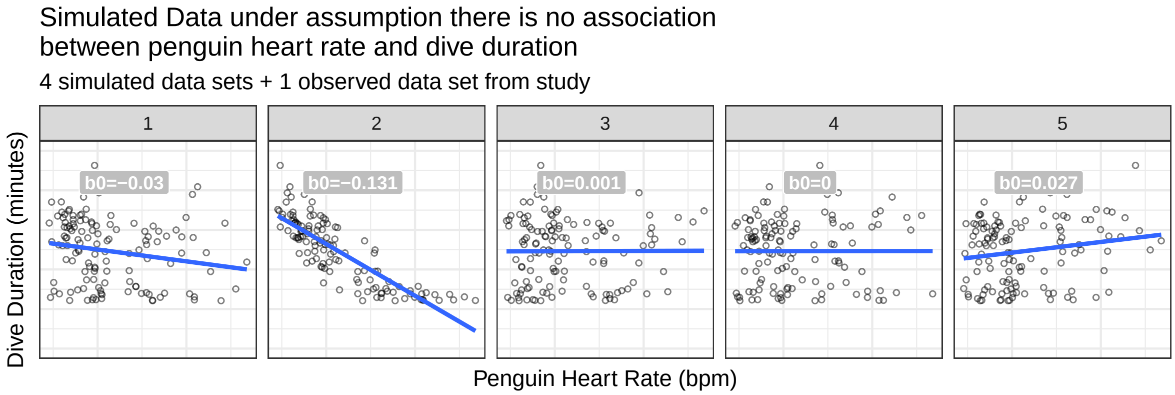
- Step 4: Calculate the ____ for the new simulated data set and add it to the dot plot.

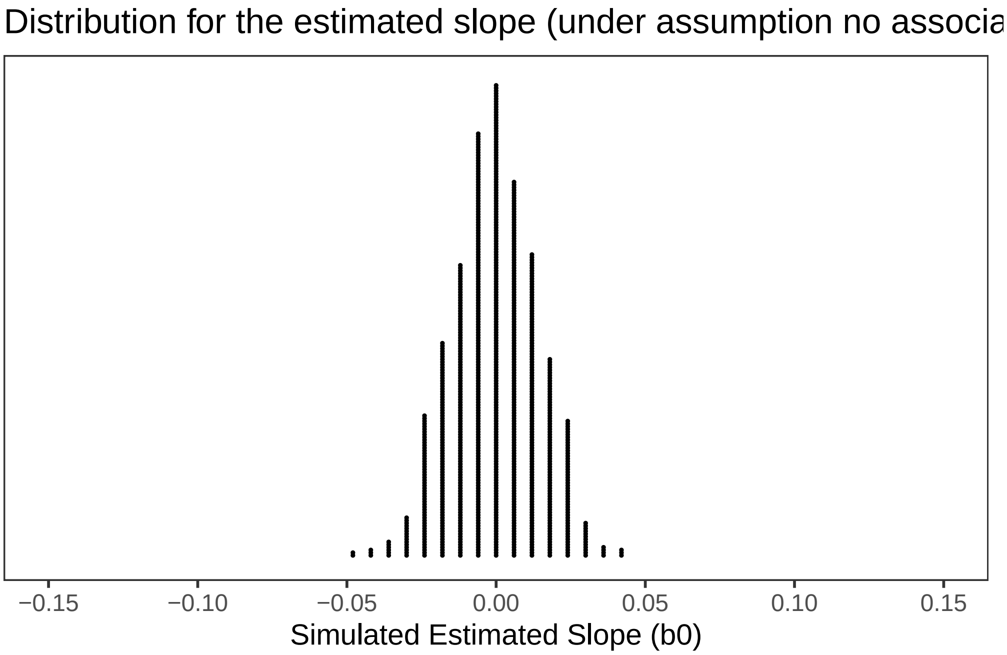
The plot below contains 4 panels of "new simualted" data sets (under the assumption there is *no association* between penguin heart rate (bpm) and dive duration (minutes), following the process above) and one panel of the actual data observed from the study.



23. Which panel contains the actual data observed in the study? Was it hard to pick out? Remember what it was like trying to pick this out.

We could take these and calculate the regression equation for each panel (simulated data set). Recall, we are really interested in the estimated slope coefficient. We can plot this value to begin creating the distribution to compare our observed slope to.





24. Take note that our observed slope was -0.13, do you believe this slope is likely to occur under the condition that there is *no association* between penguin heart rate and dive duration (i.e., the null is true)?

It turns out that these slopes vary in a predictable way following a distribution we already know – the t-distribution! Therefore, we use a t-statistic as our test statistic for conducting a hypothesis test on the slope coefficient.

Recall our regression output from earlier:

```
# A tibble: 2 x 7
  term           estimate std.error statistic  p.value conf.low conf.high
  <chr>             <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)      14.8      0.466       31.6 6.10e-61   13.8      15.7
2 Dive_HeartRate   -0.131    0.00744    -17.6 2.35e-35   -0.146    -0.116
```

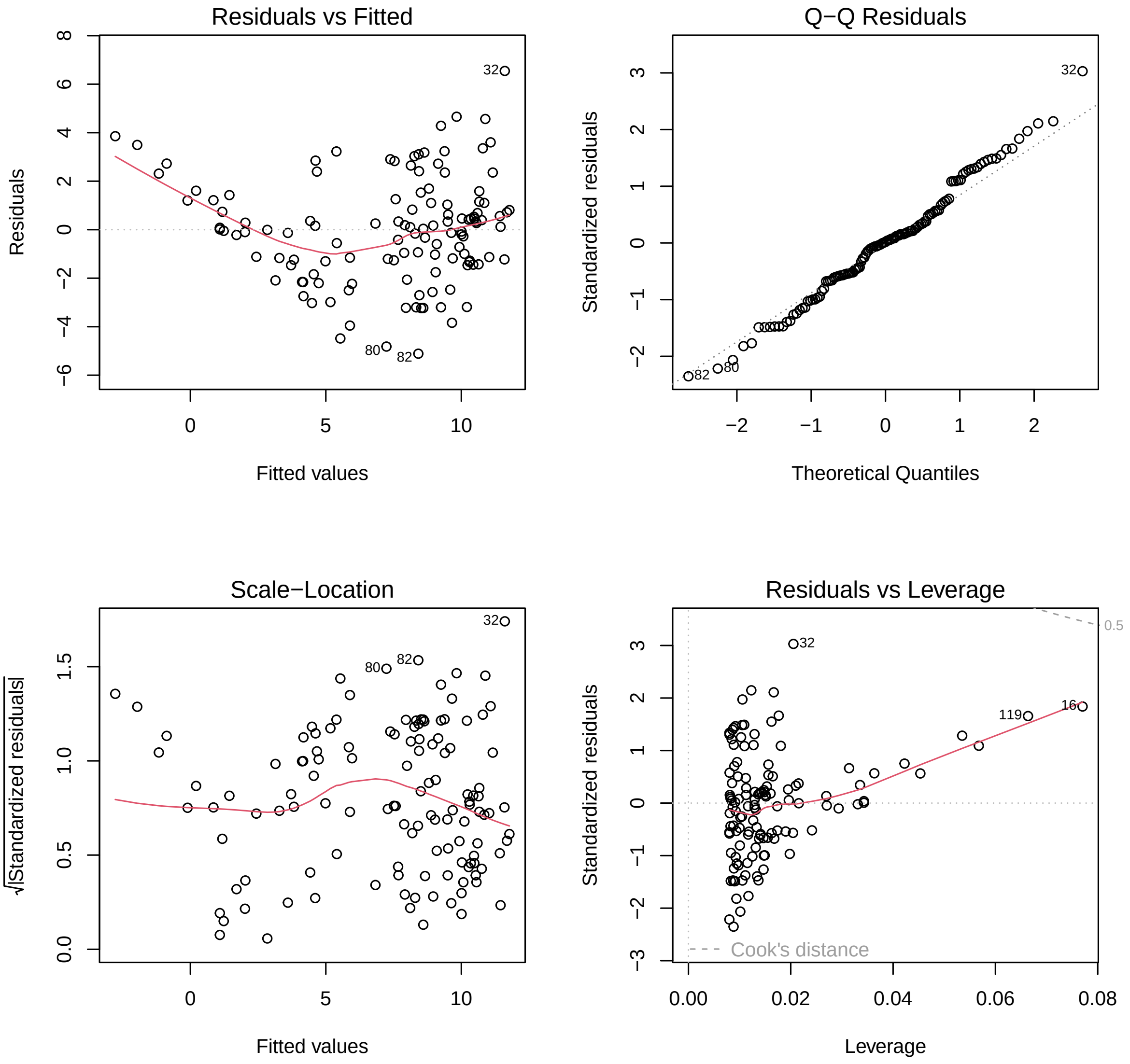
25. Based on the output above, write a conclusion for the research question in context of the problem.

Conditions for using simple linear regression (LINE)

1. Linearity
2. Independent observations
3. Normality of response variable
4. Equal variance

```
par(mfrow = c(2,2))                                                      ①
plot(diving_lm)                                                          ②
```

① Set the plotting window to give a 2x2 grid of plots.
② Use the `plot(model_name)` function to plot the residuals for model diagnostics.



```
par(mfrow = c(1,1))                                                    ③
```

③ Reset the plotting window back to a 1x1 grid.

27. Note the 95% confidence interval for the true slope (β_1) is (-0.146, -0.116). Interpret this interval in context of the problem.

28. Why does it make sense that the 95% confidence interval does not contain 0?