Lab 4: Golf Driving Distance

Present Group Member's Names Go Here!

Remember the questions you should answer in Canvas are boldfaced! The Canvas questions may be extensions/requirer thinking deeper about these questions.

Golf driving distance

In golf the goal is to complete a hole with as few strokes as possible. A long driving distance to start a hole can help minimize the strokes necessary to complete the hole, as long as that drive stays on the fairway. Data were collecting on 354 PGA and LGPA players in 2008. For each player, the average driving distance (yards), fairway accuracy (percentage), and sex was measured. Use these data to assess, "Does the accuracy of a professional golfer change when they hit the ball farther?"

```
golf <- read_csv("data/golf.csv")
```

Scenario/Context Setup Reminders (fill in the blanks)

- Observational Unit: _____
- Explanatory Variable (and type): _____
- Response Variable (and type): _____
- Population of Interest: _____
- Parameter of Interest: _____

Ask a research question

1. Write out the null and alternative hypotheses in words.

Creating a Scatterplot

Use the skeleton R code below, fill in the remaining pieces to create a scatterplot to examine the relationship between the driving distance and percent accuracy. You should use `Driving_Distance` and `Percent_Accuracy` for `x` and `y`, respectively.

```
ggplot(data = golf,
       mapping = aes(x = _______, y = _______)
       ) +
  geom_point() +
  labs(x = "_______",
       y = "_______")
```

2. Does it appear that there is a relationship between driving distance and percent accuracy? Explain. Note: `Driving Distance` should be on the x-axis.

Adding a regression line

You can add a regression line on top of the scatterplot by adding an additional layer to your plot. Specifically, `geom_smooth()` adds a smoothed line on top of a plot. There are many different types of lines we could use, but we want a straight line. So, we need to tell `geom_smooth()` to use the `"lm"` method to get a straight line!

3. Add one more line of code to your scatterplot (at the end of the code above!) to plot a linear regression line to the plot! *Hint: look at week 3 & 4 R resource note cards.*

Conditions for the least squares line

When performing inference on a least squares line, the follow conditions are generally required (recall LINE from the reading):

Independent observations (for both simulation-based and theory-based methods): individual data points must be independent.

- Check this assumption by investigating the sampling method and determining if the observational units are related in any way.

4. Do you believe this condition (independent observations) is violated? Why or why not?

Linearity (for both simulation-based and theory-based methods): the data should follow a linear trend.

- Check this assumption by examining the scatterplot of the two variables, the relationship on the scatterplot should appear linear.

5. Do you believe this condition (linearity) is violated? Why or why not?

Equal/Constant variability (for theory-based methods only): the variability of points around the least squares line remains roughly constant

- Check this assumption by examining the scatterplot of the two variables, the variability in the points around the regression line should be approximately the same for all of the values of x. When I say the "variability in the points," I mean you should be looking at the **vertical** spread of the points.

Nearly normal residuals (for theory-based methods only): residuals must be nearly normal.

- Check this assumption by examining a histogram of the residuals **or** a histogram of the response variable (y). The distribution of either of these values should appear approximately normal.

Using theory-based methods

To use a t-distribution as an approximation for the true sampling distribution, we need to verify that the last two conditions **are not** violated. If these conditions are violated, then we will need to use a simulation-based method (e.g. permutation or bootstrapping) instead.

6. Based on the scatterplot, do you believe the condition of equal / constant variability is violated? Why or why not?

To check the condition of normality, we can use either the distribution of responses or the distribution or residuals. To obtain the residuals we have to first fit the linear regression!

To find the equation of the regression line, we use the `lm()` function. This function takes two arguments (1) the variables we are using for our line and (2) the dataset the variables live in. We have to specify the variables using `y ~ x` notation.

Using the skeleton code below, fill in the explanatory and response variables for our regression.

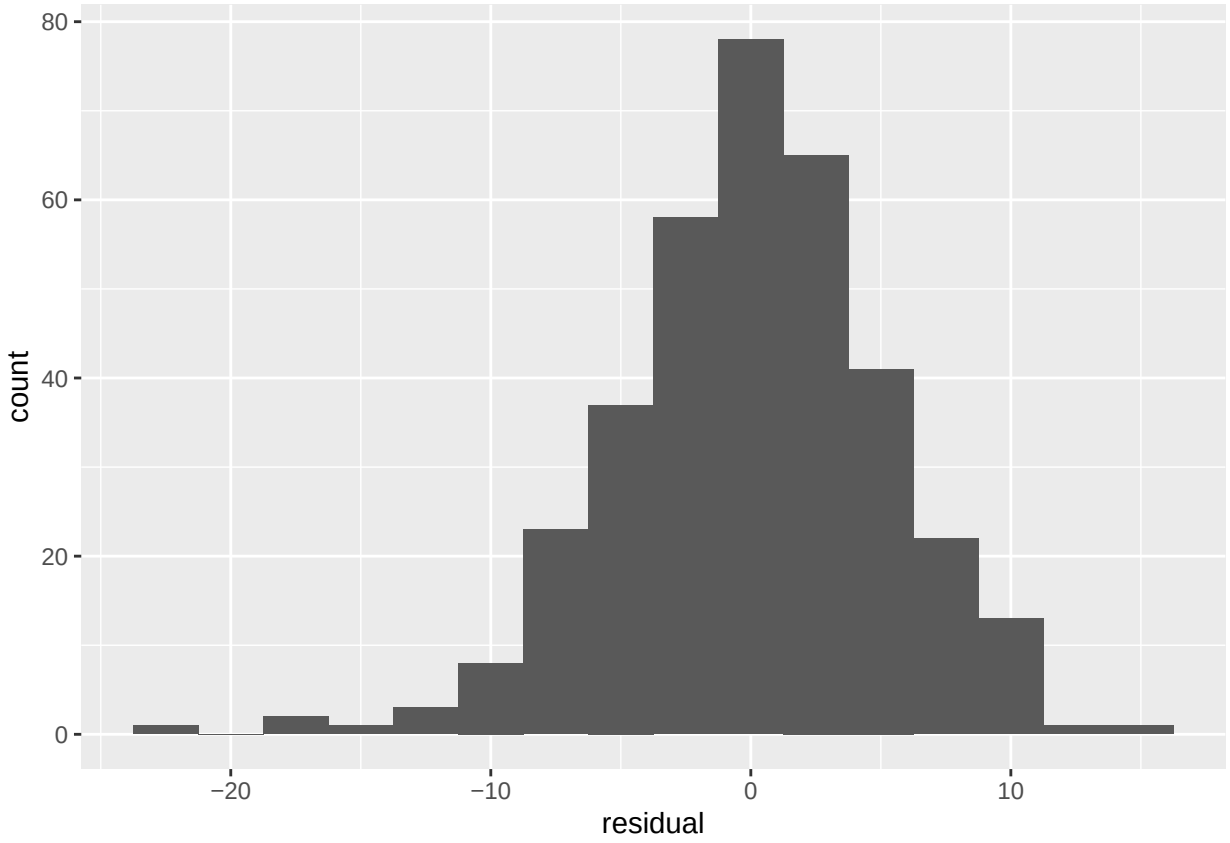
```
golf_lm <- lm(________ ~ ________, data = golf)
```

Alright, now that we've found the equation of the regression line, we can find the values of the residuals.

7. How would you find the residual for a golfer in these data?

To obtain the residuals, we use the `get_regression_points()` function.

```
get_regression_points(golf_lm) %>%
  ggplot(mapping = aes(x = residual)) +
  geom_histogram(binwidth = 2.5)
```



Let's compare this to the distribution of the Percent Accuracy!

8. Create a histogram of the Percent Accuracy variable.

```
# Histogram code goes here.
```

9. Do you believe the condition of normality is violated? Why or why not?

10. Do you believe it is "safe" to use the** t**-distribution to approximate the sampling distribution? Justify your answer.

Obtaining the regression line, p-value, and confidence interval

To obtain information about the regression line, we will use the `get_regression_table()` function. The function takes one input, the name of the linear regression. Remember, we stored our linear regression in an object named, `golf_lm`, so that is what we will input here!

```
get_regression_table(_______)
```

11. Using the output from the code above, write the equation of the regression line in the context of the problem.

12. Interpret the estimated slope in context of the problem.

Use statistical inferential methods to draw inferences from the data

Hypothesis test

To find the value of the standardized statistic to test the slope we will use,

$$T = \frac{\text{slope estimate} - 0}{SE} = \frac{b_1}{SE(b_1)}.$$

We will use the output from our regression table to obtain the estimate for slope and the standard error of the slope.

13. What are the values of b_1 and $SE(b_1)$?

14. Calculate the t-statistic for slope. Where is this value located on the regression table?

15. The p-value reported in the regression table is the two-sided p-value for testing if the slope is different from 0. Report the p-value for this test.

16. It is standard to assume a 95% confidence level. Based on the p-value, how much evidence is there against the null hypothesis? *(This is asking you to think about the strength of statistical significant evidence and what that means you conclude. What are you comparing your p-value to? What then do you conclude?.)*

Confidence interval

Recall that a confidence interval is calculated by adding and subtracting the margin of error to the point estimate.

$$\text{point estimate} \pm t^*_{df} \times SE(\text{estimate}).$$

When the point estimate is a regression slope, this formula becomes

$$b_1 \pm t^*_{df} \times SE(b_1).$$

The t^* multiplier comes from a t-distribution with $n - 2$ degrees of freedom. Recall for a 95% confidence interval, we use the 97.5% percentile (95% of the distribution is in the middle, leaving 2.5% in each tail).

17. The sample size for this study is 354. How many degrees of freedom will we use?

Fortunately, the regression table also outputs a 95% confidence interval. This interval was calculated using a t-distribution with the number of degrees of freedom you just calculated.

18. Report the lower and upper values of the confidence interval for the true slope. *e.g. (lower bound #, upper bound #)*

19. Interpret the 95% confidence interval in context of the problem.

20. Based on your 95% confidence interval, what decision about your hypotheses would you make? *(Hint: Reject or Fail to Reject)* Why?

21. Does your decision based on your confidence interval match the decision you made based on your p-value in question 16? Is this surprising? Explain.

22. Write a conclusion in context of the data. *(Tell me in words what you think is going on; look back at your null and alternative hypotheses.)*