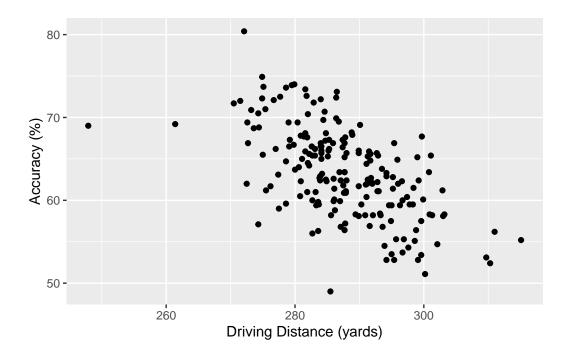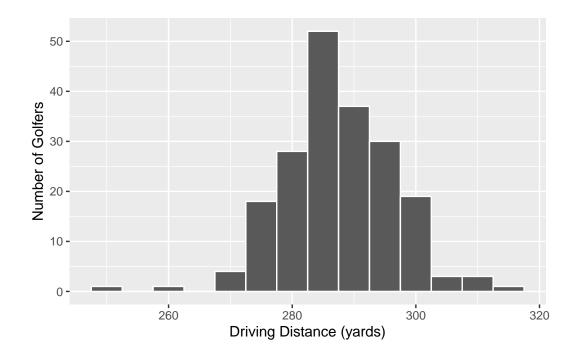# Homework 9: Golf Driving Distances
Simple Linear Regression

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.

```
golf_data <- read_csv("data/golf.csv") |>
  filter(Sex == "male")
head(golf_data)
```

```
# A tibble: 6 x 3
  Driving_Distance Percent_Accuracy Sex
             <dbl>            <dbl> <chr>
1             290.             59.5 male
2             302.             54.7 male
3             287.             62.4 male
4             283.             65.4 male
5             299.             52.8 male
6             300.             51.1 male
```
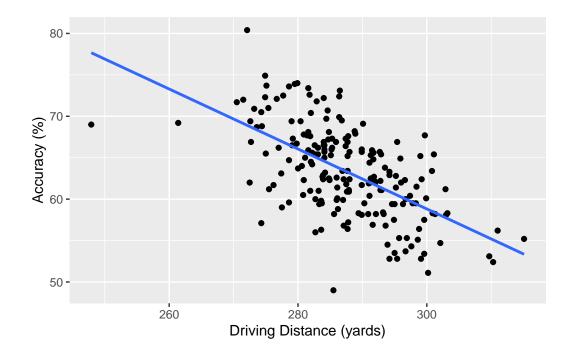
Research Question: Does the accuracy of a professional golfer change when they hit the ball farther?

```
ggplot(data = golf_data,
       mapping = aes(x = Driving_Distance,
                     y = Percent_Accuracy)
       ) +
  geom_point() +
  labs(x = "Driving Distance (yards)",
       y = "Accuracy (%)")
```

```
ggplot(data = golf_data,
       mapping = aes(x = Driving_Distance
                     )
       ) +
  geom_histogram(binwidth = 5, color = "white") +
  labs(x = "Driving Distance (yards)",
       y = "Number of Golfers")
```

1. Identify the following in context of the problem:

- Observation:

- Explanatory Variable (and data type):

- Response Variable (and data type):

- Population of Interest:

- Sample (and Sample Size):

2. Set up the null and alternative hypotheses:

- In words.

- In symbols

3. Based on the scatterplot above, does it appear that there is a relationship between driving distance and percent accuracy? Explain.

Note that we can add the line of best fit (linear regression line) on top of the scatterplot by adding an additional layer, `geom_smooth(method = 'lm', se = F)` to the code. Notice that `method = 'lm'` tells the smoothed line to be a straight line (linear regression model).

```
ggplot(data = golf_data,
       mapping = aes(x = Driving_Distance,
                     y = Percent_Accuracy)
       ) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(x = "Driving Distance (yards)",
       y = "Accuracy (%)")
```
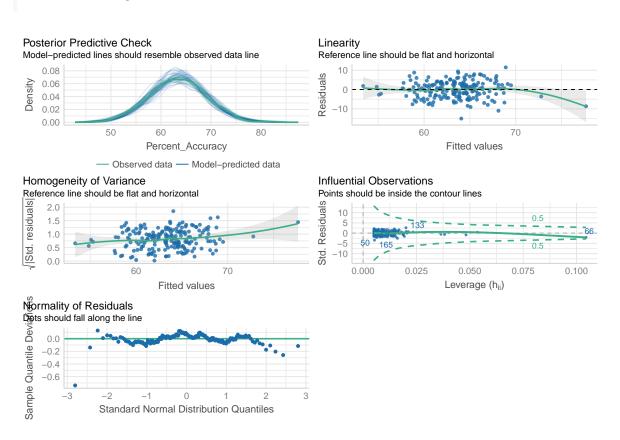

4. Set up the null and alternative hypotheses for testing for an association using the slope.

5. Fill in the code below as if you were running a linear regression model.

```
golf_model <- lm(_____ ~ _____,
                 data = golf_data)

golf_model |>
  tidy(conf.int = TRUE,
       conf.level = TRUE)
```

```
# A tibble: 2 x 7
  term            estimate std.error statistic p.value conf.low conf.high
```

<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>
1 (Intercept)	167.	10.1	16.6	<0.001	148.	187.
2 Driving_Distance	-0.362	0.035	-10.4	<0.001	-0.431	-0.293

6. Check the assumptions/conditions for using simple linear regression on our data:

```
library(easystats)
check_model(golf_model)
```


- Independent observations: 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.

- Linearity: 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.
- – Also see top right panel in the plot above.

- Equal/Constant variability: 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.
  - – Also see middle left panel in the plot above.

- Normality: residuals must be nearly normal.

  - – Check this assumption by examining a histogram of the response variable (y).
  - – Also see top and bottom left panels in the plot above.

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

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

9. State your conclusion in context of the research question. Cite evidence to support your conclusion.

10. Interpret the 95% confidence interval for the slope.

11. Suppose you have a golfer with a driving distance of 280 yards. What would you expect their accuracy to be?

12. Assume that same golfer has an actual accuracy of 71%, what is the residual for this golfer?

13. Instead, assume that same golfer has an actual accuracy of 55%, what is the residual for this golfer?

🔥 Canvas Quiz (Optional)

Make sure to complete the Homework Quiz on Canvas.