**NHL Data: Linear Regression**

1. Do you think that teams who score more goals (higher *GF*) win more games (higher PTS)?

Yes. A team that scores more goals is more likely to have more than their opponent and therefore win the game.

1. Plot *PTS vs GF*. Does there appear to be a relationship between the two variables? If so, describe it.

A graph with numbers and dots

Description automatically generated

There appears to be a fairly strong positive relationship.

1. Fit a linear model using *GF* to predict *PTS*.

lm(PTS~GF, data = nhl\_2223)

* 1. Write down the prediction equation.
  2. Interpret the slope in the context of what the equation is predicting.

For every goal a team scores, the number of points they win during the season increases by 0.5097.

* 1. Is *GF* an effective predictor of *PTS*? What part of the output did you use to make this decision?

Yes. The p-value for its individual t-test is 4.64 x 10-8, which is small. This means that you can reject the null hypothesis that the coefficient is 0, making it an effective predictor.

* 1. Calculate the residual for the Toronto Maple Leafs (*Team* = TOR).

Use the prediction equation to calculate the predicted PTS. The Leafs scored 279 goals.

Residual = actual – predicted

Residual = 111 – 100.76

Residual = 10.24 points

1. Fit a linear model to predict *PTS* using *GA.*

lm(PTS~GA, data = nhl\_2223)

* 1. Write down the prediction equation*.*
  2. Based on the coefficients, describe the relationship between *PTS* and *GA*.

There is a negative relationship. As a team lets in more goals, they get less points.

* 1. How many points is a team the lets in 219 goals predicted to win?

1. Fit a linear model using *DIFF* to predict *PTS*.

lm(PTS~DIFF, data = nhl\_2223)

* 1. Write down the prediction equation.
  2. What is the R2 value? Interpret this in context.

R2 = 0.963. 96.3% of variability in PTS can be explained by goal differential.

1. Which of the three models is the most effective for predicting *PTS*? Explain your answer using evidence from your output.

The model using *DIFF* is the most effective.

3 possible explanations:

* Since these models use a single predictor, we can look at the individual t-tests. The p-value for *DIFF* is the smallest, meaning that it is the “most different” from zero of the three predictors, making it the most effective.
* Look at the R2 values. The model using *DIFF* has the highest R2, meaning it explains the most variability in *PTS*.
* Look at the residual standard error in the output. *DIFF* has the lowest one, meaning the predicted values are closest to the actual ones.