Fastball Velocity

1. My dependent variable is the average fastball velocity for each pitch number during a game, with all pitchers included.
2. My independent variable is the pitch number during a game.
3. The shape is definitely not linear. A polynomial regression is clearly the best fit for the model between pitch number and average fastball velocity for a team at that pitch number. The graph created has two distinct turns along the line of best fit, which makes sense due to the cubic nature of the equation. Two turns makes sense in the real world sense as well. A pitcher's velocity may be low to start the game, but it starts to increase as their body becomes warmed up. The second turn in the graph at around 80 pitches is logical as well, as a pitcher's velocity begins to drop when they grow tired.
4. For the model with pitch number as the independent variable and differential as the dependent variable, there wasn't a polynomial regression model that had a high R-squared. I did a simple linear regression, a quadratic polynomial, and a cubic polynomial. The p-values for all of these models were very low, which suggests that that all three models are statistically significant. However, the highest R-squared is 3.76% from the cubic polynomial model. This suggests that the best model for this relationship is most likely a very specific formula.

Shuttle Time

1. Based on the two simple linear regression models, the R-squared between weight and shuttle time (27.34%) is much larger than the R-squared between height and shuttle time (1.54%). Even though the estimated slope coefficient is larger for height, the R-squared metrics suggest that weight is more predictive of shuttle time. The relationships are not perfectly linear, especially considering the outliers in the data. This suggests that different models are most likely more appropriate than linear ones.
2. When both predictor variables are in the multiple linear regression model, the relationships between the predictors and shuttle time change. The p-values for the coefficients of both predictor variables are extremely low, which suggests that both predictor variables are statistically significant even in the presence of the other variable. The estimated slope coefficients are small, but they are still significant. This is because the dependent variable is measured in seconds, which is typically a small number for shuttle time. The relationships changed because of the lurking variables in the original two simple linear regression models. Because each of the first two models excluded a predictor, the estimated slope coefficients for the included variable also captured the impact of the excluded variable. The multiple linear regression model controlled for this lurking variable bias.
3. My polynomial model uses weight and weight squared as the two predictor variables. The predicted shuttle times were calculated using this quadratic equation. The residuals were calculated a little bit differently than a typical regression. Because a smaller time is desired, the actual shuttle time was subtracted from the predicted shuttle time to create the residuals. For the adjusted shuttle time metric, I created percentile ranks for each residual. Player 1268 benefited the most from having their time adjusted. This player placed in the 100th percentile because they performed 1.31 seconds faster than their predicted shuttle time.

NBA Usage and Points Per Game

1. Including the interaction term Usage*Minutes provides more explanatory power than treating them as separate inputs because Usage and Minutes do not independently affect total points. The multiple linear regression model with the interaction term shows how Usage and Minutes interact with each other when predicting total points. From a basketball perspective,

Usage Rate and Minutes Played affect total points together. If a player has a high Minutes Played, but has a low Usage Rate, then their total points might be low. If a player has a high Usage Rate, but a low Minutes Played, then their total points still might be low. If they have a high number for both metrics, then their total points might increase at a unique rate.

2. A one unit increase in USG% has more of an impact on points if a player’s minutes played are high. Although the slope coefficients for each individual variable are negative, the estimated slope coefficient for the interaction term is positive. This suggests that if minutes played for a player is high, then a unit increase in USG% has a larger impact on points scored. This makes sense, as usage rate’s effect on points matters much more when minutes played are high.