Assignment 4 - Alyssa Rossi

Linear Regression is a strong statistical method used to model the relationship between two variables having a line of best fit. There is an independent variable which is known as the predictor and the dependent variable which is the outcome we'd like to predict. The goal of linear regression is to find the best fitting line that reduces the difference between the observed data points and predicted data. In this assignment, we needed to predict the *Men's 100 meters world records progression* for the years 2023 to 2030. In order for us to minimize our data, we needed to take into consideration outliers. By looking at our dataset, we can see by removing one outlier can have a significant difference in our dataset. In our *Linear Regression 1* model, we have all the dataset (including the outlier) we can see that the line of best fit has a beta coefficient β_1 = - 0.00743113 and β_0= 24.64808957 with an R-Squared of approximately 83.6% and a p-value that is less than 0.001, indicating that the model is significant. An R-Square of 83.6% is very good and explains most of the historical trend in men's 100m of world record. However, 16.4% of the variation is due to other factors or simply randomness. Looking at *Linear Regression 2* model, we can see there is a slightly higher R-Square of 87.8% with beta coefficients of β_1 = -0.00687737 and β_0= 23.551246054891937. This is because an outlier has been removed from our dataset, indicating that having higher R-Squared, cleaner residuals and more stable predictions is more suitable for our data. We can see that in the *Linear Regression 1* model that the skewness of residuals is non-normal. It is heavily-tailed with a skewness of -1.444. On the other hand, in *Linear Regression 2* model we have a skewness of -0.929 where residuals are nearly normal. We can see that without an outlier, the omnibus is 4.867 (p-value≈0.088) compared to 12.638 showing that the residuals are consistent with a normal distribution. In essence, both models provide excellent goodness of fits as most of the data surrounds the regression lines. Looking at the F-test, both models were statistically significant confirming that year is a strong predictor of *Men's world records progression*. With no outlier F-test = 150.9. Limitations of our dataset could be the lack of accounting for other variables, only having a single predictor such as "year". Other factors such as training methods, tracking technology, athletic performance were not counted for. These limitations could significantly impact the outcome of our model. Another limitation is having a small dataset, world record datapoints are small and spread out, which can lower statistical power and make the regression sensitive to outliers.

Figure 1: Predicted Values with Original Data

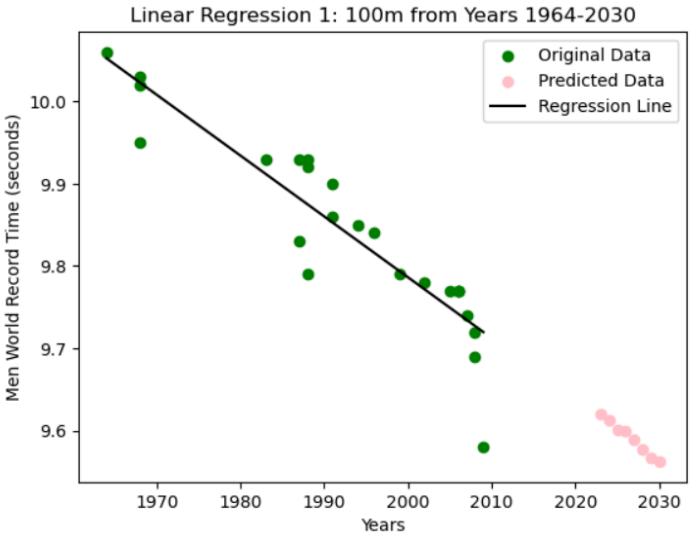

Figure 2: New Data with Predicted Values and no outliers