Nikolas Argiropoulos

Simple linear regression is a common statistical tool among medical and social science researchers. Unfortunately, since regression analysis measures the linear relationship between the means of the dependent and independent variable(s), the results can be greatly impacted by outliers. In other words, outliers can reduce the fit and provide a bias prediction. If we look at data from the *Men's 100 metres world record progression* we can see how one outlier effects the intercept, beta coefficient and R-squared. Visually, we can see that there is an outlier in *Figure 1*. When reviewing the summary statistic for *Model 1*, we see that the beta coefficients are $\beta_1 = -0.0074$ and $\beta_0 = 24.6481$. Furthermore, the model has an R^2 value is 0.836 with a p-value of less than 0.001, suggesting that the model is significant and a great fit ($F(1,22)=112.5$, $p<0.001$). In other words, this indicates that the year in which the record is recorded explains 83.6% of the variation in world record times. Viewing the summary statistics of *Model 2*, i.e. *Figure 2*, we see that the R^2 value increase to 0.878, suggesting that the outlier had an effect on model fit, but not a larger enough effect since the overall model remained significant ($F(1,21)=150.9$, $p<0.001$). The beta coefficients are $\beta_1 = -0.0069$ and $\beta_0 = 23.5512$. Furthermore, when comparing the skewness between *Model 1* (-1.444) and *Model 2* (-0.929), we see that the removal of the outlier, "normalizes" the distribution, suggesting that the outlier was skewing the data for *Model 1*. Unfortunately, both models are kurtotic as their values are great than 3. In conclusion, both models provide great fit as the majority of the data surrounds their respective regression lines suggesting that the models have small residual error and since both models are significant, we reject the null hypothesis suggesting that there is a significant linear relationship between years and world record time. Another limitation of this model is the lack of accounting for other variables. A multiple regression could look into potential factors other than the year the world record was set, for example training regiment, nutrition, and other physiological/psychological variables to explain the potential for world record time.

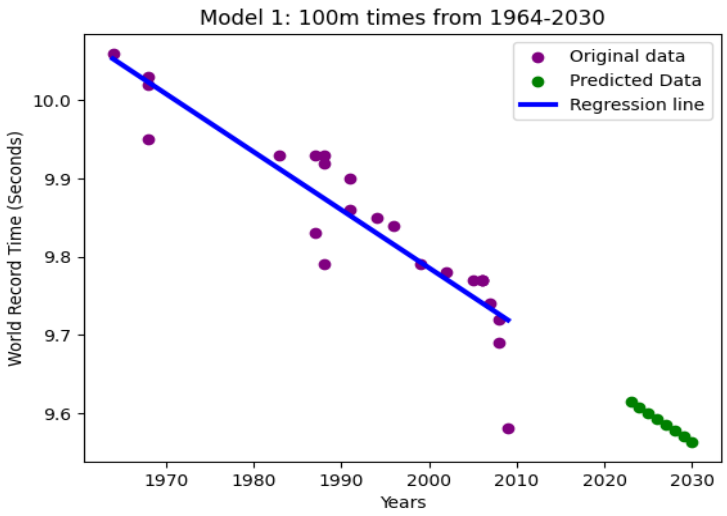

Figure 1. Original Data with Predicted Values

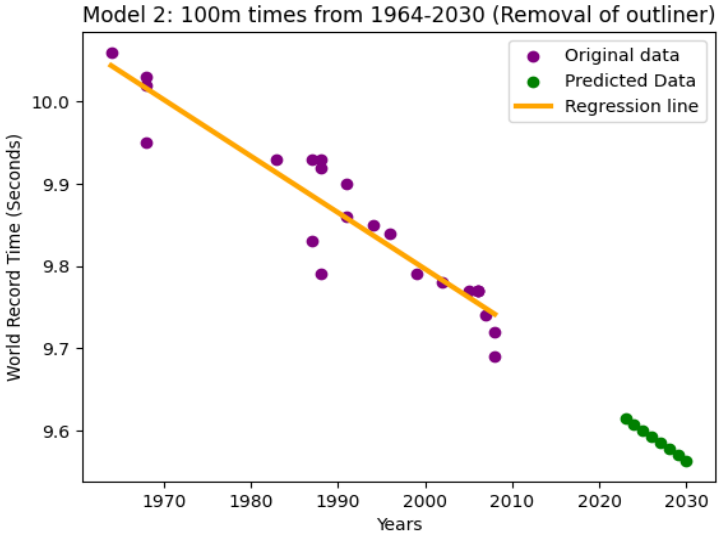

Figure 2. New Data with removal of outliers and Predicted Values