126 Data Project, Step 2

Sam Ream, Valeria Lopez, Skyler Yee

Introduction

The data was approximately what we had expected, which is shown in our calculated batting average (the number of a player's hits divided by their total number of at-bats) being around 0.246 which is close to MLB's 0.250 calculated value for the league. We sampled our data randomly to get an accurate representation of the population. We originally had hits as an independent variable but realized that we would be double counting and would make our estimators partially unidentifiable, so we removed it from our data.

Linear Model Assumptions - Runs and Doubles

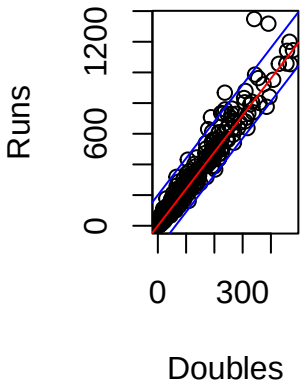

Figure 1: Scatter plot of the relationship between Runs and Doubles

It is apparent that the relationship between runs and doubles fits the assumptions of a linear model through the inspection of a graph of their relationship and apriori knowledge of the nature of the data.

- Linearity: All the points on the relationship plot above are arranged in a very linear way without transformations (a red line has been included to help demonstrate this).

- Constant Variance: Almost all of the points have a similar distance from a proposed straight line. The blue lines included in the graph above help demonstrate this fact. While 4 points in the sample do not fall within these proposed bounds, they may be outliers and only represent 0.8% of the sample.

- Independence: With the knowledge that one batter hitting the ball well enough to get to second base (a double) does not affect the likelihood of the next batter doing the same, we know that the predictors are

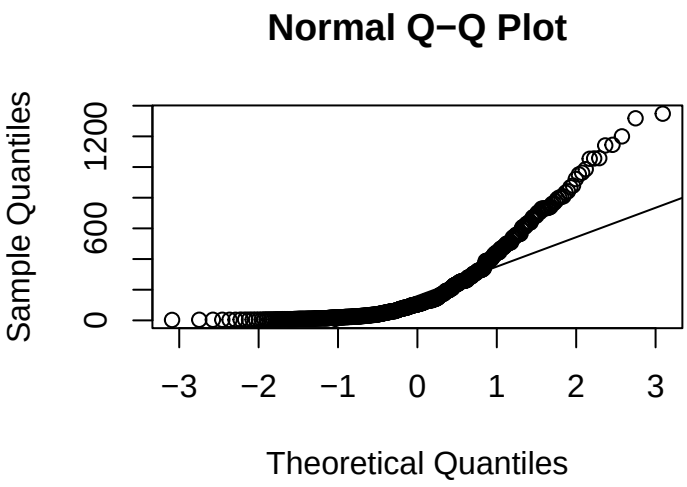

Figure 2: QQ-Plot showing that the data is not normally distributed

independent of one another. Additionally, we performed the Durbin-Watson test below which confirms the independence of our variables.

- Normality: While our errors do not appear to be normally distributed, our large sample size allows us to leverage the Central Limit Theorem to make meaningful analysis. When we tried to fit our data in such a way that the normality assumption would hold, the other three assumptions broke down.

```
##
##  Durbin-Watson test
##
## data:  batting$RUNS ~ batting$DOUBLE
## DW = 2.0821, p-value = 0.8207
## alternative hypothesis: true autocorrelation is greater than 0
```

Hypothesis Testing

Significance Test

$$H_0 : \beta_i = 0$$

$$H_a : \beta_1 \neq 0$$

$$\alpha = 0.05$$

Test Statistic $= 102.942$

P Value ≈ 0

We reject H_0 at 0.05 level. Thus, the amount of doubles a player hits is a significant predictor of how many runs the player scores.

Confidence Intervals

Confidence Interval for the Model

```
##       fit      lwr      upr
## 1 206.606 201.9707 211.2413
```

Interpretation

We are 95% confident that the mean runs per player is between 201.9707 and 211.2413.

Confidence Interval of Runs when Doubles = 200

```
##   DOUBLE
## 1    200
```

```
##        fit      lwr      upr
## 1 501.2793 397.3753 605.1833
```

Interpretation

We are 95% confident that the mean number of runs for a player with 200 doubles is between 397.3753 and 605.1833.

Fit of Model

The R^2 value of our model is 0.9551 and the adjusted R^2 is 0.9550. This means that the model explains 95.51% of the variance of the recorded events. Additionally, the residual plot in figure 3 shows how the data points share a similar spread which implies that the model is a good fit. As a result, we conclude that this model fits our data well and explains the majority of the variation in the data.

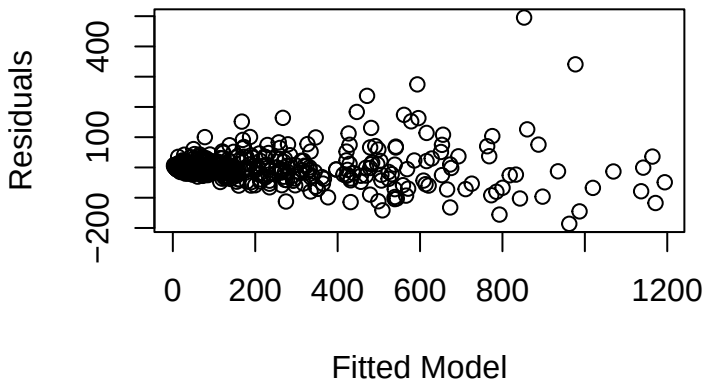

Figure 3: Residual Plot of the fitted model

Conclusion

The data is approximately what we had expected while we analyzed the relationship between our two quantitative variables, doubles and runs. Our plots were linear and the variance was constant, so we didn't have to transform the data. We used the Durbin-Watson test to test for correlation in the residuals and the value of 0 indicated independence and a positive correlation between our variables. Through hypothesis testing, we proved that doubles are a significant predictor of runs. The residual plot showed that our model was a good fit and the R^2 value of 0.9551 means that the regression explains 95.51% of the variation in our y-variable.