Data 602 - Project

Josh Brauner, Raahim Salman, Ze Yu

Feburary 14th, 2024

Part 1

Data Wrangling

In the initial phase of our project, the primary goal is to prepare the dataset for analysis. This preparation involves loading the data and then removing data points that could potentially skew or bias our results. Specifically, the dataset encompasses annual home run rates (home runs per at bat) for Barry Bonds, covering several seasons. Per the project's guidelines, it's imperative to exclude the 2001 season—the year Barry Bonds hit a record 73 home runs—from our dataset before proceeding with any further analysis. The rationale behind excluding this particular season is rooted in its outlier status within the scope of our investigation, possibly due to factors such as intentional walks, which are not reflective of the typical performance trends we aim to analyze.

```
bonds_data <- read.csv("/Users/rs/MDataSci/data601/project/bondsdata.csv")
bonds_data_filtered <- subset(bonds_data, season != 2001)
tail(bonds_data_filtered)
```

```
##    season     hrat
## 9    1995 0.065217
## 10   1996 0.081238
## 11   1997 0.075188
## 12   1998 0.067029
## 13   1999 0.095775
## 14   2000 0.102083
```

The removal of the 2001 season data point is a critical step in ensuring a more accurate and unbiased evaluation of Bonds' performance trends over the years leading up to this extraordinary season. By focusing on the years preceding 2001, we aim to construct a statistical model that predicts Bonds' home run rates without the influence of this anomalous year. This approach is intended to provide a clearer perspective on his performance trajectory, allowing us to assess whether there is evidence of an unusual improvement that could be attributed to external factors, such as the speculated use of steroids.

Model Building

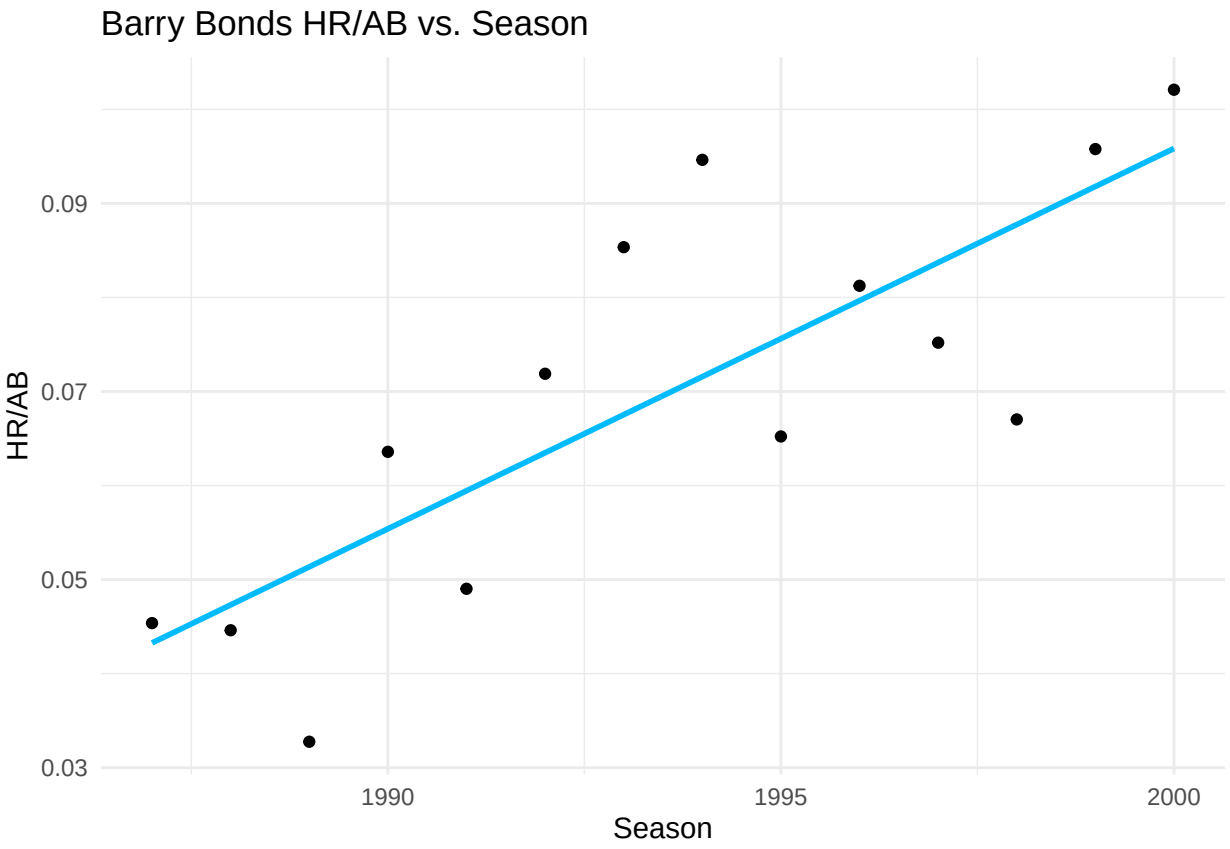
The statistical model we plan to build, $HRAT_i = A + B \times Year_i + e_i$, where $HRAT_i$ represents the home run rate in year i, and $Year_i$ denotes the year of the season, is designed to elucidate the relationship between time (years) and Bonds' home run rates. Through quantifying the trend over the specified years, this linear regression model serves as a tool for analyzing potential significant deviations in performance. Such deviations, if present, could align with steroid use, under the premise that such use would manifest as an atypical increase in home run rates.

```
model <- lm(hrat ~ season, data = bonds_data_filtered)
model$coef
```

```
##  (Intercept)       season
```

```
## -7.992499290  0.004044169
```

The linear regression equation representing the model is $\hat{HRAT}_i = -7.992 + 0.004 \times Year_i$.



Importantly, to ensure the integrity of the conclusions drawn from this model, we will verify the assumptions underlying linear regression analysis, including linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of error terms. These checks are fundamental to confirming that our model is accurately specified and that the inferences and predictions derived from it are reliable.

Model Assumption and Validation

Correlation Coefficient Check:

```
cor(bonds_data_filtered$season, bonds_data_filtered$hrat)
```

```
## [1] 0.7981544
```

A correlation coefficient of 0.7981544 indicates a strong positive linear relationship between the year and the home runs per at bat (HR/AB) for Barry Bonds, excluding the 2001 season.

Significance of Coefficient Estimates

In the context of analyzing the significance of the coefficient estimate for the year in predicting home runs per at bat (HR/AB) for Barry Bonds (excluding the 2001 season), we can formulate the null hypothesis (H_0) and the alternative hypothesis (H_1) as follows, with an alpha level (α) of 0.05:

Null hypothesis: $H_0 : \beta_1 = 0$, $H_0 : \beta_0 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$, $H_A : \beta_0 \neq 0$

```
summary(model)
```

```
##
## Call:
## lm(formula = hrat ~ season, data = bonds_data_filtered)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.020722 -0.009931  0.001841  0.007701  0.023055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.9924993  1.7566775  -4.550 0.000666 ***
## season       0.0040442  0.0008812   4.589 0.000622 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01329 on 12 degrees of freedom
## Multiple R-squared:  0.6371, Adjusted R-squared:  0.6068
## F-statistic: 21.06 on 1 and 12 DF,  p-value: 0.0006222
```

```r
coef(summary(model))[, "t value"]
```

```
## (Intercept)      season
##   -4.549782    4.589384
```

```r
coef(summary(model))[, "Pr(>|t|)"]
```

```
##  (Intercept)       season
## 0.0006664296 0.0006222474
```

Intercept

- **t-value for Intercept**: -4.549782
- **p-value for Intercept**: 0.0006664296

The intercept's t-value is significantly negative, and the corresponding p-value is much less than the alpha level of 0.05. This statistically significant result suggests that the intercept is significantly different from zero. Therefore we reject the null hypothesis $H_0 : B_0 = 0$ in favor of the alternative hypothesis $H_1 : B_0 \neq 0$\$.

Season

- **t-value for Season**: 4.589384
- **p-value for Season**: 0.0006222474

The t-value for the year coefficient is significantly positive, indicating a positive relationship between the year and HR/AB ratio for Barry Bonds in the dataset analyzed. The p-value associated with this t-value is much less than 0.05, strongly suggesting that we reject the null hypothesis $H_0 : B_1 = 0$ in favor of the alternative hypothesis $H_1 : B_1 \neq 0$.

Implications

The statistical analysis of the year's coefficient reveals that there is a significant linear relationship between the year and the HR/AB ratio. This means that the year significantly predicts the HR/AB ratio for Barry Bonds in the years leading up to 2001, excluding the 2001 season itself. The positive t-value indicates that as the year increases, so does the HR/AB ratio, suggesting an improvement in Bonds' performance in hitting home runs per at-bat attempt over time.

Given the alpha level of 0.05 and the very low p-values obtained for both the intercept and the year coefficient, our analysis provides strong statistical evidence to support the conclusion that there was a significant trend

in Barry Bonds' home run rates per at-bat over the years analyzed.

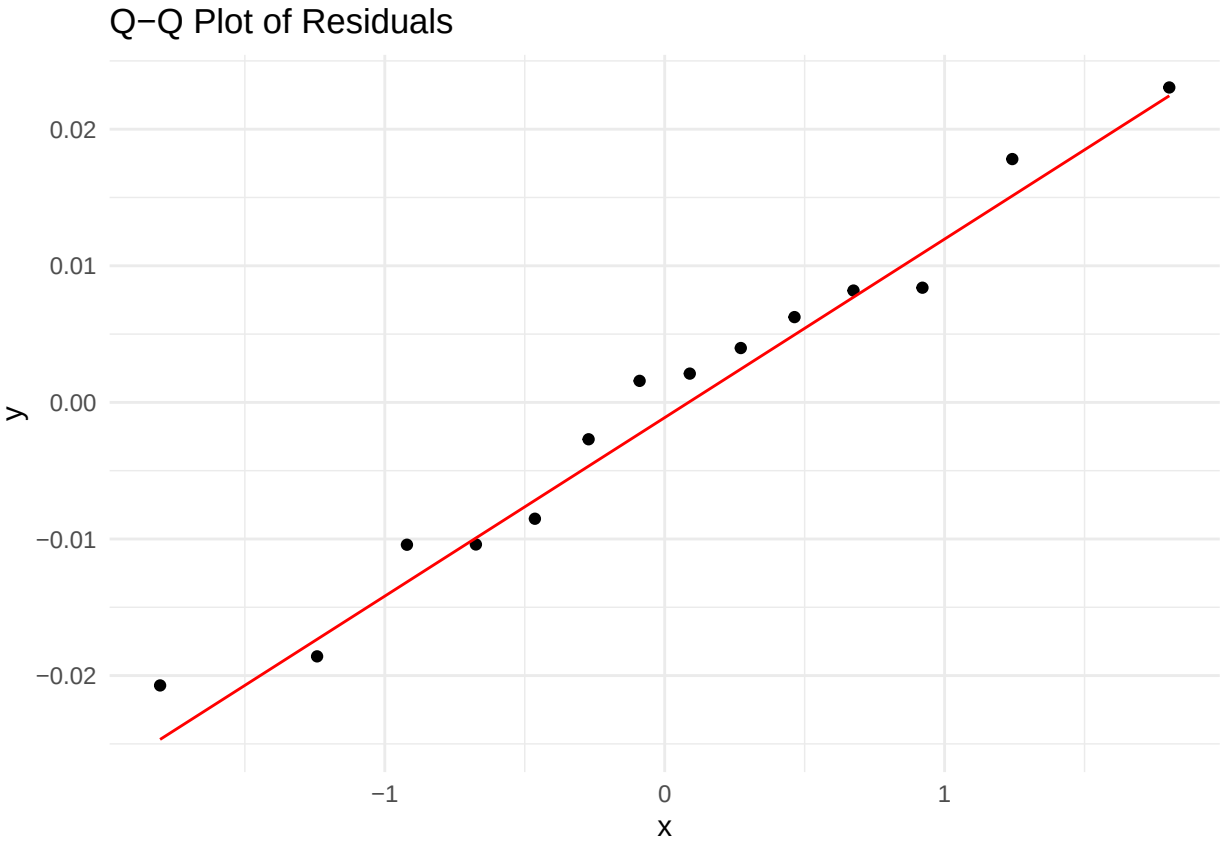
The R-squared value of 0.6371 in your linear regression model indicates that approximately 63.71% of the variance in the dependent variable is explained by the independent variable.

Residual Analysis

Normality of Residuals

In our specific analysis concerning Barry Bonds' home run rates (HR/AB) as a function of the season (year), the normality of residuals implies that the deviations from the predicted home run rates are random and follow a normal distribution. This condition supports the premise that our linear model is appropriately capturing the relationship between the year and HR/AB without systematic bias.

The evidence supporting the normality of residuals in our analysis comes from both graphical and statistical methods:



In our case, the residuals align closely with the reference line in the Q-Q plot, indicating that their distribution resembles a normal distribution. This evidence suggests that the residuals from our linear model do not deviate significantly from normality.

```
shapiro.test(residuals)
```

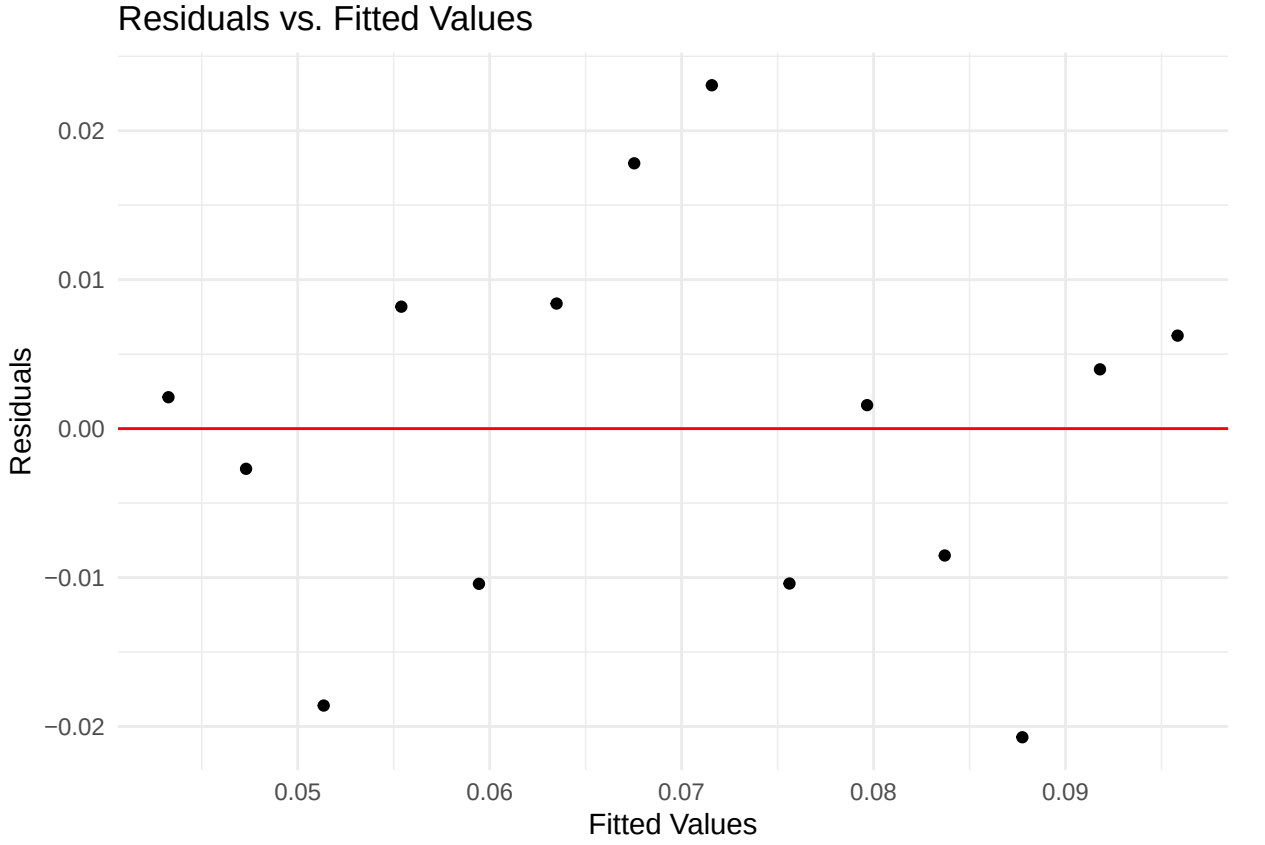
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.97132, p-value = 0.8938
```

The Shapiro-Wilk test is a statistical test designed to assess the normality of a dataset. In our analysis, the Shapiro-Wilk test for the residuals yields a p-value of 0.8938. Given that this p-value is significantly greater

than the conventional threshold of 0.05, we fail to reject the null hypothesis that the residuals are normally distributed. This statistical evidence further substantiates the claim that the residuals of our model are normal.

Homoscedasticity

The examination of homoscedasticity is a critical step in validating the assumptions underlying a linear regression model. Homoscedasticity refers to the condition where the residuals (the differences between observed and predicted values) have constant variance across all levels of the independent variable(s). This assumption ensures that the model's predictive accuracy is uniform across the range of the independent variable, which in this context is the season (year) for Barry Bonds' home run rates (HR/AB).




```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.04849, df = 1, p-value = 0.8257
```

The Breusch-Pagan (BP) test is a statistical procedure designed to test for heteroscedasticity — the presence of non-constant variance in the error terms of a regression model. In the context of analyzing Barry Bonds' home run rates (HR/AB) as a function of the season (year), the BP test's p-value of 0.8257 significantly exceeds the common alpha level threshold of 0.05. This high p-value indicates that there is insufficient evidence to reject the null hypothesis of the BP test, which states that the error variances are homoscedastic.

The even distribution of residuals across the predicted values, as evidenced by the scatter plot, allows us to conclude that the condition of homoscedasticity holds for our linear regression model. This finding is crucial because it means that our model meets another important assumption of linear regression, reinforcing its

validity and the reliability of its predictions and inferences.

Given that our analysis has confirmed both key assumptions of linear regression—normality of residuals and homoscedasticity—we can assert that our linear regression model is valid. This validity implies that the model is well-founded and that the statistical inferences drawn from it, such as confidence intervals and hypothesis tests on the regression coefficients, are based on solid assumptions.

Prediction of Home Run Rate in 2001 Season

```
new_data <- data.frame(season = 2001)
predicted_hrab <- predict(model, newdata = new_data, interval = "predict")
print(predicted_hrab)
```

```
##          fit        lwr       upr
## 1 0.09988334 0.06662845 0.1331382
```

The 95% prediction interval for Bonds' HRAT in the 2001 season, ranging from 0.06663 to 0.13314, represents the range within which we would expect Bonds' HRAT to fall, given the trends observed in the data from previous seasons. The fact that Bonds' actual HRAT of 0.153400 falls outside this interval suggests that his performance in 2001 was not only exceptional but also statistically significant in terms of deviation from predicted trends.

While the statistical analysis does not directly address whether Barry Bonds was using steroids in the 2001 season, the significant discrepancy between the predicted and actual HRATs, and the fact that the actual HRAT lies outside the 95% prediction interval, suggest that there were factors influencing Bonds' performance that year beyond what could be expected based on historical trends.

In conclusion, while our linear regression model and subsequent analysis provide valuable insights into Barry Bonds' performance trends leading up to the 2001 season, they also highlight an extraordinary deviation in his 2001 performance that warrants further investigation. The statistical evidence suggests that Bonds' HRAT in 2001 was significantly higher than expected based on past performance, pointing to the presence of additional factors that contributed to this anomaly.

It's important to note that statistical analysis alone cannot prove the use of performance-enhancing substances. Such conclusions would require corroborative evidence beyond the scope of this statistical model. However, our analysis does underscore the exceptional nature of Bonds' 2001 season within the context of his career performance trends.