Josh Warner, Tim Mellman, Jimin Lee, Harrison Keyser, Malcolm Newmark
Dr. José E. Figueroa-López
Statistics and Data Science 439, Linear Statistical Models
9 December 2024

Multivariate Statistical Analysis of MLB Home Run Distance

I. Introduction & Objectives

In this project, we attempted to create a linear model for the distance of home run balls hit during the 2024 MLB baseball season. We modeled the dependent variable *Home Run Distance* using a variety of predictors, namely *Pitch Velocity*, *Launch Speed*, *Launch Angle*, *Bat Speed*, and *Swing Length*. *Home Run Distance* is the distance from home plate that a home run ball travels. *Pitch Velocity* refers to how fast the pitch was thrown by the pitcher in miles per hour. *Launch Speed* describes how fast the ball was hit by the batter in miles per hour. *Launch Angle* depicts the angle above horizontal in degrees that a ball travels. *Bat Speed* denotes how fast the batter swung in miles per hour. *Swing Lengt*h indicates how long the batter's swing path, from the initial position of the bat to the point of contact with the ball, was in feet.

Our dataset was adopted from Baseball Savant, a website that keeps extensive records of a broad array of baseball statistics. Our data includes N=5445 observations, with each corresponding to a single home run. Due to incomplete information on 485 of the observations, we ultimately used N=4960 points in our analysis. Note that inside-the-park home runs, in which a batter scored on their own batted ball that did not go over the fence, are included.

After the creation and exploration of the basic model itself, we had numerous other objectives for this project. We first wanted to see if a linear model was appropriate for such data, and therefore took numerous diagnostics to test whether we can use a Gauss-Markov linear model to appropriately estimate home run distance from the aforementioned predictors. Then, we explored the possibility of applying a transformation to sharpen our full model. After, we looked for serial correlation, particularly relating to the time in the season in which the ball was hit. We further explored whether the variance of the errors depended linearly on a single predictor, in this case, *launch angle*. We lastly assessed the potential for multicollinearity between different predictors.

Once we analyzed the full model, we used a variety of model selection techniques to determine which predictors contributed most significantly to the total home run distance. We used exhaustive selection parameterized by R^2, R^2_{adj}, AIC, BIC, and Mallow's C_p, as well as backward and forward selection, to produce a reduced model for home run distance. We finally computed diagnostics on this reduced model.

The remainder of this report is structured as follows. Section II presents our statistical analysis, which includes IIa. Initial Data Analysis, IIb. Running the initial full model, IIc, Diagnostics of the full model, including issues of collinearity, IId. Correcting inadequacies in the models, IIe. Model Selection, and IIf. Model Diagnostics of the reduced model. Section III contains our conclusions and interpretations, including our final reduced model.

II. Statistical Analysis

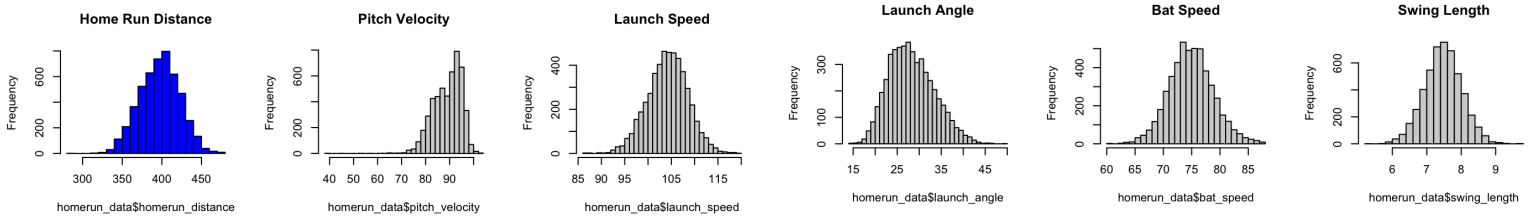
A. Initial Data Analysis

We first conducted an exploratory analysis of our dataset in R. This involved us first having to omit any observations with incomplete data, bringing our total sample size N to 4960, of the 5445 home runs hit during the 2024 season. We also excluded variables we felt bore no

resemblance to home run distance, such as the release position of the pitch in the X, Y, and Z axes. This left us with 5 predictors of interest: *Pitch Velocity*, *Launch Speed*, *Launch Angle*, *Bat Speed*, and *Swing Length*.

For our initial analysis, we constructed a histogram of each variable, including the response, home run distance. Our main goal here was to discern the normality of each variable. The results of our histograms are shown below:

These plots suggest approximately normal distributions for our response *Home Run Distance*, as well as our predictors of *Launch Speed*, *Bat Speed*, and *Swing Length*. Two possibly problematic predictors could be *Pitch Velocity* and *Launch Angle* (right-tailed), which are left-tailed and right-tailed respectively, and thus not normally distributed.



Next, we created a scatterplot matrix for each variable. Our main goal here was to determine what, if any, relationships exist between each of our predictors with each other, and the response. A strong linear relationship between predictors can be suggestive of multicollinearity, which can potentially bias our model. A strong linear relationship between one of the predictors and the response can show that the predictor may have a particularly significant effect on the overall model. This scatterplot matrix is shown to the right →

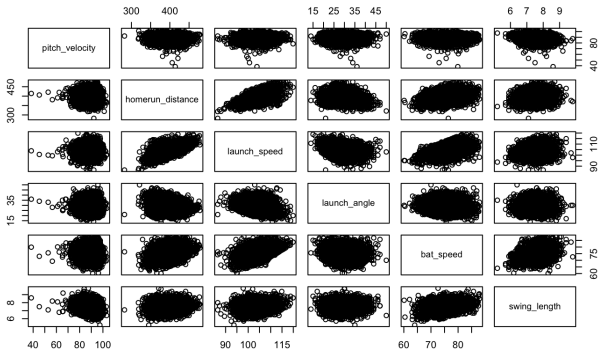
```
Call:
lm(formula = homerun_distance ~ pitch_velocity + launch_speed +
    launch_angle + bat_speed + swing_length, data = homerun_data)

Residuals:
    Min      1Q  Median      3Q     Max
-83.512 -10.714   1.551  12.890  62.915

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.40908    9.08772   2.026  0.04285 *
pitch_velocity -0.21420    0.04779  -4.483 7.54e-06 ***
launch_speed    3.93431    0.07588  51.846  < 2e-16 ***
launch_angle    0.17706    0.05541   3.195  0.00141 **
bat_speed      -0.07449    0.08282  -0.899  0.36849
swing_length   -1.49272    0.54508  -2.739  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.9 on 4954 degrees of freedom
Multiple R-squared:  0.4296,    Adjusted R-squared:  0.4291
F-statistic: 746.3 on 5 and 4954 DF,  p-value: < 2.2e-16
```

Note the scatterplots revealed a few distinguishable relationships. Most notably, there appeared to be a strong positive linear relationship between *Launch Speed* and *Home Run Distance*. From this cursory observation, we predicted that *Launch Speed* would be the most significant regressor. Additionally, we noted a moderate positive linear relationship between *Launch Speed and Bat Speed*, suggesting some correlation may exist between these two predictors.



B. Initial Full Model

To begin the process of developing a linear model for the dataset, we ran a full regression preview. We used *Pitch velocity*, *Launch Speed*, *Launch Angle*, *Bat Speed*, and *Swing Length* as the predictors, and *Home Run Distance* as the response. The R output of our initial model is shown to the right →

From this output, we can write our initial model as ***Home Run Distance* = 18.491 - 0.214*Pitch Velocity + 3.934*Launch Speed + 0.178*Launch Angle - 0.074*Bat Speed - 1.493*Swing Length**

Note that our overall regression is highly significant, with a p-value close to 0. Our R^2 goodness of fit is 0.430, which signifies that 43% of the variance in home run distance is accounted for by our model. Together, these suggest our regression provides a useful, but certainly not comprehensive explanation of home run distance.

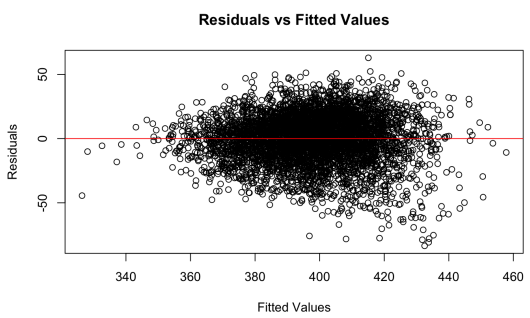
Looking at each regressor specifically, we see that *Pitch Velocity* and *Launch Speed* are highly significant, and remain so regardless of the α used for significance. *Launch Angle* and *Swing Length* are also significant for any $\alpha < 0.01$. *Bat Speed* is not significant at any reasonable α.

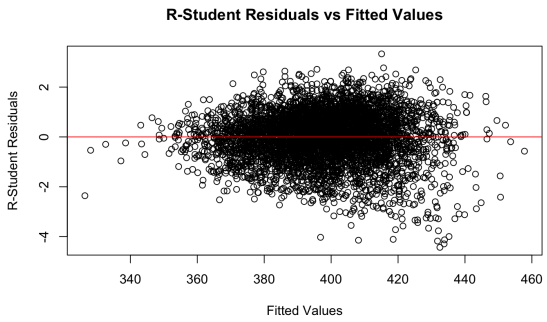
The sign of the coefficients generally makes sense- a slower pitch, a faster hit ball, and a higher launch angle all logically would lead to the ball being hit further. However, we were surprised that bat speed and swing length both had negative coefficients. We assumed a faster and longer swing would lead to further homers, but they actually were correlated with slightly shorter distances.

C. Diagnostics for Full Model

We proceeded to conduct a variety of diagnostics for our model, including checks for constant variance, normality, large leverage points, outlying points in y, influential points, the existence of serial correlation, and the possibility that the variance of our errors is linearly dependent on a specific predictor.

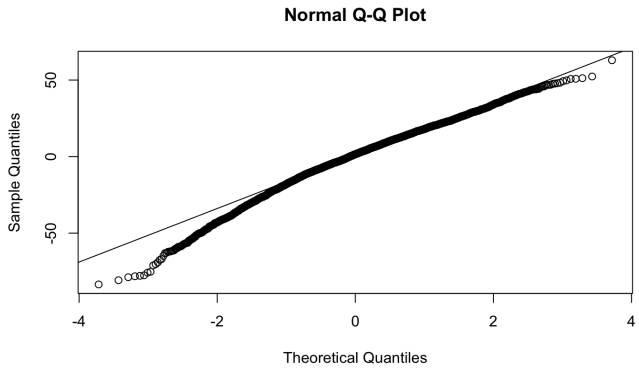
We first checked the constant variance assumption for our errors using a residuals vs fitted plot and R-student vs fitted plot. If variance were constant, we would expect each residual and R-student residual to be spread roughly evenly around 0 across the whole plot. Our plots are below.





Note that we see some outward funneling in both plots as our fitted values increase. This led us to test more formally for heteroscedasticity in the data. We conducted a Breusch-Pagan (BP) test, which returned a BP test statistic of 174.65 with 5 degrees of freedom, and a p-value near 0. We therefore concluded the errors are indeed heteroscedastic.

We next tested the normality assumption of our data via usage of a QQ plot. If the data did indeed follow a normal distribution, we would expect the points to follow a straight line. Our normal QQ plot is shown below



Note that on both ends of the QQ plot, the sample quantiles lay well below the theoretical ones. This suggests our data is likely short-tailed and more importantly does not meet the normality assumption.

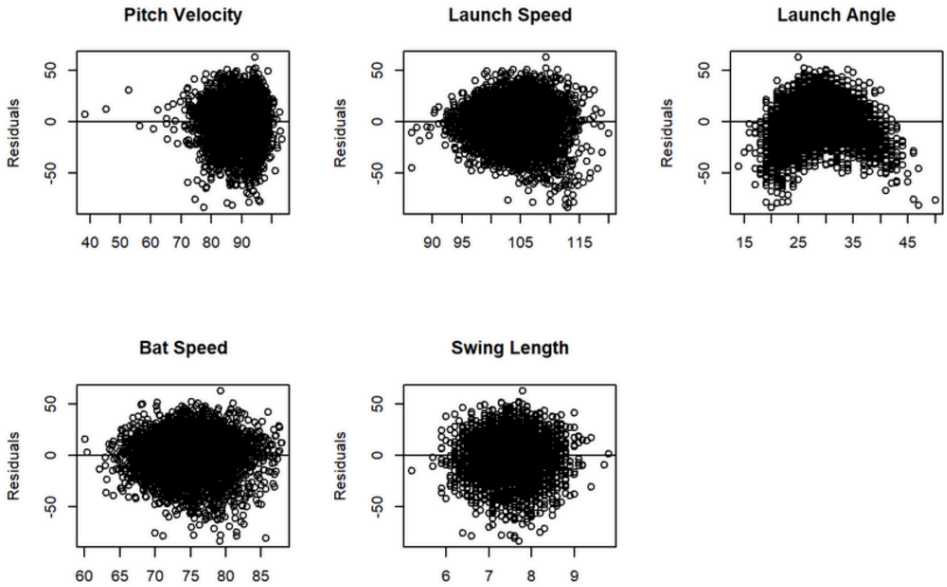
We then look for large leverage, outlying, and influential points. Large leverage occurs when the predictors take on a value well outside the bulk of the dataset. Outliers are points where values of the response deviate greatly from the rest of the dataset and their fitted value. Influential points are those which have an unusually large effect on the model. Using Bonferroni's correction, we find 264 large leverage points and 1 outlier out of our 4960 points. Additionally, we do not find any influential points.

We next looked for the possibility of serial correlation between errors, or that the errors are correlated with one another. Serial correlation can be problematic because it would lead to a violation of the independent error assumption necessary to employ the Gauss-Markov model. We attempted to sort our data by time, but due to the issue of introducing duplicates, we instead opted to sort the data by *Launch Speed*. In order to assess for serial correlation, we used a Durbin-Watson test. If there were no serial correlation, we would expect the D-W statistic to take on a value of 2. Note that this test was conducted on a reciprocal transformation of our model, for reasons discussed in the following section Our results are shown below

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5663815     0.8615496       0
##  Alternative hypothesis: rho != 0
```

This test returned a D-W statistic of 0.862, with a p-value of 0. Additionally, note that the ρ= 0.566. These both suggest positive serial correlation exists, and we will correct for it using a generalized least squares estimator.

We then examined whether the variance of the errors may be dependent on a specific predictor. To do this, we plotted the residuals of the model against each predictor. The plots are shown below.



Note that the residuals appear generally homoscedastic and unbiased with respect to most predictors, except for *Launch Speed* and *Launch Angl*e. With respect to *Launch Speed*, the residuals exhibit heteroskedasticity but unbiasedness. In contrast, with respect to *Launch Angle*, the residuals appear both heteroskedastic and biased.

Lastly, we evaluated the model for multicollinearity, which occurs when multiple predictors are correlated with one another. Multicollinearity can be a problem since a relationship between predictors can inflate standard errors, make fitted coefficients unreliable, and make the model excessively sensitive to measurement errors. In order to test for multicollinearity, we used a Condition Numbers Test. If no multicollinearity were to exist, we would expect all condition numbers to be below 30. The results of our condition numbers are below:

1.00000 28.30905 30.03318 56.83544 306.52465

Note that three of the condition numbers are above 30, and one is exceedingly large, with a value above 300. This suggests multicollinearity does exist. A second test for multicollinearity is the Variance Inflation Factor (VIF). The VIF estimates how much the regression coefficient is inflated due to collinearity. If no multicollinearity were to exist, we would expect all the VIFs to be below 10. The results of our VIFs are shown below

pitch_velocity	launch_speed	launch_angle	bat_speed	swing_length
1.083491	1.488177	1.123148	1.458378	1.203293

Note that the VIFs are all well below 10, and in fact below 1.5. This result would suggest that multicollinearity does not exist.

Given our Condition Number and VIF tests produced contrasting results, we came to the conclusion that multicollinearity likely exists to a mild degree. In terms of explaining this contrasting result, we believe perhaps non-linear relationships between the variables, which are not captured by the VIFs may exist, or that just a few of the predictors exhibit collinear relationships, which again would lead to an overall low VIF inflation.

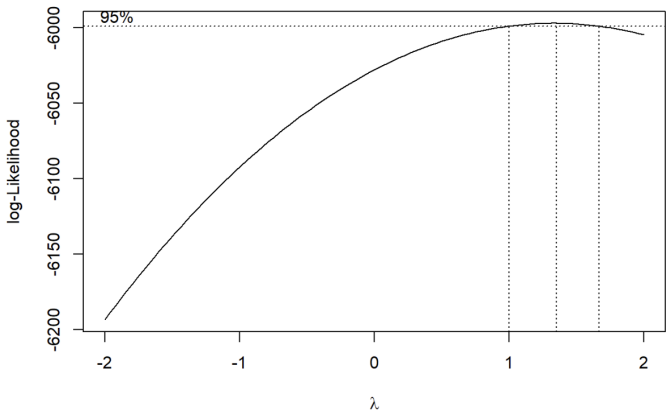
D. Correction of Model Inadequacies

In order to improve our model, we performed numerous procedures to correct potential inadequacies. To meet this end, we explored various transformations, applied a general least squares estimator, and applied a weighted least squares along the lines of a specific predictor.

We first applied a Box-Cox transformation, which entails using a power transformation on the response, in our case *Home Run Distance*, with the goal of improving homoscedasticity errors. The Box-Cox transformation takes the form:

$$y' = g_\lambda(y) = \begin{cases} y^\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases}$$

For this transformation, we received an optimal λ= 1.354, as illustrated in the graph below.



Using this transformation, our new model becomes ***Home Run Distance*$^{1.354}$= 18.491 - 0.214**Pitch Velocity* + 3.934**Launch Speed* + 0.178**Launch Angle* - 0.074**Bat Speed* - 1.493**Swing Length*.** We conducted a BP test for the Box-Cox transformed model, and this actually returned a worse BP statistic of 205.150.

The Box-Cox transformation did improve residual normally. However, given the Box-Cox transformation made the heteroscedasticity worse, we explored a few alternative transformations. We used a logarithmic, reciprocal, and square-root transformation, and conducted a BP test for each. The BPs for the logarithmic, reciprocal, and square-root transformation were 98.947, 43.428, and 134.610 respectively. This suggests the reciprocal transformation produces the most homoscedastic errors. Our best-transformed model is, therefore:

1 / *Home Run Distance*= 0.004941 + 0.000001416Pitch Velocity* - 0.00002529**Launch speed* - 0.0000009777**Launch Angle* + 0.0000007327**Bat Speed* + 0.000008857**Swing Length*.** Note that a tradeoff of using this transformation is a slight reduction in residual normality.

We next applied a General Least Squares (GLS) estimator with an AR(1) structure to reduce correct serial correlation in the errors by launch angle. The AR(1) structure accounts for the error in the previous terrible. The GLS model in R produced the output displayed in the figure below.

Using this output, we can write the formula for the GLS model as **1 / *Home Run Distance* = 0.005095 + 0.000000543**Pitch Velocity* - 0.000025252**Launch Speed* - 0.000000541**Launch Angle* + 0.000000194**Bat Speed* + 0.000001915**Swing Length***. This model produced an R^2= 0.424, which is no improvement over our untransformed reciprocal model, which had an R^2= 0.425, and is actually lower than the R^2 for our initial model, which was 0.430. However, the model should correct for serial correlation based on *Launch Speed* and therefore may still serve as an improvement over our initial full model.

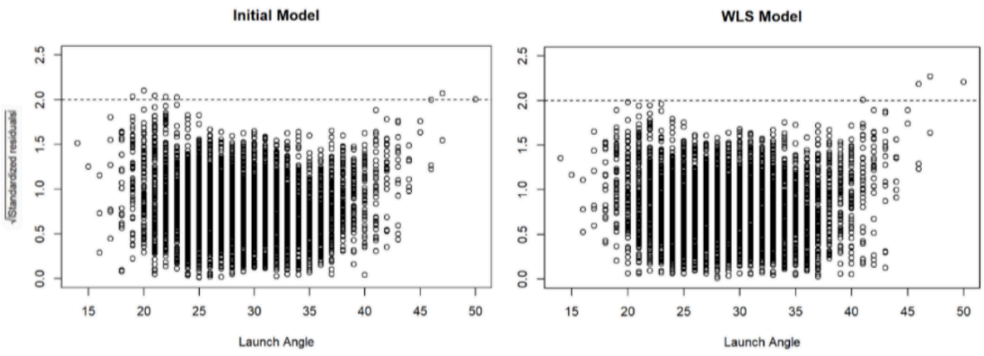
```
## Generalized least squares fit by REML
##   Model: I(1/homerun_distance) ~ pitch_velocity + launch_speed +
##   Data: homerun_data
##        AIC       BIC   logLik
##   -77780.94 -77728.87 38898.47
##
## Correlation Structure: AR(1)
##  Formula: ~row_index
##  Parameter estimate(s):
##       Phi
## 0.6453011
##
## Coefficients:
##                       Value    Std.Error   t-value p-value
## (Intercept)     0.005095050 9.388924e-05  54.26660  0.0000
## pitch_velocity  0.000000543 1.982900e-07   2.73858  0.0062
## launch_speed   -0.000025252 8.816700e-07 -28.64128  0.0000
## launch_angle   -0.000000541 2.300300e-07  -2.35091  0.0188
## bat_speed       0.000000194 3.432300e-07   0.56474  0.5723
## swing_length    0.000001915 2.285480e-06   0.83776  0.4022
##
##  Correlation:
##                (Intr) ptch_v lnch_s lnch_n bt_spd
## pitch_velocity -0.198
## launch_speed   -0.930 -0.043
## launch_angle   -0.143 -0.011  0.115
## bat_speed      -0.057  0.031 -0.169 -0.095
## swing_length   -0.135  0.246 -0.014 -0.065 -0.278
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.5899146 -0.7004914 -0.1274405  0.5423320  4.6816263
##
## Residual standard error: 0.0001224763
## Degrees of freedom: 4960 total; 4954 residual
```

Lastly, we applied the Weighted Least Squares estimate to account for the variance of errors depending on a specific predictor. Due to the bias and heteroskedasticity of the errors being the greatest for *Launch Angle*, we used *Launch Angl*e as the anchor for our weighted model. Using the procedure provided in the appendix, we derived the formula for the weights:

$$w = 1/\hat{s}_{home\ run\ distance} = 1/(e^{9.5920} x_{launch\ angle}^{-1.0138})$$

Plugging in this weight, we received the following formula for our WLS regression: ***Home Run Distance*= 30.49612 - 0.24576**Pitch Velocity* + 3.97114**Launch Speed* - 0.18916**Launch Angle* - 0.12145 **Bat Speed* - 1.32054**Swing Length.*** This model produced an R^2= 0.456, which is a slight improvement from R^2=0.430 of our initial model.

To the right is a plot comparing the studentized residuals (scaled by square root) of the initial model

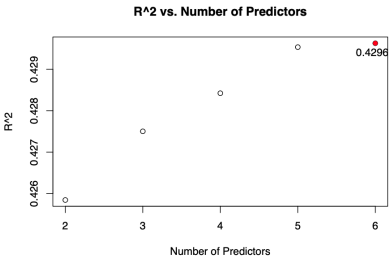


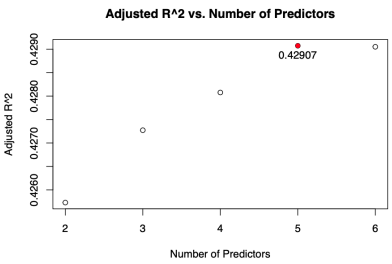
versus the weighted least squares model.

We examined the scale-location plots against the launch angle of the initial model and WLS model. After applying WLS, the small y-values for observations with large launch angles decreased and became more uniform across the x-axis. This diminished the rising pattern previously seen between x-values of 35 and 45 in the initial model plot. Overall, the WLS model plot demonstrates a more consistent representation of y-values across the x-axis, displaying reduced curvature and stronger homoscedasticity.

E. Model Selection

Our analysis so far has revealed that *Bat Speed* does not have a significant effect on *Home Run Distance*, given the summary statistics, but it's so far been unclear if a model excluding that variable (or any other variable) would in fact be a better model of *Home Run Distance* than the full model. Therefore, we used two categories of methods - exhaustive model selection search, and backward elimination and forward selection - to perform model selection by removing insignificant variables in order to balance goodness of fit with model simplicity.

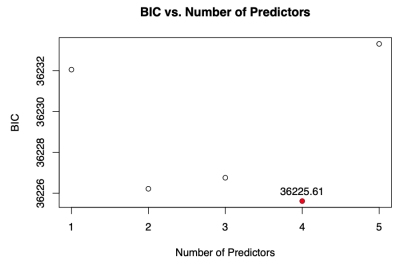


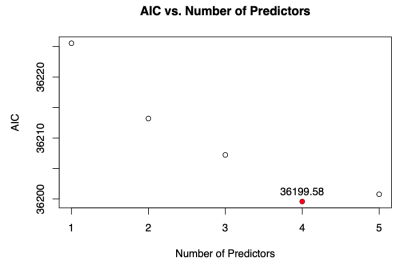


Beginning with an exhaustive model selection search, we used R to select the top model with one, two, three, and four variables in addition to the full five-predictor model and compared the selection criteria statistics visually to find the optimal model with each selection method.

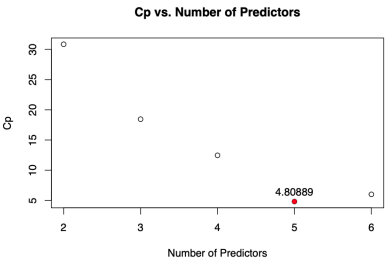
First looking at the models' R^2 values (above), the model with the highest R^2 value was the full model, which had $R^2 = 0.430$. This was expected: R^2 always increases when a new predictor is added. A more accurate comparison is the Adjusted R^2 (above), which accounts for the number of predictors. The model with the highest-performing model was the one with all predictors except *Bat Speed*, which supports the claim that *bat speed* isn't necessary for predicting *Home Run Distance*. The R^2_{adj} for our four-predictor model was 0.429_j In fact, the Adjusted R^2 for the full model is only 0.00002 higher, suggesting the two models are very similar.

AIC and BIC (right →) are both measures that penalize based on the number of predictors, where the model with the lowest value is the best model. AIC's best model is the model with all predictors except *Bat Speed*, with the order of the next-best models matching that of the R^2_{adj}. BIC, on the other hand, has the same resulting best model, but its penalty that also accounts for dataset size results in the next-best models being ordered differently. In fact, the second-best model as determined by BIC





has just two predictors, *Pitch Velocity* and *Launch Speed* — notably, the two most significant regressors from the original significance table, with a p-value of near-zero.



The final model selection technique is to take each model's Mallow's C_p value (right →), which is ideally close to zero. This finds the same ideal model as the other techniques (save R^2) and with the same ranking as Adjusted R^2 and AIC.

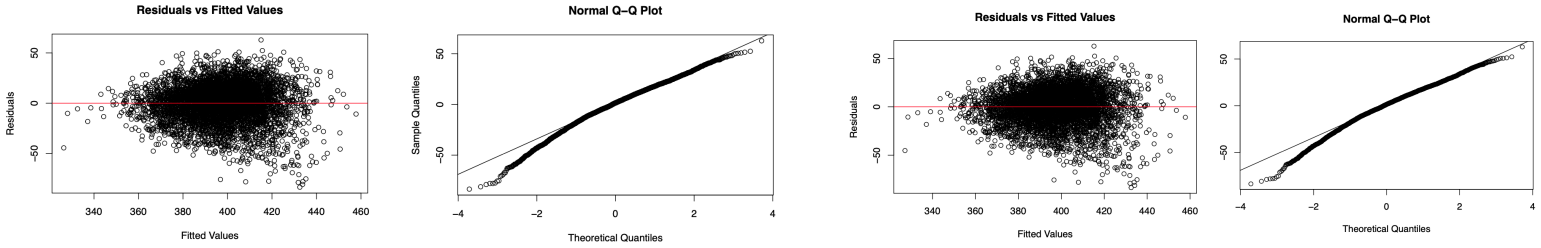
We then used backward elimination and forward selection to test if they returned the same ideal model. Backward elimination starts with the full model and removes predictors with the highest p-value above 0.05, until all predictors are significant — after removing *Bat Speed* (with a p-value of 0.3695) as a predictor, all remaining predictors had a p-value below 0.05 (the highest being *Swing Length* with 0.0019), meaning they are all significant predictors of *Home Run Distance*, which is consistent with the other selection criteria methods we used.

Likewise, forward selection results in the same findings. This process adds predictors one by one based on the highest AIC value (when automated with the step() operation as we did here). The best model is once again the one with all predictors except for *Bat Speed*, which has an AIC value of 29,163.64.

F. Diagnostics for Reduced Model

After determining an ideal reduced model - using all predictor variables except *Bat Speed* - we ran the same five model diagnostics as before and compared them with the results from the full model.

Our findings here were fairly simple — every measure was almost exactly the same for the full and reduced models. The Residuals vs Fitted Values and Normal Q-Q plots are effectively identical, with the former suggesting some evidence of heteroscedasticity and funneling and the latter being clearly concave-down with a very low p-value, suggesting that the residuals violate the normality assumption and are left skewed. All four plots are shown below, with the left panel showing the full model and the right panel showing the reduced one. The left two figures plot the full model, and the right two figures plot the reduced model.



After performing the calculations for large leverage, outlying, and influential points, both the full and reduced models have one outlying point (accounting for Bonferroni's correction) and zero

influential points. The full model has 264 large leverage points, and the reduced model has a very similar 266 large leverage points.

All in all, removing the *Bat Speed* predictor returns almost the exact same model as including it, which confirms this is the best way to reduce complexity without sacrificing goodness of fit.

III. Conclusion and Interpretations

In conclusion, we landed on a model that we believe accurately predicts home run distance as a function of relevant variables. Our final reduced model is

Home Run Distance* = 17.184 - 0.213*Pitch Velocity* + 3.902****Launch Speed* + 0.173****Launch Angle* - 1.628****Swing Length***.

In other words, our model predicts home run distance in feet as being 17.184 - 0.213 times the pitch velocity in miles per hour + 3.902 times the launch speed in miles per hour + 0.173 the launch angle - 1.628 times the swing length in feet. Note that *Bat Speed* was not significant and we removed it from the reduced model. This means that a slower pitch, faster batted-ball speed, higher launch angle, and more compact swing length all contributed to a home run ball going further, and how fast a batter swung made no impact on the distance.

Note that as we completed this project, a few minor points of issue or concern did emerge that are worthy of brief discussion. Firstly we had a lengthy debate over whether the errors in the original model could be considered homoscedastic. While some outward funneling emerged with increasing distance values, it was not clear whether this rose to the level of a pattern or simply reflected randomness in the data. We performed more formal tests which concluded the errors were indeed heteroscedasticity, but we remain cautious in this conclusion. Second, in our assessment of serial correlation with the time of the season, we came across the problem that numerous points in our data were duplicated– that they were hit in the same inning of a game on the same day. We believe this effect was relatively minor, and therefore just ordered duplicates alphabetically by the last name of the player. Third, our exploration of multicollinearity produced conflicting results. Via the usage of the Condition Number, a correlation appeared to exist between the variables. However, usage of the Variance Inflation factor (VIF) produced no such result. We, therefore, concluded mild multicollinearity may exist between some of the predictors, but are again cautious in our conclusion. Despite these minor issues, the dataset was large, thorough, and complete, which gave us confidence in our conclusions.

This model can be of use to many baseball teams in attempts to increase the distance their players hit the ball or decrease this distance for opponents. Our model suggests a shorter, more compact swing, more aggressive contact with the ball, an uppercut, and a keen eye for the offspeed pitch can all make a ball travel further. In turn, pitching more fastballs designed to induce weak and downward contact can help to decide the distance a ball is batted.