Logistic Regression Analysis on Toronto Blue Jays: Predicting Base Hits

Kristina Herard

Department of Mathematics & Statistics, University of Calgary

Introduction

- Logistic regression can be performed to predict whether a batted ball becomes a base hit by analyzing baseball performance data.

Research Question: what variables are most significant when predicting whether a batted ball becomes a base hit for the Toronto Blue Jays?

Data

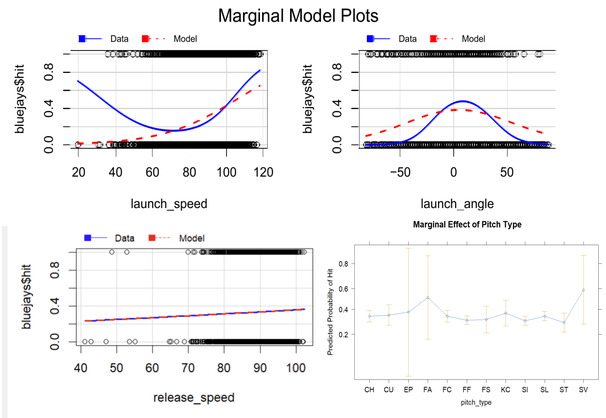
- 4,236 data points
- 118 variables
- Data: Toronto Blue Jays Statcast Data (2022 season)

Initial Model

Variables Used:

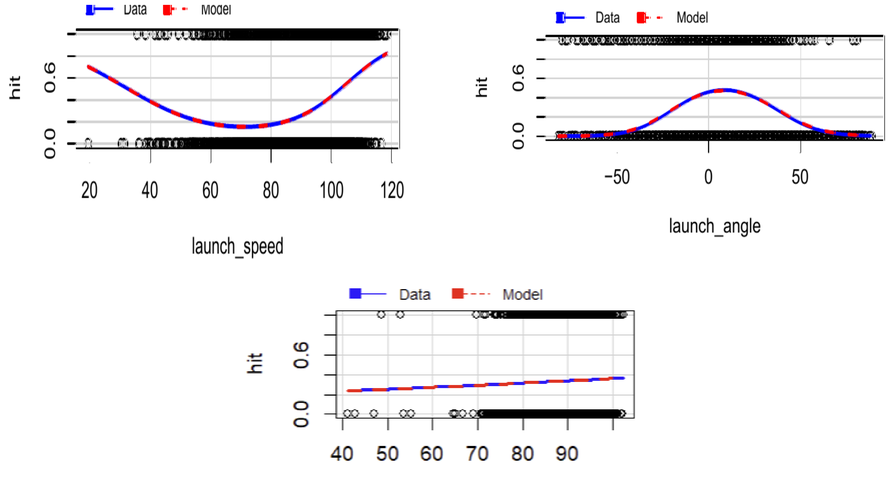
- Outcome variable:
  - Hit (0 = no base hit, 1 = base hit)
- Predictors:
  - Launch_speed (exit velocity)
  - Launch_angle
  - Release_speed
  - Pitch_type

In the initial model, the marginal model plots do not quite fit the model.
The data shows strong nonlinear relationships with U-shaped effect of launch speed, and bell-shaped effect on launch angle. This means the model is too rigid, and not flexible enough.
However, the plot for release speed shows that the data fits the model perfectly



Transformations

For model 2, a polynomial term was added to correct the plots and ensure that the model fits the data



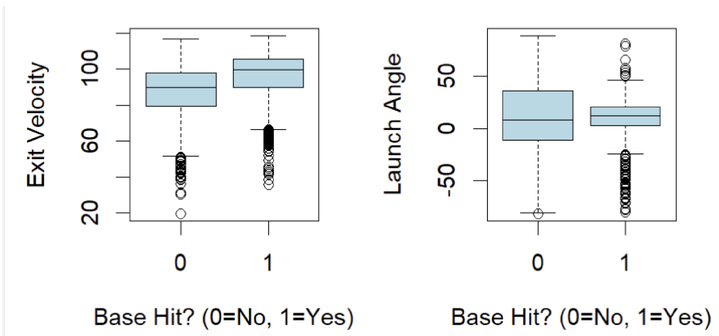
A VIF test was performed on model 2, and the model was found to have no collinearity between launch angle and launch speed, but high VIF values for the other predictors suggests multicollinearity between them.

Variable Selection

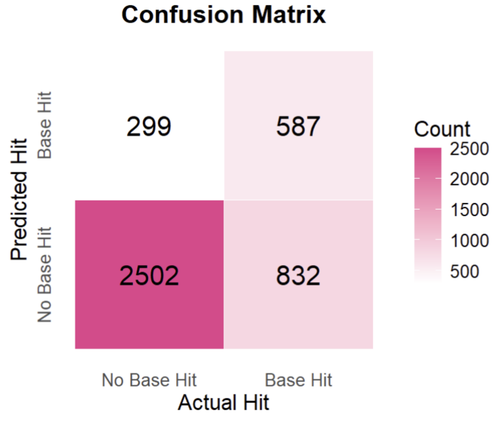
The variable selection used for this model was AIC. This variable selection method shows the lowest AIC value when pitch type and release speed are eliminated from the model, meaning the only 2 statistically significant predictors for predicting a base hit are launch speed (exit velocity) and launch angle.

This method is most applicable to this model as the model is large, and AIC holds less penalty for model complexity than BIC does.

The box plot for this model shows that exit velocity is left skewed, meaning the higher the exit velocity, the more likely a base hit will occur. Similarly, launch angle is very slightly left skewed, meaning when the launch angle is between 0 and 25°, a base hit is more likely to occur



Model Performance



Precision = 0.414: When the model predicts a hit, it is only correct 41% of the time. This means the model is not reliable in predicting actual hits.

Recall = 0.662: Of the actual hits in the dataset, this model correctly predicted 62% of them.

These values are relatively expected as a "hit" in baseball is rare (most plate appearances are outs)

An ANOVA test was run to test which model is best for inference. The p-value=0.7678, meaning that the extra predictors in the initial model are not significant enough to include in the final model, and the model will predict equally well with less model complexity.

Conclusions and Limitations

In conclusion, the key predictive factors include exit velocity and launch angle. Model 3 is the best model for inference based on the ANOVA test, variable selection, marginal model plots, and lack of collinearity.

Computing the average values of all batted balls:

- Launch speed = 90.25 mph
- Launch angle = 11.22°

Predicted probability of a hit for a batted ball with average speed and average angle:

- Approximately 0.317, or 31.7% chance of a batted ball becoming a hit.

Acknowledgements

Statcast. (n.d.-a). *Statcast Search*. baseballsavant.com. https://baseballsavant.mlb.com/statcast_search