**JEREMY PETERSON DUTOYA**

**PR. MENG-SHEN HSIEH**

**PMBA-8358-OLA: DATA-DRIVEN STRATEGIES FOR BUSINESS**

Logistic Regression Model Homework Assignment

Predicting the Baseball World Series Champion

*In this assignment, we will data on baseball teams in Major League Baseball (MLB) to try to predict the World Series winner at the beginning of the playoffs.*

*The data for this assignment is in the file Baseball.csv. Each observation corresponds to a team that made it to the playoffs each year. This dataset has 13 variables, which are defined in the following table:*

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Team | A code for the name of the team |
| League | The Major League Baseball league the team belongs to, either AL (American League), or NL (National League) |
| Year | The year of the corresponding record |
| RS | The number of runs scored by the team in that year |
| RA | The number of runs allowed by the team in that year |
| W | The number of regular-season wins by the team in that year |
| OBP | The on-base percentage of the team in that year |
| SLQ | The slugging percentage of the team in that year |
| BA | The batting average of the team in that year |
| RankSeason | The ranking of the regular season record for the teams (1 is best) in that year |
| RankPlayoffs | The ranking of the team in the playoffs in that year (the team winning the World Series gets a RankPlayoffs value of 1) |
| NumCompetitors | The number of teams in the playoffs in that year (ranges from 2 to 10) |
| WonWorldSeries | Whether or not the team won the World Series (1 if they won, and 0 otherwise) |

***Objective:*** *build a* ***logistic regression model*** *to predict the World Series winner.*

*Perform the following tasks and answer the questions:*

1. *When we are not sure which of the variables are useful for predicting the winning outcome, it is useful to build a model with only one single predicting variable and assess if that variable is statistically significant and delivers a good prediction performance.*

*Build a logistic regression model to predict* ***WonWorldSeries*** *using* ***one single*** *predicting variable from the following:* ***RS, RA, W, OBP, SLG, BA, RankSeason, NumCompetitors****, and* ***League****. Keep the entire data as the training data (uncheck* ***Partition*** *box in rattle R) and fit the model to the training data.*

**

*You should create 9 logistic regression models.*

1. *Among these 9 models,* ***for which model is the predicting variable significant?*** *In other words, which predicting variables are significant? Assume level of significance*

To determine which predicting variable is significant, we can fit 9 logistic regression models with each variable as the sole predictor variable. We will use a significance level of α = 0.10.

After fitting the 9 logistic regression models, we can examine the p-value for each predictor variable. If the p-value is less than 0.10, we can reject the null hypothesis that the predictor variable is not significant and conclude that the predictor variable is significant.

Here are the results:

* RS: p-value = 0.201 (Not significant)
* RA: p-value = 0.0262 (Significant)
* W: p-value = 0.0577 (Significant)
* OBP: p-value = 0.296 (Not significant)
* SLG: p-value = 0.0504 (Significant)
* BA: p-value = 0.839 (Not significant)
* RankSeason: p-value = 0.0438 (Significant)
* NumCompetitors: p-value = 0.000678 (Significant)
* League: p-value = 0.626 (Not significant)

From these results, we can see that the predicting variables RA, W, SLG, RankSeason, NumCompetitors are significant at α = 0.10, while the predicting variables RS, OBP, BA, and League are not significant at α = 0.10.

See below for the detail of each variable model:

*RS:* Text

Description automatically generated

RA: Text

Description automatically generated

W: Text, letter

Description automatically generated

OBP: Text

Description automatically generated with medium confidence

SLG: Text

Description automatically generated with low confidence

BA: Text

Description automatically generated with low confidence

RankSeason: A picture containing text

Description automatically generated

NumCompetitors: Text, letter

Description automatically generated

League: Text, letter

Description automatically generated

1. *Among those significant variables, which variable has a positive association with* ***the likelihood of winning the World Series****? Which variable has a negative association?*

Based on the logistic regression models obtained, we can determine which variables have a significant positive or negative association with the likelihood of winning the World Series.

Among the significant variables, we have:

* RA: This variable has a **negative association** with the likelihood of winning the World Series. This means that teams with lower values of RA (i.e., better pitching performance) are more likely to win the World Series.
* W: This variable has a **positive association** with the likelihood of winning the World Series. This means that teams with higher values of W (i.e., more wins) are more likely to win the World Series.
* SLG: This variable has a **negative association** with the likelihood of winning the World Series. This means that teams with higher values of SLG (i.e., better hitting performance) are not necessarily more likely to win the World Series. In baseball, slugging percentage (SLG) is a measure of a player's hitting performance that calculates the total number of bases a player records per at-bat. Specifically, SLG is calculated as: SLG = (Singles + 2 x Doubles + 3 x Triples + 4 x Home Runs) / At-bats. SLG gives more weight to extra-base hits (doubles, triples, and home runs) than to singles. A high SLG indicates a player with a strong hitting performance, as they are able to hit for extra bases more frequently. However, in the context of predicting the likelihood of winning the World Series, a negative association between SLG and winning suggests that having a high SLG does not guarantee winning the championship. This could be due to various reasons, such as a team's overall pitching performance, defense, or other factors that may also impact a team's success. Therefore, while having a strong hitting performance (as indicated by a high SLG) is important, it may not be the sole determinant of a team's success in winning the World Series.
* RankSeason: This variable has a **negative association** with the likelihood of winning the World Series. This means that teams with lower ranks in the season (i.e., better overall performance, 1st, 2nd for example) are more likely to win the World Series.
* NumCompetitors: This variable has a **negative association** with the likelihood of winning the World Series.

It's important to note that the interpretation of association may differ based on the specific logistic regression model and other factors such as data and contextual understanding.

1. *Now build a logistic regression model using those significant predicting variables from Q1 to predict* ***WonWorldSeries.***

***Instructions:***

1. *Split the data into training and test data, each contains 70% and 30% of the entire data respectively.*

**

1. *Fit a logistic regression model using the* ***statistically significant*** *variables from Q1 to the* ***training*** *data. Present your fitted model with the following columns:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std.Error | z value | Pr(> |Z|) |

Text

Description automatically generated

1. *Assume level of significance In this model, which predicting variables are significant? Which are insignificant? Why would some of the variables that are significant in the single predictor model in Q1 now become insignificant? Justify your answer.*

*In this logistic regression model, we used all the predicting variables that were significant when selected alone (RA, W, SLG, BA, RankSeason, and NumCompetitors) to predict the outcome variable (WonWorldSeries). The model was fit using the training data after partitioning the dataset 70/30/0, and the significance level was set at α = 0.10.*

*The results of the model show that only one predictor variable, NumCompetitors, is significant at α = 0.10. This means that the other variables (RA, W, SLG, and RankSeason) are not significant at this level.*

*The reason why some variables that were significant in the single predictor models in Q1 became insignificant in the multiple predictor model is because of multicollinearity. Multicollinearity occurs when predictor variables are highly correlated with each other, and it can make it difficult to determine the individual contribution of each variable in the model. When this happens, it becomes difficult to distinguish the effect of each predictor variable on the outcome, as the variables are no longer independent. In this case, it is possible that some of the variables are highly correlated with each other, and when they are included in the same model, their individual effects are reduced or eliminated.*

*In conclusion, only the NumCompetitors variable is significant in this logistic regression model at α = 0.10. The other variables (RA, W, SLG, and RankSeason) and the intercept are not significant. The insignificance of some variables may be due to multicollinearity, which reduces their individual contributions to the model.*

1. *Refit your model again after removing the insignificant variables from the model in (ii). Present your fitted model with the following columns:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std.Error | z value | Pr(> |Z|) |

A picture containing text

Description automatically generated

1. *Use the test data to evaluate the prediction performance of your model in (iv). Create a ROC curve and present the curve. What is the AUC (area under the ROC curve)? Do you think the model does a good job in predicting the winner of the World Series? Justify your answer.*

*Chart, line chart

Description automatically generated*

*The resulting ROC curve shows the TPR (True Positive Rate) on the y-axis and the FPR (False Positive Rate) on the x-axis. The closer the curve is to the top left corner, the better the model's performance. The AUC (Area Under the Curve) summarizes the ROC curve's overall performance, with an AUC of 1 indicating a perfect model and an AUC of 0.5 indicating a model with no predictive ability (random guessing).*

*In our case, the AUC is 0.62, which is only a little better than 0.5 and random guessing. This indicates that the logistic regression model is moderate to Poor ad so not very effective at predicting the winner of the World Series.*

1. *From the ROC curve you create, if the* ***false positive rate*** *is 0.25, what is the approximate* ***true positive rate****? In such a case, what are the corresponding values of* ***sensitivity*** *and* ***specificity****?*

Given that the false positive rate is 0.25, we can look at the ROC curve and find the point where the curve intersects with the horizontal line y=0.25. From that point, we can move vertically until we hit the diagonal line, which represents the random guessing line. The point where the curve intersects with the diagonal line represents an approximate true positive rate (TPR) for a false positive rate (FPR) of 0.25. The point of intersection is at approximately (0.25, 0.4) on the ROC curve, the approximate TPR is 0.4.

* Sensitivity = True Positive Rate = TPR = 0.4
* Specificity = 1 - False Positive Rate =1-0.25 = 0.75