PREDICTING THE BASEBALL WORLD SERIES CHAMPION

Last week, in the Moneyball lecture, we discussed how regular season performance is not strongly correlated with winning the World Series in baseball. In this homework question, we'll use the same data to investigate how well we can predict the World Series winner at the beginning of the playoffs.

To begin, load the dataset baseball.csv (/c4x/MITx/15.071x/asset/baseball.csv) into R using the read.csv function, and call the data frame "baseball". This is the same data file we used during the Moneyball lecture, and the data comes from Baseball-Reference.com (http://www.baseball-reference.com/).

As a reminder, this dataset contains data concerning a baseball team's performance in a given year. It has the following variables:

- **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
- **SLG**: The slugging percentage of the team in that year
- **BA**: The batting average of the team in that year
- **Playoffs**: Whether the team made the playoffs in that year (1 for yes, 0 for no)
- **RankSeason**: Among the playoff teams in that year, the ranking of their regular season records (1 is best)
- **RankPlayoffs**: Among the playoff teams in that year, how well they fared in the playoffs. The team winning the World Series gets a RankPlayoffs of 1.
- **G**: The number of games a team played in that year
- **OOBP**: The team's opponents' on-base percentage in that year
- **OSLG**: The team's opponents' slugging percentage in that year

PROBLEM 1.1 - LIMITING TO TEAMS MAKING THE PLAYOFFS (1/1 point)

Each row in the baseball dataset represents a team in a particular year.

How many team/year pairs are there in the whole dataset?

1232

\[1232\]

Answer: 1232

EXPLANATION

You can read the dataset into R by using the following command:

baseball = read.csv("baseball.csv")

Then nrow(baseball) or str(baseball) both show that there are 1232 team/year pairs.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.2 - LIMITING TO TEAMS MAKING THE PLAYOFFS (1/1 point)

Though the dataset contains data from 1962 until 2012, we removed several years with shorter-than-usual seasons. Using the table() function, identify the total number of years included in this dataset.

47

\[47\]

Answer: 47

EXPLANATION

table(baseball$Year) contains 47 years (1972, 1981, 1994, and 1995 are missing). You can count the number of years in the table, or the command length(table(baseball$Year)) directly provides the answer.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.3 - LIMITING TO TEAMS MAKING THE PLAYOFFS (1/1 point)

Because we're only analyzing teams that made the playoffs, use the subset() function to **replace baseball** with a data frame limited to teams that made the playoffs (so your subsetted data frame should still be called "baseball"). How many team/year pairs are included in the new dataset?

244

\[244\]

Answer: 244

EXPLANATION

baseball = subset(baseball, Playoffs == 1) limits the dataset, and the nrow() or str() functions can be used to identify that 244 team/year pairs remain.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.4 - LIMITING TO TEAMS MAKING THE PLAYOFFS (1/1 point)

Through the years, different numbers of teams have been invited to the playoffs. Which of the following has been the number of teams making the playoffs in some season?

- [x] 2
- [x] 4
- [ ] 6
- [x] 8
- [x] 10
- [ ] 12

EXPLANATION

Using table(baseball$Year), we can see at least one season had 2, 4, 8, and 10 contenders. A fancier approach would be to use table(table(baseball$Year)).

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 2.1 - ADDING AN IMPORTANT PREDICTOR (1/1 point)

It's much harder to win the World Series if there are 10 teams competing for the championship versus just two. Therefore, we will add the predictor variable NumCompetitors to the baseball data frame. NumCompetitors will contain the number of total teams making the playoffs in the year of a particular team/year pair. For instance, NumCompetitors should be 2 for the 1962 New York Yankees, but it should be 8 for the 1998 Boston Red Sox.

We start by storing the output of the table() function that counts the number of playoff teams from each year:

PlayoffTable = table(baseball$Year)

You can output the table with the following command:

PlayoffTable

We will use this stored table to look up the number of teams in the playoffs in the year of each team/year pair.

Just as we can use the names() function to get the names of a data frame's columns, we can use it to get the names of the entries in a table. What best describes the output of names(PlayoffTable)?

- Vector of years stored as numbers (type num)
- Vector of years stored as strings (type chr) ✔
- Vector of frequencies stored as numbers (type num)
- Vector of frequencies stored as strings (type chr)

EXPLANATION

From the call str(names(PlayoffTable)) we see PlayoffTable has names of type chr, which are the years of the teams in the dataset.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.2 - ADDING AN IMPORTANT PREDICTOR (1/1 point)

Given a vector of names, the table will return a vector of frequencies. Which function call returns the number of playoff teams in 1990 and 2001? (HINT: If you are not sure how these commands work, go ahead and try them out in your R console!)

- PlayoffTable(1990, 2001)
- PlayoffTable(c(1990, 2001))
- PlayoffTable("1990", "2001")
- PlayoffTable(c("1990", "2001"))
- PlayoffTable[1990, 2001]
- PlayoffTable[c(1990, 2001)]
- PlayoffTable["1990", "2001"]
- PlayoffTable[c("1990", "2001")] ✔

EXPLANATION

Because PlayoffTable is an object and not a function, we look up elements in it with square brackets instead of parentheses. We build the vector of years to be passed with the c() function. Because the names of PlayoffTable are strings and not numbers, we need to pass "1990" and "2001".

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.3 - ADDING AN IMPORTANT PREDICTOR (1/1 point)

Putting it all together, we want to look up the number of teams in the playoffs for each team/year pair in the dataset, and store it as a new variable named NumCompetitors in the baseball data frame. While of the following function calls accomplishes this? (HINT: Test out the functions if you are not sure what they do.)

- baseball$NumCompetitors = PlayoffTable(baseball$Year)
- baseball$NumCompetitors = PlayoffTable[baseball$Year]
- baseball$NumCompetitors = PlayoffTable(as.character(baseball$Year))
- baseball$NumCompetitors = PlayoffTable[as.character(baseball$Year)] ✔

EXPLANATION

Because PlayoffTable is an object and not a function, we look up elements in it with square brackets instead of parentheses. as.character() is needed to convert the Year variable in the dataset to a string, which we know from the previous parts is needed to look up elements in a table. If you're not sure what a function does, remember you can look it up with the ? function. For instance, you could type ?as.character to look up information about as.character.

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.4 - ADDING AN IMPORTANT PREDICTOR (1/1 point)

Add the NumCompetitors variable to your baseball data frame. How many playoff team/year pairs are there in our dataset from years where 8 teams were invited to the playoffs?

128

\[128\]

Answer: 128

EXPLANATION

You can add the NumCompetitors variable to the baseball data frame with the following command:

baseball$NumCompetitors = PlayoffTable[as.character(baseball$Year)]

Then you can obtain the number of team/year pairs with 8 teams in the playoffs by running table(baseball$NumCompetitors)

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.1 - BIVARIATE MODELS FOR PREDICTING WORLD SERIES WINNER (1/1 point)

In this problem, we seek to predict whether a team won the World Series; in our dataset this is denoted with a RankPlayoffs value of 1. Add a variable named WorldSeries to the baseball data frame, by typing the following command in your R console:

baseball$WorldSeries = as.numeric(baseball$RankPlayoffs == 1)

WorldSeries takes value 1 if a team won the World Series in the indicated year and a 0 otherwise. How many observations do we have in our dataset where a team did NOT win the World Series?

197

\[197\]

Answer: 197

EXPLANATION

You can create the WorldSeries variable by running the command:

baseball$WorldSeries = as.numeric(baseball$RankPlayoffs == 1)

Then, if you create the table:

table(baseball$WorldSeries)

You can see that there are 197 teams that did not win the World Series.

Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.2 - BIVARIATE MODELS FOR PREDICTING WORLD SERIES WINNER (1 point possible)

When we're not sure which of our variables are useful in predicting a particular outcome, it's often helpful to build bivariate models, which are models that predict the outcome using a single independent variable. Which of the following variables is a significant predictor of the WorldSeries variable in a bivariate logistic regression model? To determine significance, remember to look at the stars in the summary output of the model. We'll define an independent variable as significant if there is at least one star at the end of the coefficients row for that variable (this is equivalent to the probability column having a value smaller than 0.05). Note that you have to build 12 models to answer this question! Use the entire dataset baseball to build the models.

- [x] Year ✔
- [x] RS
- [x] RA ✔
- [ ] W
- [ ] OBP
- [x] SLG
- [ ] BA
- [x] RankSeason ✔
- [ ] OOBP
- [ ] OSLG
- [x] NumCompetitors ✔
- [x] League

EXPLANATION

The results come from building each bivariate model and looking at its summary. For instance, the result for the variable Year can be obtained by running summary(glm(WorldSeries~Year, data=baseball, family="binomial")). You can save time on repeated model building by using the up arrow in your R terminal. The W and SLG variables were both nearly significant, with p = 0.0577 and 0.0504, respectively.

Hide Answer *You have used 3 of 3 submissions*

PROBLEM 4.1 - MULTIVARIATE MODELS FOR PREDICTING WORLD SERIES WINNER (1/1 point)

In this section, we'll consider multivariate models that combine the variables we found to be significant in bivariate models. Build a model using all of the variables that you found to be significant in the bivariate models. How many variables are significant in the combined model?

0

\[0\]

Answer: 0

EXPLANATION

You can create a model with all of the significant variables from the bivariate models (Year, RA, RankSeason, and NumCompetitors) by using the following command:

LogModel = glm(WorldSeries ~ Year + RA + RankSeason + NumCompetitors, data=baseball, family=binomial)

Looking at summary(LogModel), you can see that none of the variables are significant in the multivariate model!

Help

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.2 - MULTIVARIATE MODELS FOR PREDICTING WORLD SERIES WINNER (1/1 point)

Often, variables that were significant in bivariate models are no longer significant in multivariate analysis due to correlation between the variables. Which of the following variable pairs have a high degree of correlation (a correlation greater than 0.8 or less than -0.8)?

- [ ] Year/RA
- [ ] Year/RankSeason
- [x] Year/NumCompetitors ✔
- [ ] RA/RankSeason
- [ ] RA/NumCompetitors
- [ ] RankSeason/NumCompetitors

EXPLANATION

To test the correlation between two variables, use a command like cor(baseball$Year, baseball$RA). While every pair was at least moderately correlated, the only strongly correlated pair was Year/NumCompetitors, with correlation coefficient 0.914.

As a shortcut, you can compute all pair-wise correlations between these variables with:

cor(baseball[c("Year", "RA", "RankSeason", "NumCompetitors")])

Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.3 - MULTIVARIATE MODELS FOR PREDICTING WORLD SERIES WINNER (1/1 point)

Build all six of the two variable models listed in the previous problem. Together with the four bivariate models, you should have 10 different logistic regression models. Which model has the bes value (the minimum AIC value)?

- ( ) Year
- ( ) RA =
- ( ) RankSeason
- (•) NumCompetitors ✔
- ( ) Year/RA
- ( ) Year/RankSeason
- ( ) Year/NumCompetitors
- ( ) RA/RankSeason
- ( ) RA/NumCompetitors
- ( ) RankSeason/NumCompetitors

EXPLANATION

The two-variable models can be built with the following commands:

Model1 = glm(WorldSeries ~ Year + RA, data=baseball, family=binomial)

Model2 = glm(WorldSeries ~ Year + RankSeason, data=baseball, family=binomial)

Model3 = glm(WorldSeries ~ Year + NumCompetitors, data=baseball, family=binomial)

Model4 = glm(WorldSeries ~ RA + RankSeason, data=baseball, family=binomial)

Model5 = glm(WorldSeries ~ RA + NumCompetitors, data=baseball, family=binomial)

Model6 = glm(WorldSeries ~ RankSeason + NumCompetitors, data=baseball, family=binomial)

None of the models with two independent variables had both variables significant, so none seem promising as compared to a simple bivariate model. Indeed the model with the lowest AIC value is the model with just NumCompetitors as the independent variable.

This seems to confirm the claim made by Billy Beane in Moneyball that all that matters in the Playoffs is luck, since NumCompetitors has nothing to do with the quality of the teams!

Hide Answer *You have used 1 of 2 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

New Post