Dimensionless Technologies Private Limited
Visit us at: www.dimensionless.in
- info@dimensionless.in
- 9923170071, 8108094992

PREDICTING THE BASEBALL WORLD SERIES CHAMPION

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 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.

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

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

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

Problem 1.2

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.

Problem 1.3

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?

Problem 1.4

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? Select all that apply.

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

Problem 2.1 - Adding an Important Predictor

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)

Problem 2.2 - Adding an Important Predictor

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")]

Problem 2.3 - Adding an Important Predictor

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. Which 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)]

Problem 2.4 - Adding an Important Predictor

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?

Problem 3.1 - Bivariate Models for Predicting World Series Winner

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:

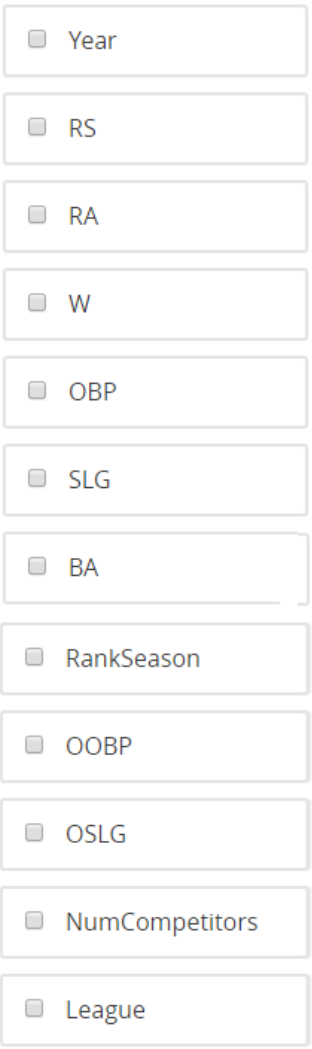
```
baseball$WorldSeries = as.numeric(baseball$RankPlayoffs == 1)
```

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?

Problem 3.2 - Bivariate Models for Predicting World Series Winner

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. (Select all that apply.)



Problem 4.1 - Multivariate Models for Predicting World Series Winner

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?

Problem 4.2 - Multivariate Models for Predicting World Series Winner

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)? Select all that apply.

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

Dimensionless Technologies Private Limited
Visit us at: www.dimensionless.in
- info@dimensionless.in
- 9923170071, 8108094992

DIMENSIONLESS TECHNOLOGY

Problem 4.3 - Multivariate Models for Predicting World Series Winner

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 best AIC value (the minimum AIC value)?

- Year
- RA
- RankSeason
- NumCompetitors
- Year/RA
- Year/RankSeason
- Year/NumCompetitors
- RA/RankSeason
- RA/NumCompetitors
- RankSeason/NumCompetitors