

Homework 1

Group 1





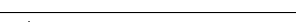
Contents

1	DATA EXPLORATION	1
2	DATA PREPARATION	3
3	BUILD MODELS	4
4	Select Models	5
5	Appendix A	5
5.1	Data Dictionary	5
5.2	R code used in document	5

1 DATA EXPLORATION

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

Descriptive Statistics 16 Variables 2276 Observations

TARGET_WINS												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
2276	0	108	1	80.79	54.0	61.0	71.0	82.0	92.0	99.5	104.0	
lowest : 0 12 14 17 21, highest: 128 129 134 135 146												
TEAM_BATTING_H												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
2276	0	569	1	1469	1282	1315	1383	1454	1537	1636	1695	
lowest : 891 992 1009 1116 1122, highest: 2333 2343 2372 2496 2554												
TEAM_BATTING_2B												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
2276	0	240	1	241.2	167	182	208	238	273	303	320	
lowest : 69 112 113 118 123, highest: 382 392 393 403 458												
TEAM_BATTING_3B												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
2276	0	144	1	55.25	23	27	34	47	72	96	108	
lowest : 0 8 9 11 12, highest: 166 190 197 200 223												
TEAM_BATTING_HR												
n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95	
2276	0	243	1	99.61	14.0	20.0	42.0	102.0	147.0	179.5	199.0	
lowest : 0 3 4 5 6, highest: 247 249 257 260 264												

TEAM_BATTING_BB

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2276	0	533	1	501.6	248.2	363.5	451.0	512.0	580.0	635.0	670.2

lowest : 0 12 29 34 45, highest: 815 819 824 860 878

TEAM_BATTING_SO

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2174	102	822	1	735.6	359	421	548	750	930	1049	1103

lowest : 0 66 67 72 74, highest: 1303 1320 1326 1335 1399

TEAM_BASERUN_SB

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2145	131	348	1	124.8	35.0	44.0	66.0	101.0	156.0	231.0	301.8

lowest : 0 14 18 19 20, highest: 562 567 632 654 697

TEAM_BASERUN_CS

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1504	772	128	1	52.8	24	30	38	49	62	77	91

lowest : 0 7 11 12 14, highest: 171 186 193 200 201

TEAM_BATTING_HBP

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
191	2085	55	1	59.36	40.0	44.0	50.5	58.0	67.0	76.0	82.5

lowest : 29 30 35 38 39, highest: 87 88 89 90 95

TEAM_PITCHING_H

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2276	0	843	1	1779	1316	1356	1419	1518	1682	2058	2563

lowest : 1137 1168 1184 1187 1202
highest: 16038 16871 20088 24057 30132

TEAM_PITCHING_HR

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2276	0	256	1	105.7	18.0	25.0	50.0	107.0	150.0	187.0	209.2

lowest : 0 3 4 5 6, highest: 291 297 301 320 343

TEAM_PITCHING_BB

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2276	0	535	1	553	377.0	417.5	476.0	536.5	611.0	693.5	757.0

lowest : 0 119 124 131 140, highest: 2169 2396 2840 2876 3645

TEAM_PITCHING_SO

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2174	102	823	1	817.7	421.3	490.0	615.0	813.5	968.0	1095.0	1173.0

lowest : 0 181 205 208 252
highest: 3450 4224 5456 12758 19278

TEAM_FIELDING_E

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
2276	0	549	1	246.5	100.0	109.0	127.0	159.0	249.2	542.0	716.0

lowest : 65 66 68 72 74, highest: 1567 1728 1740 1890 1898

TEAM_FIELDING_DP

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1990	286	144	1	146.4	98	109	131	149	164	178	186

lowest : 52 64 68 71 72, highest: 215 218 219 225 228

- Mean / Standard Deviation / Median
- Bar Chart or Box Plot of the data
- Is the data correlated to the target variable (or to other variables?)
- Are any of the variables missing and need to be imputed "fixed"?

2 DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

3 BUILD MODELS

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

4 Select Models

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. For the multiple linear regression model, will you use a metric such as Adjusted R², RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R², (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

5 Appendix A

5.1 Data Dictionary

VARIABLE.NAME..	DEFINITION	THEORETICAL.EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	NA
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

5.2 R code used in document