# world money_ball - kaggle dataset

*patrick mcdevitt (& collaboration with preeti swaminathan)*

*04 juin 2017*

## money_ball

example of ordinary logistic regression model building techniques

### 1. DATA EXPLORATION (40 points)

This data set contains approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

| Name | _ext | Number | Mean | St Dev | Min | Max |
|---|---|---|---|---|---|---|
| index | | 2,246 | 1,265,817 | 734,970 | 1 | 2,534 |
| target_wins | | 2,246 | 81,178 | 14,596 | 39 | 123 |
| team_batting | _h | 2,238 | 1,465,629 | 128,777 | 992 | 2,192 |
| team_batting | _2b | 2,246 | 241,694 | 46,136 | 113 | 458 |
| team_batting | _3b | 2,243 | 54,875 | 27,273 | 0 | 190 |
| team_batting | _hr | 2,246 | 100,715 | 60,156 | 0 | 264 |
| team_batting | _bb | 2,246 | 505,620 | 116,931 | 45 | 878 |
| team_batting | _so | 2,145 | 741,267 | 242,532 | 0 | 1,399 |
| team_baserun | _sb | 2,126 | 123,686 | 84,771 | 18 | 562 |
| team_baserun | _cs | 1,490 | 52,299 | 20,955 | 11 | 166 |
| team_batting | _hbp | 191 | 59,356 | 12,967 | 29 | 95 |
| team_pitching | _h | 2,234 | 1,650,051 | 548,521 | 1,137 | 7,093 |
| team_pitching | _hr | 2,246 | 106,649 | 61,037 | 0 | 343 |
| team_pitching | _bb | 2,238 | 547,450 | 114,549 | 119 | 1,123 |
| team_pitching | _so | 2,141 | 801,531 | 252,156 | 0 | 3,450 |
| team_fielding | _e | 2,240 | 232,420 | 192,012 | 65 | 1,374 |
| team_fielding | _dp | 1,983 | 146,538 | 26,123 | 52 | 228 |
| team_batting | _1b | 2,240 | 1,069,086 | 116,599 | 709 | 1,706 |
| team_batting | _ab | 2,238 | 5,788,027 | 130,347 | 5,317 | 6,517 |
| team_batting | _slg | 2,246 | 0,366 | 0,038 | 0,244 | 0,481 |

| Name | _ext | Number | Mean | St Dev | Min | Max |
|---|---|---|---|---|---|---|
| team_batting | _obp | 2,245 | 0,311 | 0,019 | 0,235 | 0,392 |

Describe the size and the variables in the MONEYBALL data set so that a manager can understand it. 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.

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

## 2. DATA PREPARATION (40 Points)

Each variable was assessed for several relevant characteristics to decide potential improvements for incorporation (or not) into the regression model.

- what is the percentage of missing values in that column ?
- is there a strong correlation between each column and the other potential independent variables ?
- what is the single parameter correlation to the dependent variable (target wins) ?
- what is the skewness for each variable ?
- is there a generally linear relationship to the target_wins ?
- based on a linear single parameter regression fit, are the residuals from this fit uniformly distributed ?

All of the above questions resulted in making the following decisions / actions on the data set :

---

1. **team_batting_hbp**, **team_baserun_cs** - dropped due to > 90% and 30% of missing values . . . there is little likelihood to make an informed and useful imputation correction for those data columns

---

2. **team_pitching_hr** - dropped due to 97% correlaction with team_batting_hr. There is no obvious explanation for this. Perhaps there is a mistake in the data tabulation that mixed pitching home runs allowed with batting home runs achieved or perhaps there is a tim. There is no

obvious reason to expect that a team's offense that hits a lot (or a few) home runs that their defense (pitching staff) also allows (or restricts) home runs to approximately the same level. In any case, for two independent characteristics that are 97% correlated, it is not value added to include both characteristics in the regression model.

---

3. **team_batting_h**, **team_batting_1b** (hits) - decomposed to create new predictor "team_batting_1b" (singles), with the idea that doubles (_2b), triples (_3b), and home runs (_hr) are all a part of "hits", it is perhaps useful to separate out the potential influence of singles (_1b) from the other hits, and drop hits (_h) from the independent predictors data set. There is, at the least, no loss of information from this choice.

---

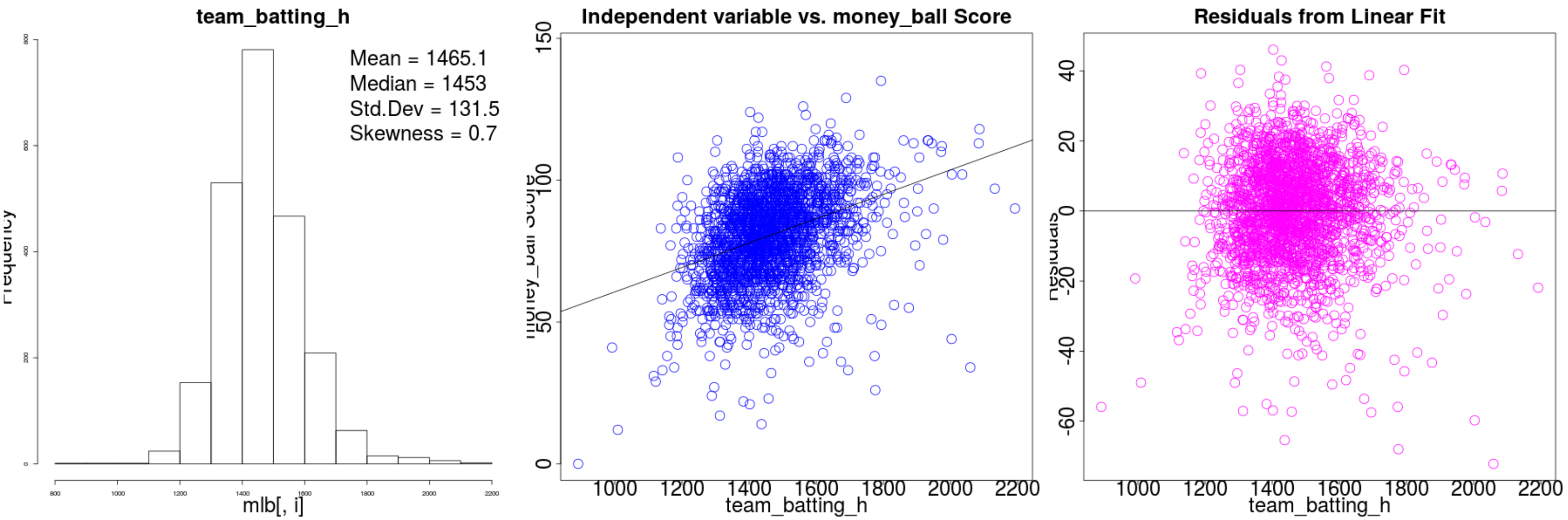
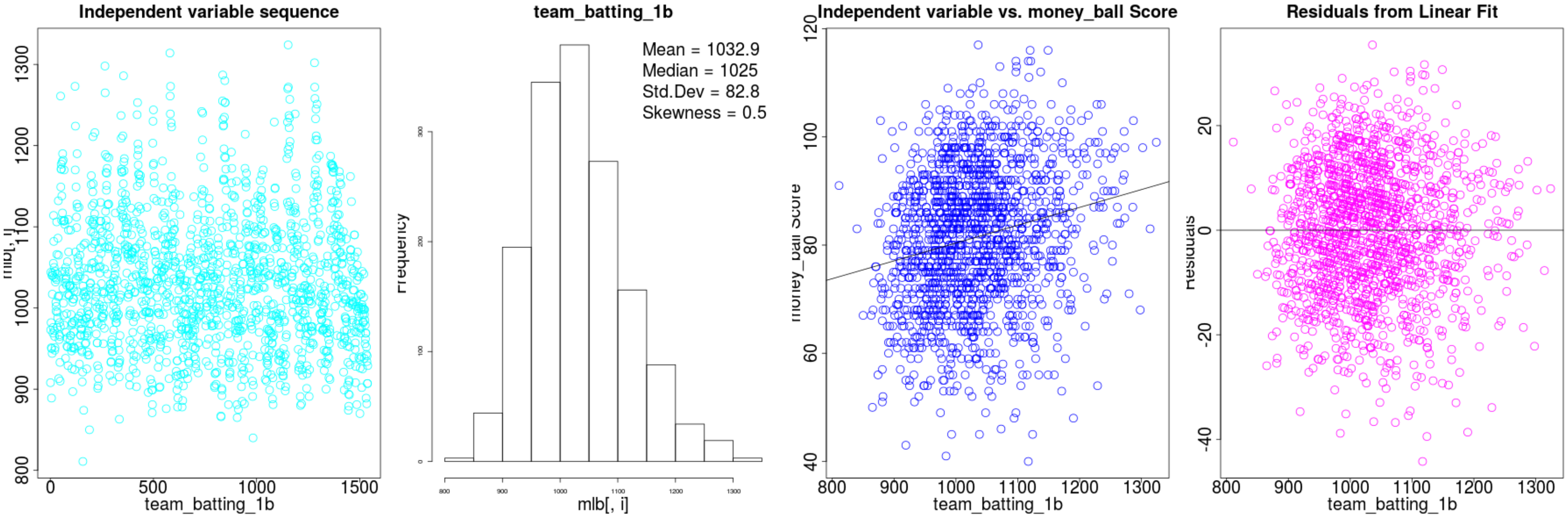


Figure 1: ###Provided data set : Team Batting Hits

---



***

4. **team__batting__slg -** new predictor added (created) from some of the existing columns. "Slugging Percentage" (SLG) is traditionally a high value predictor in SABR metrics evaluations of win predictions : http://www.baseball-reference.com/bullpen/Slugging_percentage. Although all of the characteritics that constitue the generally accepted definition of slugging percentage, (at-bats is not a part of the data set), an estimate of at-bats can be inferred based on hits, outs produced from on-base runners, and minimum number af at-bats in a game.

5. **team__batting__obp -** new predictor added. Similar to "slugging percentage", "on base percentage" (OBP) is another widely metric in SABR analyses for win predictions, http://www.baseball-reference.com/bullpen/On_base_percentage. Again, not all of the fundamental elements of OBP are included in the data set, so reasonable estimates, based on available data, were made for hit-by-pitch, sacrifice fly, and at-bats to create an estimate of OBP.

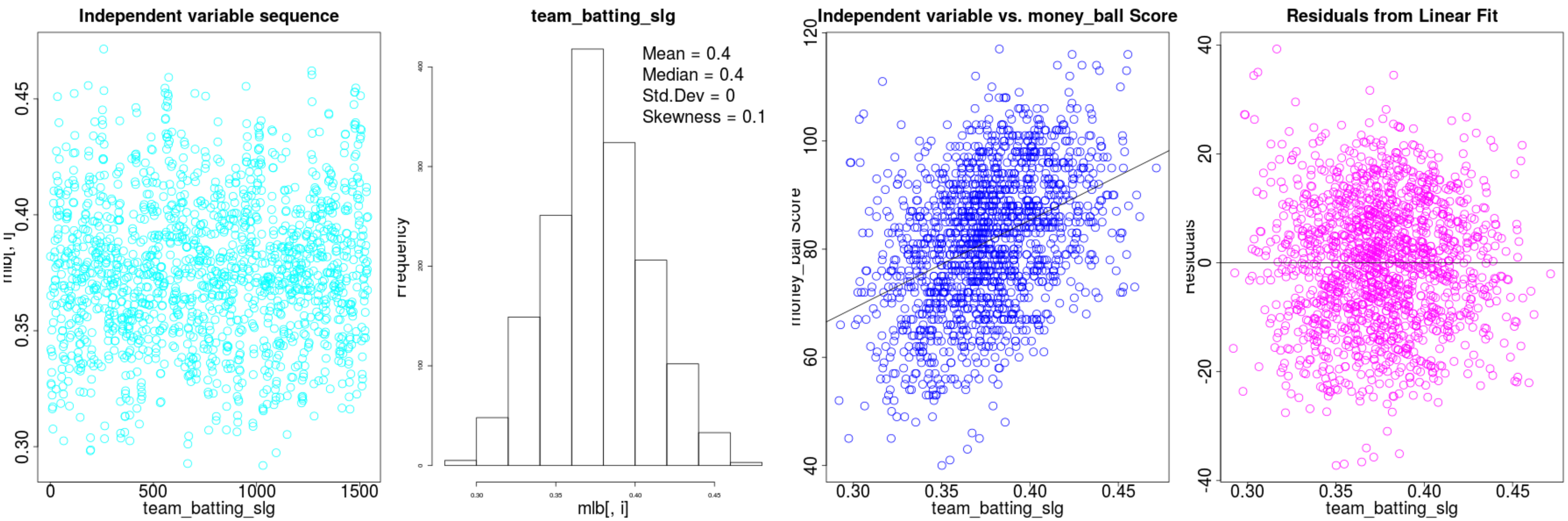
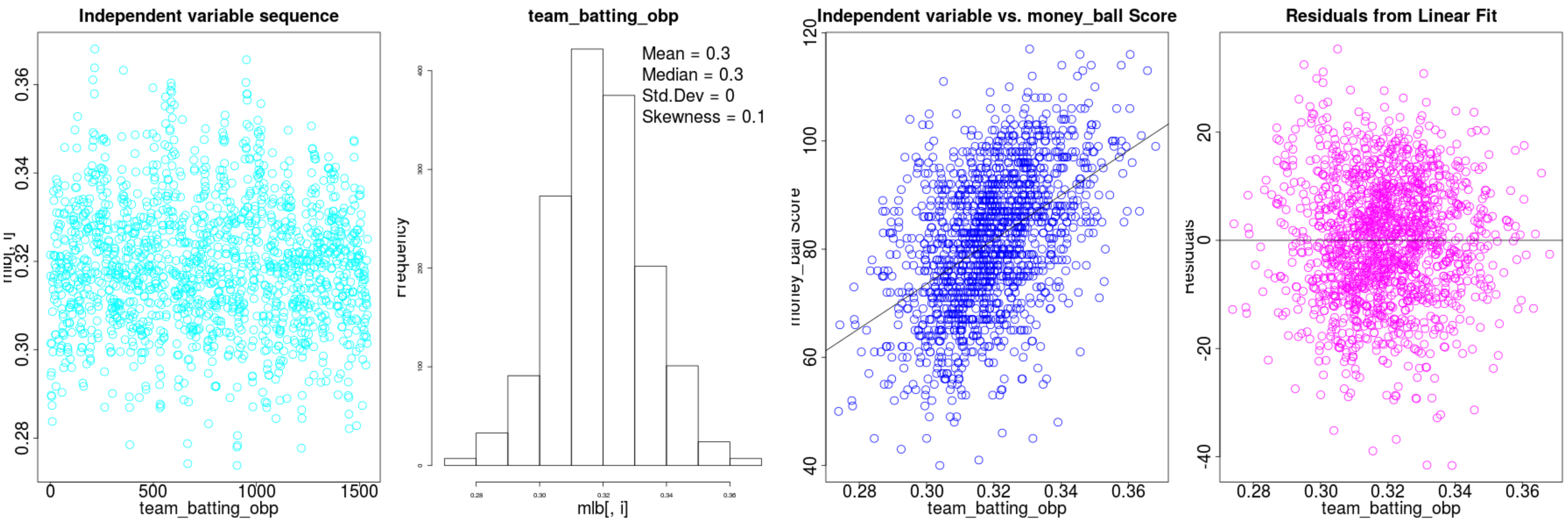


Figure 2: ###Fabricated data set : Team Batting Singles



***

6. **team_pitching_h**, **team_pitching_123b** (hits) - decomposed to create new predictor "team_pitching_123b" (singles, doubles, triples), with the idea that singles (_1b), doubles (_2b), triples (_3b), and home runs (_hr) are all a part of "hits", it is perhaps useful to separate out the potential influence of non-home-runs (_h) from the other hits, and then drop hits (_h) from the independent predictors data set. Similar to item (3), there is no loss of information from this choice.

---

7. **imputation** - here is the percentage of missing values by predictor :

| Variable | % NAs |
|---|---|
| team_batting_hbp | 91.6 |
| team_baserun_cs | 33.9 |
| team_fielding_dp | 12.6 |
| team_baserun_sb | 05.8 |
| team_batting_so | 04.5 |
| team_pitching_so | 04.5 |
| all remaining | 00.0 |

As stated above (item 1) team_batting_hbp and team_baserun_cs were dropped from the data set based on high percentage of missing values. For the remaining predictors in the above table, the missing values were *imputed* as the *mean* value of that column.

---

8. **transformation** corrections - the skewness of each predictor was measured. For every predictor for which skewness was > 0.9 (right skewed), the predictor was log-transformed. For every predictor for which skewness was < -0.9 (left skewed), the predictor was sqrt-transformed. The list of such transformed predictors is as follows :

| Variable |
|---|
| log_team_batting_3b |
| log_team_baserun_sb |
| log_team_pitching_bb |
| log_team_pitching_so |
| log_team_fielding_e |
| log_team_batting_1b |
| log_team_pitching_123b |
| sqrt_team_batting_bb |

---

9. **outlier removals** - several values are clearly outside the experience of MLB historical values, so they were eliminated from the data set. As an example, there were several values of *target_wins* which are greater than or less than the historical recorded most wins or most losses of any team in recorded baseball history. So, any *target_win* that was greater than 76% victories in a 162 game season or less than 24% victories in a 162 game season were eliminated from the data set. Similarly, other values outside historical experience were removed.

---

10. **additional transformations** - based on visual inspection of the remaining predictors, their single parameter linear regression characteristic, and the residuals from those linear fits, some additional transformations were completed :

**That comprises the ten-rules of data preparation completed on this data set. There is an auspicious sort of conjuction in having ten rules of data preparation for a data set that is predicting baseball wins - this went into extra innings !**

---

## 3. BUILD MODELS (40 Points)

Build at least three different LINEAR REGRESSION using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques.

Describe the techniques you used. If 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 model, 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.

**Baseline Model**

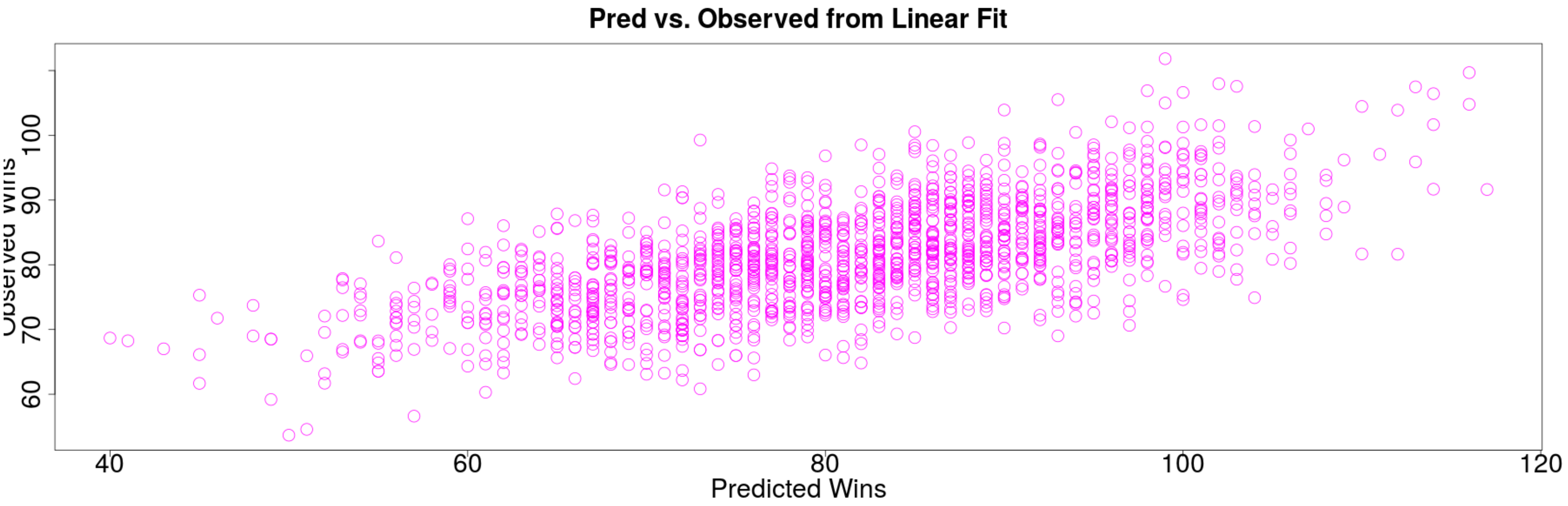


Figure 3: 1

**Summary of Models Evaluated :**

| Model | $r^2$ | Adj-$r^2$ | N features | F-stat | MSE | RSE | Features retained |
|---|---|---|---|---|---|---|---|
| Base | | | | | | | |
| Full Interaction | | | | | | | |
| SAS - Fwd | | | | | | | |

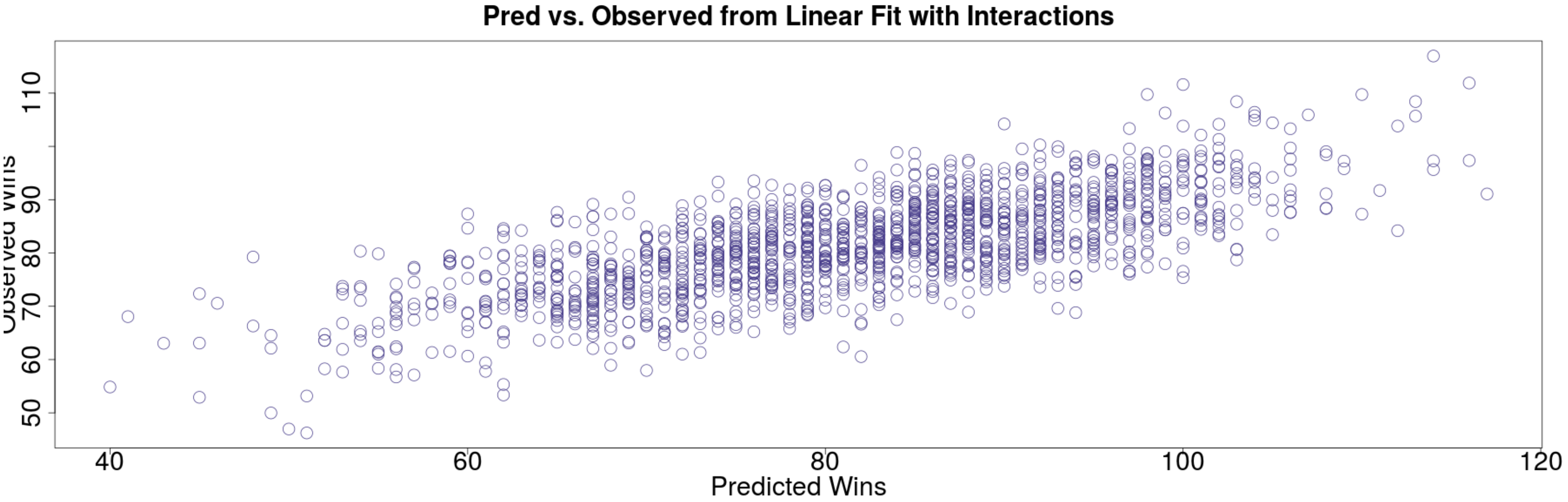

Pred vs. Observed from Linear Fit with Interactions
Observed Wins
50
60
70
80
90
110
40
60
80
100
120
Predicted Wins

Figure 4: 2

```
SAS - LARS          |
```

## 4. SELECT MODELS (40 Points)

Decide on the criteria for selecting the "Best Model". Will you use a metric such as Adjusted R-Square or AIC? Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model. STAND ALONE SCORING PROGRAM (40 POINTS)

### WRITE MODEL DEPLOYMENT CODE (40 Points)

Write a Stand Alone SAS data step that will score new data and predict the number of wins. The variable with the Predicted number of Wins should be named:

P_TARGET_WINS

The SAS data step will need to include:

a. All the variable transformations such as fixing missing values
b. The regression formula

SCORED DATA FILE (50 POINTS)

SCORE THE MONEYBALL_TEST DATA SET (50 Points)

Use the stand alone program that you wrote in the previous section. Score the data file MONEYBALL_TEST. Create a file that has only TWO variables for each record:

INDEX P_TARGET_WINS

The first variable, INDEX, will allow me to match my grading key to your predicted value. If I cannot do this, you won't get a grade. So please include this value. The second value, P_TARGET_WINS is the number of wins you believe the team will have in season based upon the data given to you.

Your values will be compared against … A Perfect Model Instructor's Model Performance of Other Students Predict the Average value for everybody (MEAN) Random Model Worst Possible Model

If your model is not better than simply using an AVERAGE value, you will receive negative points If your model is not better than generating a RANDOM value, you will receive a LOT of negative points If your model is not better than the WORST model, then it will be a WHOLE LOT of negative points.