DATA 621 Homework 1

Linear Regression

Trishita Nath

DATA EXPLORATION

The training data set contains 17 columns and 2276 rows, covering baseball team performance statistics from the years 1871 to 2006 inclusive. The data has been adjusted to match the performance of a typical 162 game season. The data-set was entirely numerical and contained no categorical variables.

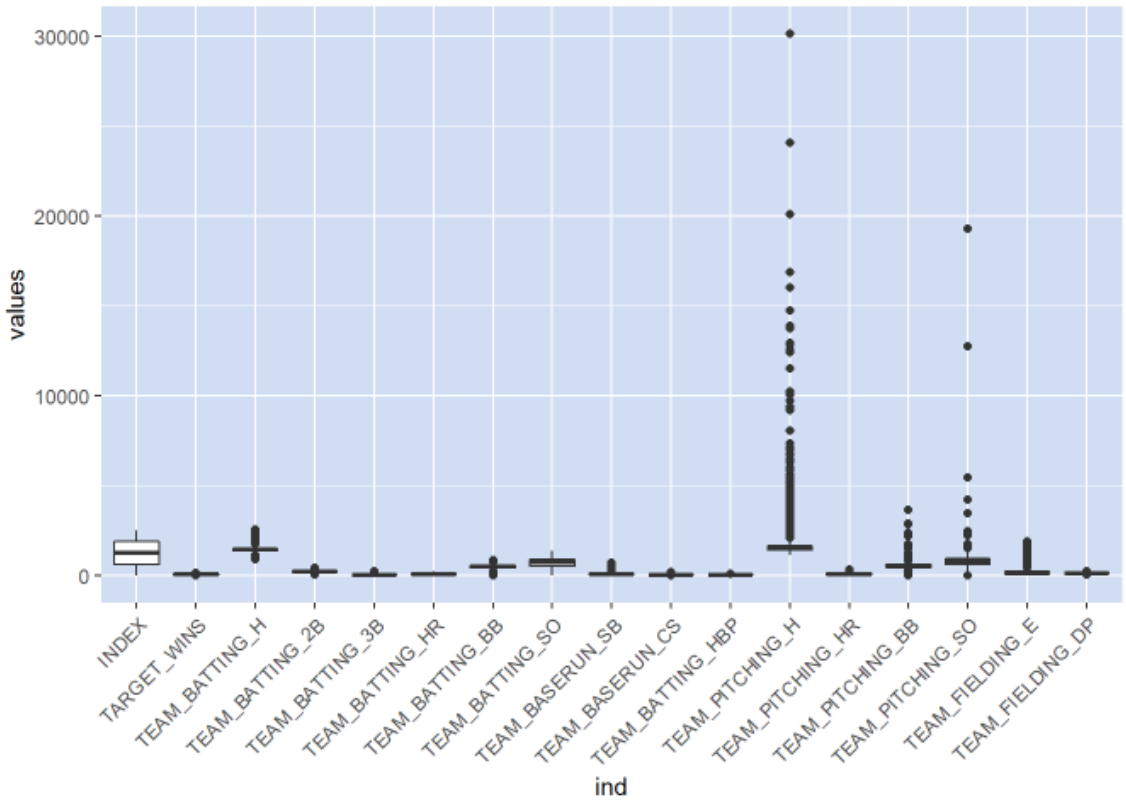
Summary stats

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
TARGET_WINS	2	191	80.92670	12.115013	82	81.11765	13.3434	43	116	73	-0.1698314
TEAM_BATTING_H	3	191	1478.62827	76.147869	1477	1477.42484	74.1300	1308	1667	359	0.1302702
TEAM_BATTING_2B	4	191	297.19895	26.329335	296	296.62745	25.2042	201	373	172	0.0915189
TEAM_BATTING_3B	5	191	30.74346	9.043878	29	30.13072	8.8956	12	61	49	0.7007420
TEAM_BATTING_HR	6	191	178.05236	32.413243	175	176.81046	35.5824	116	260	144	0.2980673
TEAM_BATTING_BB	7	191	543.31937	74.842133	535	541.31373	74.1300	365	775	410	0.3115199
TEAM_BATTING_SO	8	191	1051.02618	104.156382	1050	1046.95425	97.8516	805	1399	594	0.3985050
TEAM_BASERUN_SB	9	191	90.90576	29.916401	87	89.06536	29.6520	31	177	146	0.5553966
TEAM_BASERUN_CS	10	191	39.94241	11.898334	38	39.49020	11.8608	12	74	62	0.3468509
TEAM_BATTING_HBP	11	191	59.35602	12.967123	58	58.86275	11.8608	29	95	66	0.3185754
TEAM_PITCHING_H	12	191	1479.70157	75.788625	1480	1478.50327	72.6474	1312	1667	355	0.1279056
TEAM_PITCHING_HR	13	191	178.17801	32.391678	175	176.93464	35.5824	116	260	144	0.2989191
TEAM_PITCHING_BB	14	191	543.71728	74.916681	537	541.74510	72.6474	367	775	408	0.3144366
TEAM_PITCHING_SO	15	191	1051.81675	104.347208	1052	1047.80392	97.8516	805	1399	594	0.3945586
TEAM_FIELDING_E	16	191	107.05236	16.632162	106	106.58170	17.7912	65	145	80	0.1780432
TEAM_FIELDING_DP	17	191	152.33508	17.611682	152	152.04575	19.2738	113	204	91	0.2164822

From the data above, there are multiple variables with missing values, with TEAM-BATTING_HBP being the highest.

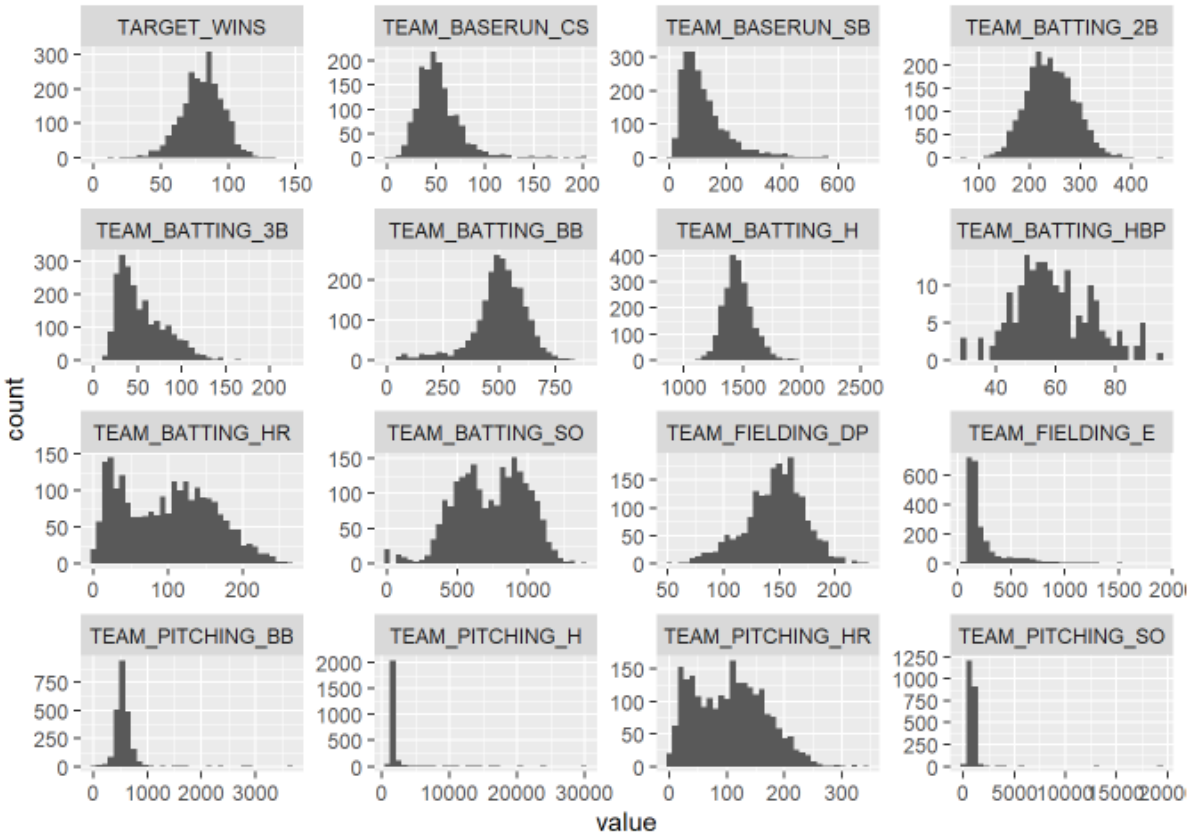
Boxplots

The boxplots below help show the spread of data within the dataset, and show various outliers. TEAM_PITCHING_H seems to have the highest spread with the most outliers.

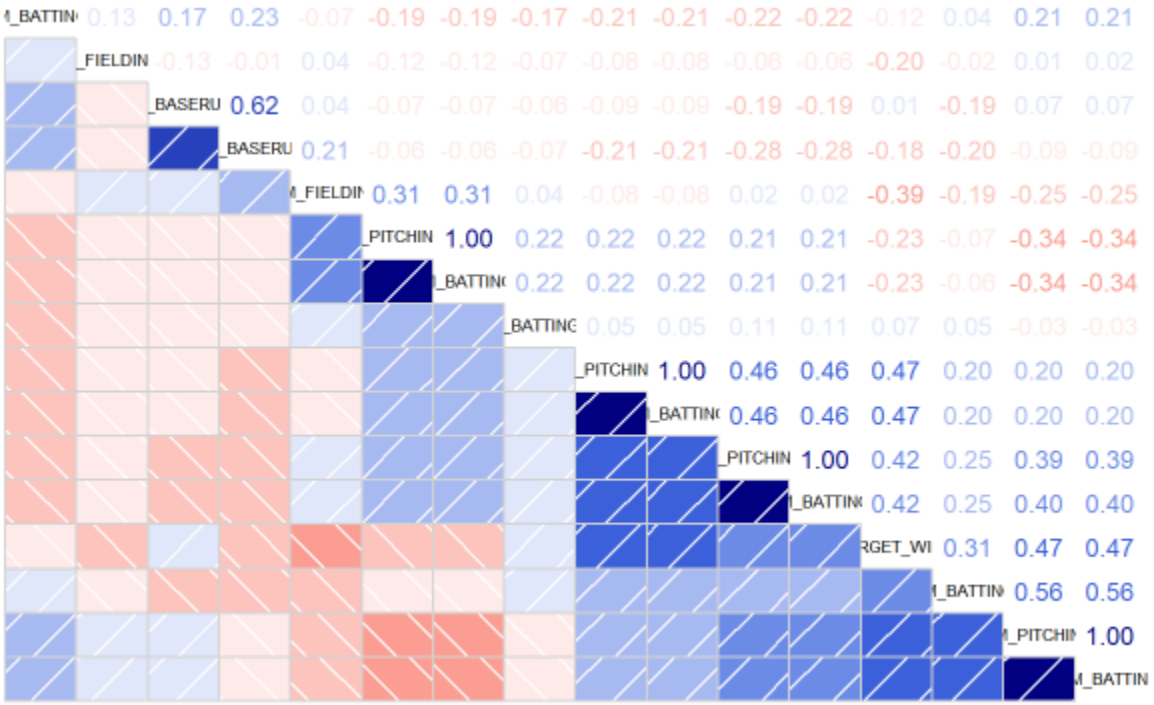


Histograms

Histograms below show distributions of the variables:



Multicollinearity



When considering features for my models, I will take into account the correlations among features so as to avoid including pairs with strong correlations.

DATA PREPARATION

Removed Fields

TEAM_BATTING_HBP is missing over 90% of its values. I will remove this variable. Variables TEAM_PITCHING_HR and TEAM_BATTING_HR are also very closely correlated with each other. I will drop the TEAM_PITCHING_HR from the dataset.

Imputation

I performed imputation via prediction using the MICE (Multivariate Imputation) library using a random forest prediction method. Variables that exceed the established threshold will be discarded to avoid collinearity issues.

```
vif(imputed)
```

```
##           Variables      VIF
## 1       TARGET_WINS 1.502612
## 2    TEAM_BATTING_H 3.976095
## 3   TEAM_BATTING_2B 2.459919
## 4   TEAM_BATTING_3B 3.063608
## 5   TEAM_BATTING_HR 4.848725
## 6   TEAM_BATTING_BB 5.433299
## 7   TEAM_BATTING_SO 5.188786
## 8   TEAM_BASERUN_SB 2.548269
## 9   TEAM_BASERUN_CS 2.306627
## 10  TEAM_PITCHING_H 3.736032
## 11 TEAM_PITCHING_BB 4.598620
## 12 TEAM_PITCHING_SO 2.861129
## 13  TEAM_FIELDING_E 4.816228
## 14 TEAM_FIELDING_DP 1.882724
```

```
v1 <- vifstep(imputed, th=10)
```

Final output is as follows:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	ske
TARGET_WINS	1	2276	80.79086	15.75215	82.0	81.31229	14.8260	0	146	146	-0.39872
TEAM_BATTING_H	2	2276	1469.26977	144.59120	1454.0	1459.04116	114.1602	891	2554	1663	1.57133
TEAM_BATTING_2B	3	2276	241.24692	46.80141	238.0	240.39627	47.4432	69	458	389	0.21510
TEAM_BATTING_3B	4	2276	55.25000	27.93856	47.0	52.17563	23.7216	0	223	223	1.10946
TEAM_BATTING_HR	5	2276	99.61204	60.54687	102.0	97.38529	78.5778	0	264	264	0.18604
TEAM_BATTING_BB	6	2276	501.55888	122.67086	512.0	512.18331	94.8864	0	878	878	-1.02575
TEAM_BATTING_SO	7	2276	728.61160	246.30216	734.5	733.56476	280.9527	0	1399	1399	-0.23610
TEAM_BASERUN_SB	8	2276	132.02768	96.16855	105.0	116.24698	66.7170	0	697	697	1.85734
TEAM_BASERUN_CS	9	2276	69.13576	42.22830	55.0	62.29254	25.2042	0	201	201	1.45251
TEAM_PITCHING_H	10	2276	1779.21046	1406.84293	1518.0	1555.89517	174.9468	1137	30132	28995	10.32951
TEAM_PITCHING_BB	11	2276	553.00791	166.35736	536.5	542.62459	98.5929	0	3645	3645	6.74389
TEAM_PITCHING_SO	12	2276	812.25000	551.15033	801.5	789.38090	257.2311	0	19278	19278	21.69267
TEAM_FIELDING_E	13	2276	246.48067	227.77097	159.0	193.43798	62.2692	65	1898	1833	2.99046
TEAM_FIELDING_DP	14	2276	142.08304	28.45915	146.0	143.27607	26.6868	52	228	176	-0.36700

Build Models

I will build 3 different linear regression models, to determine which one provides the best prediction for the number of wins. These are:

- All variables
- Only significant variables
- Backwards elimination of each variable

Model 1: All Variables

After the data has been imputed, all of the variables will be tested to determine the base model they provided:

```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.821  -8.487   0.129   8.326  60.384
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      27.0874692  5.2326389   5.177 2.46e-07 ***
## TEAM_BATTING_H    0.0469766  0.0035928  13.075  < 2e-16 ***
## TEAM_BATTING_2B  -0.0198842  0.0090450  -2.198 0.028024 *
## TEAM_BATTING_3B   0.0429280  0.0169030   2.540 0.011162 *
## TEAM_BATTING_HR   0.0737002  0.0097034   7.595 4.46e-14 ***
## TEAM_BATTING_BB   0.0128160  0.0051270   2.500 0.012499 *
## TEAM_BATTING_SO  -0.0127208  0.0024845  -5.120 3.31e-07 ***
## TEAM_BASERUN_SB   0.0385042  0.0044113   8.729  < 2e-16 ***
## TEAM_BASERUN_CS   0.0010757  0.0097175   0.111 0.911867
## TEAM_PITCHING_H  -0.0001948  0.0003712  -0.525 0.599834
## TEAM_PITCHING_BB -0.0025510  0.0034825  -0.733 0.463929
## TEAM_PITCHING_SO  0.0027762  0.0008272   3.356 0.000803 ***
## TEAM_FIELDING_E  -0.0305350  0.0025229 -12.103  < 2e-16 ***
## TEAM_FIELDING_DP -0.1092527  0.0128228  -8.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.89 on 2262 degrees of freedom
## Multiple R-squared:  0.3345, Adjusted R-squared:  0.3307
## F-statistic: 87.45 on 13 and 2262 DF,  p-value: < 2.2e-16
```

F-statistic is 87.45, R-squared is 0.3307.

Model 2: Highly Significant Variables Only

This model will focus only on the variables that are statistically significant. Variables will be chosen based on their significance level from the R output:

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_3B +
##     TEAM_BATTING_HR + TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_SO +
##     TEAM_FIELDING_E + TEAM_FIELDING_DP, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.663  -8.435   0.179   8.387  59.742
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      36.3332598  4.5257753   8.028 1.57e-15 ***
## TEAM_BATTING_H    0.0406282  0.0026570  15.291  < 2e-16 ***
## TEAM_BATTING_3B   0.0532329  0.0163026   3.265  0.00111 **
## TEAM_BATTING_HR   0.0806071  0.0091762   8.784  < 2e-16 ***
## TEAM_BATTING_SO  -0.0138112  0.0023188  -5.956 2.98e-09 ***
## TEAM_BASERUN_SB   0.0422772  0.0038259  11.050  < 2e-16 ***
## TEAM_PITCHING_SO  0.0018721  0.0005711   3.278  0.00106 **
## TEAM_FIELDING_E  -0.0344358  0.0018068 -19.058  < 2e-16 ***
## TEAM_FIELDING_DP -0.1038711  0.0125473  -8.278  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.92 on 2267 degrees of freedom
## Multiple R-squared:  0.3299, Adjusted R-squared:  0.3275
## F-statistic: 139.5 on 8 and 2267 DF,  p-value: < 2.2e-16
```

F-statistic is 139.5, R-squared is 0.3275. The F-statistic is better than the first model but the R-squared drops slightly.

Model 3: Backwards Elimination

Variables are removed one by one to determine best fit model. After each variable is removed, the model is re-run until the most optimal output values are produced. Only the final output is shown.

```
##
## Call:
## lm(formula = TARGET_WINS ~ TEAM_BATTING_H + TEAM_BASERUN_SB +
##     TEAM_FIELDING_E + TEAM_BATTING_HR, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.176  -9.038  -0.037   8.484  51.048
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.203121   2.907188   1.790 0.073628 .
## TEAM_BATTING_H   0.050129   0.002027  24.734  < 2e-16 ***
## TEAM_BASERUN_SB  0.048187   0.003426  14.064  < 2e-16 ***
## TEAM_FIELDING_E -0.027104   0.001655 -16.378  < 2e-16 ***
## TEAM_BATTING_HR  0.022628   0.005964   3.794 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.22 on 2271 degrees of freedom
## Multiple R-squared:  0.2964, Adjusted R-squared:  0.2952
## F-statistic: 239.2 on 4 and 2271 DF,  p-value: < 2.2e-16
```

F-statistic is 239.2, R-squared is 0.2964 The F-statistic is larger than both of the other two models, however the R-squared is slightly lower than the other two

SELECT MODELS

My model of choice would be Model 3 due to the following reasons:

- There is greater statistical significance with the third model relative to the others and uses less unnecessary variables to compute our prediction without compromising the adjusted R^2 value.
- Model 3 seems to have lower VIF scores than Models 1 and 2.
- It has better adjustment for multicollinearity

Predictions

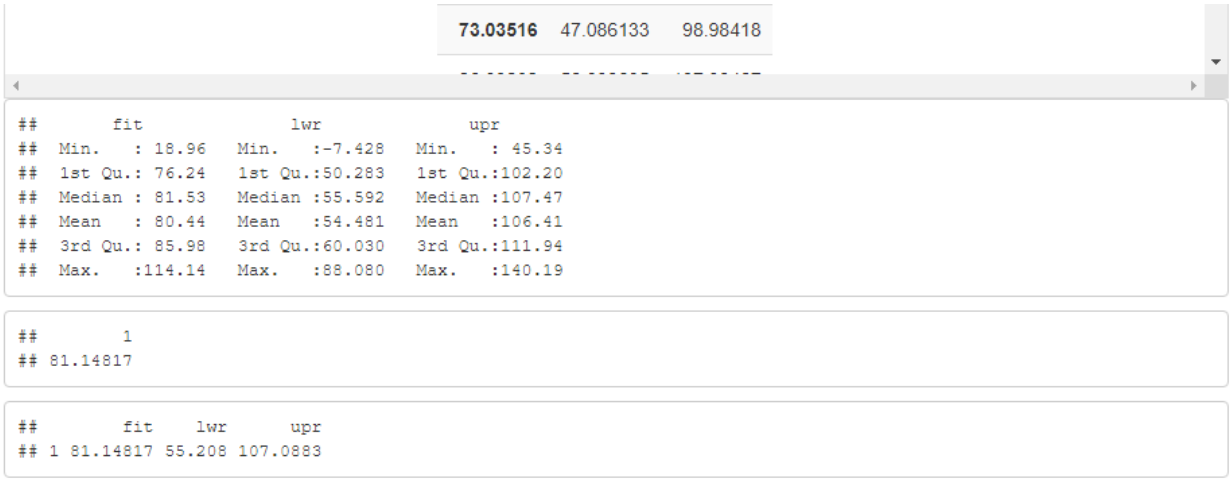
The evaluation dataset would be subjected to the same processing as the training data set.

Stats

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
TEAM_BATTING_H	1	259	1469.38996	150.65523	1455	1463.68421	114.1602	819	2170	1351	0.587613
TEAM_BATTING_2B	2	259	241.32046	49.51612	239	242.32536	48.9258	44	376	332	-0.327328
TEAM_BATTING_3B	3	259	55.91120	27.14410	52	52.94737	26.6868	14	155	141	0.979028
TEAM_BATTING_HR	4	259	95.63320	56.33221	101	93.67943	66.7170	0	242	242	0.171236
TEAM_BATTING_BB	5	259	498.95753	120.59215	509	505.98086	94.8864	15	792	777	-0.920991
TEAM_BATTING_SO	6	259	701.34749	245.20399	680	708.99043	262.4202	0	1268	1268	-0.326644
TEAM_BASERUN_SB	7	259	128.40154	98.10097	95	112.63636	63.7518	0	580	580	1.785447
TEAM_BASERUN_CS	8	259	62.32046	33.21322	55	58.13397	23.7216	0	154	154	1.132532
TEAM_PITCHING_H	9	259	1813.46332	1662.91308	1515	1554.25359	173.4642	1155	22768	21613	9.276479
TEAM_PITCHING_BB	10	259	552.41699	172.95006	526	536.46411	97.8516	136	2008	1872	4.111377
TEAM_PITCHING_SO	11	259	795.91120	613.15636	736	763.76077	246.1116	0	9963	9963	12.874591
TEAM_FIELDING_E	12	259	249.74903	230.90260	163	197.36364	59.3040	73	1568	1495	3.088726
TEAM_FIELDING_DP	13	259	142.79537	27.52350	146	143.98565	25.2042	69	204	135	-0.396554

After imputing and cleaning the data, using the predict function and Model 3, the following are the predicted values including prediction intervals:

fit	lwr	upr
66.87967	40.918671	92.84067
67.34438	41.384721	93.30404
75.85163	49.906773	101.79649
89.71960	63.770921	115.66827
71.03047	45.071817	96.98913
70.57429	44.617863	96.53071
82.03138	56.057257	108.00550
75.40230	49.454234	101.35036
69.93792	43.975706	95.90013
74.06368	48.110912	100.01644
75.53109	49.578306	101.48387
82.15377	56.206463	108.10108
78.37499	52.420100	104.32988
80.28636	54.335374	106.23734



	73.03516	47.086133	98.98418

```
##       fit              lwr              upr
##  Min.   : 18.96   Min.   :-7.428   Min.   : 45.34
##  1st Qu.: 76.24   1st Qu.:50.283   1st Qu.:102.20
##  Median : 81.53   Median :55.592   Median :107.47
##  Mean   : 80.44   Mean   :54.481   Mean   :106.41
##  3rd Qu.: 85.98   3rd Qu.:60.030   3rd Qu.:111.94
##  Max.   :114.14   Max.   :88.080   Max.   :140.19
```

```
##        1
## 81.14817
```

```
##        fit    lwr      upr
## 1 81.14817 55.208 107.0883
```

References

- A Modern Approach to Regression with R: Simon Sheather
- Linear Models with R: Julian Faraway.
- R package vignette, mixtools: An R Package for Analyzing Finite Mixture Models
- Detecting Multicolinearity with VIF

Appendix

Moneyball Dataset Columns

- INDEX: Identification Variable(Do not use)
- TARGET_WINS: Number of wins
- TEAM_BATTING_H : Base Hits by batters (1B,2B,3B,HR)
- TEAM_BATTING_2B: Doubles by batters (2B)
- TEAM_BATTING_3B: Triples by batters (3B)
- TEAM_BATTING_HR: Home runs by batters (4B)
- TEAM_BATTING_BB: Walks by batters
- TEAM_BATTING_HBP: Batters hit by pitch (get a free base)
- TEAM_BATTING_SO: Strikeouts by batters
- TEAM_BASERUN_SB: Stolen bases
- TEAM_BASERUN_CS: Caught stealing
- TEAM_FIELDING_E: Errors

- TEAM_FIELDING_DP: Double Plays
- TEAM_PITCHING_BB: Walks allowed
- TEAM_PITCHING_H: Hits allowed
- TEAM_PITCHING_HR: Homeruns allowed
- TEAM_PITCHING_SO: Strikeouts by pitchers

R Code

GitHub: https://github.com/nathtrish334/Data-621/blob/main/Homework01.rmd

RPubs: https://rpubs.com/trishitanath/814750