Data 621 Homework 1

Group 3: Amanda Arce, Austin Chan, Jithendra Seneviratne, Sheryl Piechocki

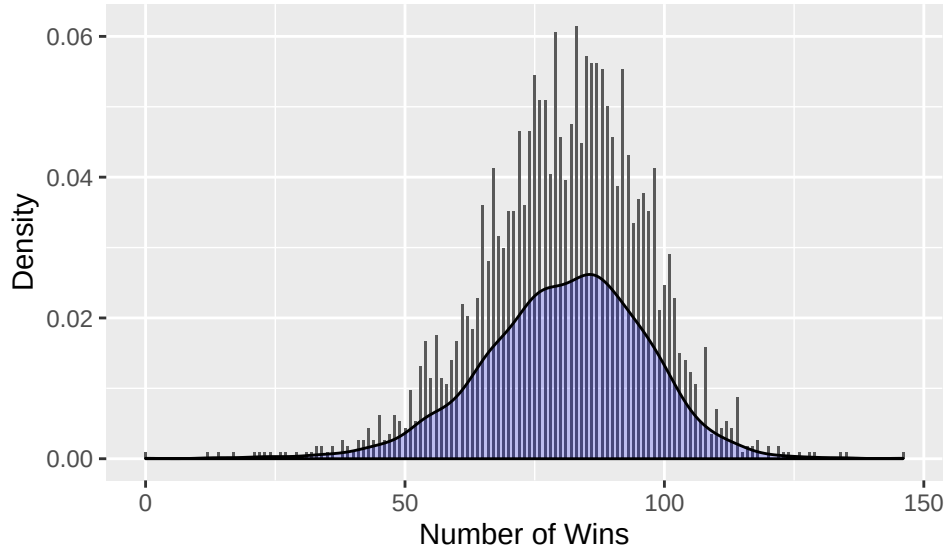
2/15/2020

Objective: Build a multiple linear regression model on the training data to predict the number of wins for baseball teams.

1. DATA EXPLORATION

The data used in this analysis, consists of performance statistics for baseball teams from the years 1871-2006. Each record represents the performanace of one team for one year. There are 2,276 records and 17 baseball statistics, including the target variable wins. Statistics include batting information, such as hits, doubles, triples, homeruns, strikeouts, and walks. Also, given are pitching statistics of hits allowed, walks allowed, homeruns allowed, and strikeouts by pitchers. Other information regarding errors, stolen bases, caught stealing, hit by pitch, and double plays is also available.

The distribution of the target variable, Wins is below. It appears to be normally distributed, with a mean of 80.79 and standard deviation 15.75.

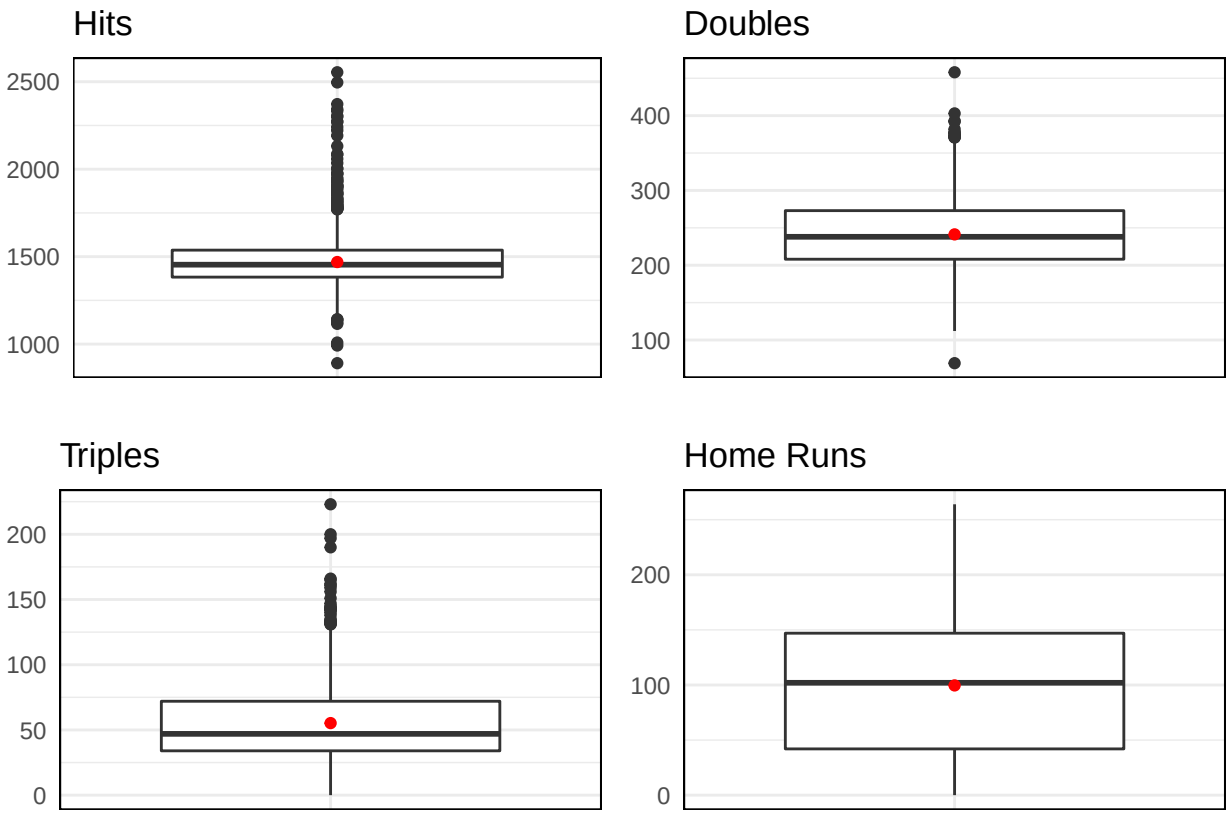


Summary statistics for each independent variable are provided below. The variables Caught Stealing and Hit by Pitch have a large number of missing values and therefore will be excluded from all subsequent analysis.

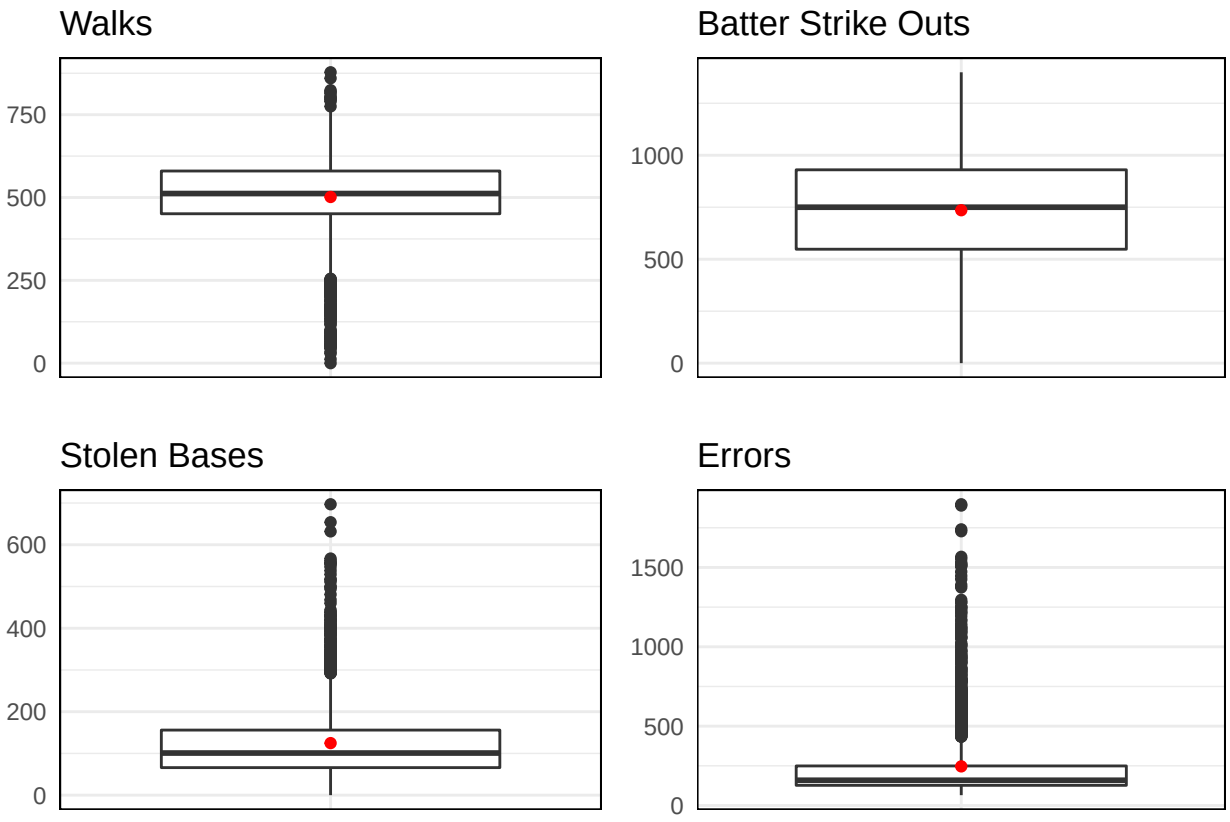
```
##       Hits         Doubles         Triples         Home Runs
##  Min.   : 891   Min.   : 69.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:1383   1st Qu.:208.0   1st Qu.: 34.00   1st Qu.: 42.00
##  Median :1454   Median :238.0   Median : 47.00   Median :102.00
##  Mean   :1469   Mean   :241.2   Mean   : 55.25   Mean   : 99.61
##  3rd Qu.:1537   3rd Qu.:273.0   3rd Qu.: 72.00   3rd Qu.:147.00
```

```
##  Max.   :2554   Max.   :458.0   Max.   :223.00   Max.   :264.00
##
##      Walks         Batter SO       Stolen Bases   Caught Stealing
##  Min.   :  0.0   Min.   :   0.0   Min.   :  0.0   Min.   :  0.0
##  1st Qu.:451.0   1st Qu.: 548.0   1st Qu.: 66.0   1st Qu.: 38.0
##  Median :512.0   Median : 750.0   Median :101.0   Median : 49.0
##  Mean   :501.6   Mean   : 735.6   Mean   :124.8   Mean   : 52.8
##  3rd Qu.:580.0   3rd Qu.: 930.0   3rd Qu.:156.0   3rd Qu.: 62.0
##  Max.   :878.0   Max.   :1399.0   Max.   :697.0   Max.   :201.0
##                  NA's   :102      NA's   :131     NA's   :772
##   Hit by Pitch     Hits Allow    Home Runs Allow  Walks Allow
##  Min.   :29.00   Min.   : 1137   Min.   :  0.0   Min.   :   0.0
##  1st Qu.:50.50   1st Qu.: 1419   1st Qu.: 50.0   1st Qu.: 476.0
##  Median :58.00   Median : 1518   Median :107.0   Median : 536.5
##  Mean   :59.36   Mean   : 1779   Mean   :105.7   Mean   : 553.0
##  3rd Qu.:67.00   3rd Qu.: 1682   3rd Qu.:150.0   3rd Qu.: 611.0
##  Max.   :95.00   Max.   :30132   Max.   :343.0   Max.   :3645.0
##  NA's   :2085
##    Pitcher SO          Errors        Double Plays
##  Min.   :    0.0   Min.   :  65.0   Min.   : 52.0
##  1st Qu.:  615.0   1st Qu.: 127.0   1st Qu.:131.0
##  Median :  813.5   Median : 159.0   Median :149.0
##  Mean   :  817.7   Mean   : 246.5   Mean   :146.4
##  3rd Qu.:  968.0   3rd Qu.: 249.2   3rd Qu.:164.0
##  Max.   :19278.0   Max.   :1898.0   Max.   :228.0
##  NA's   :102                        NA's   :286
```

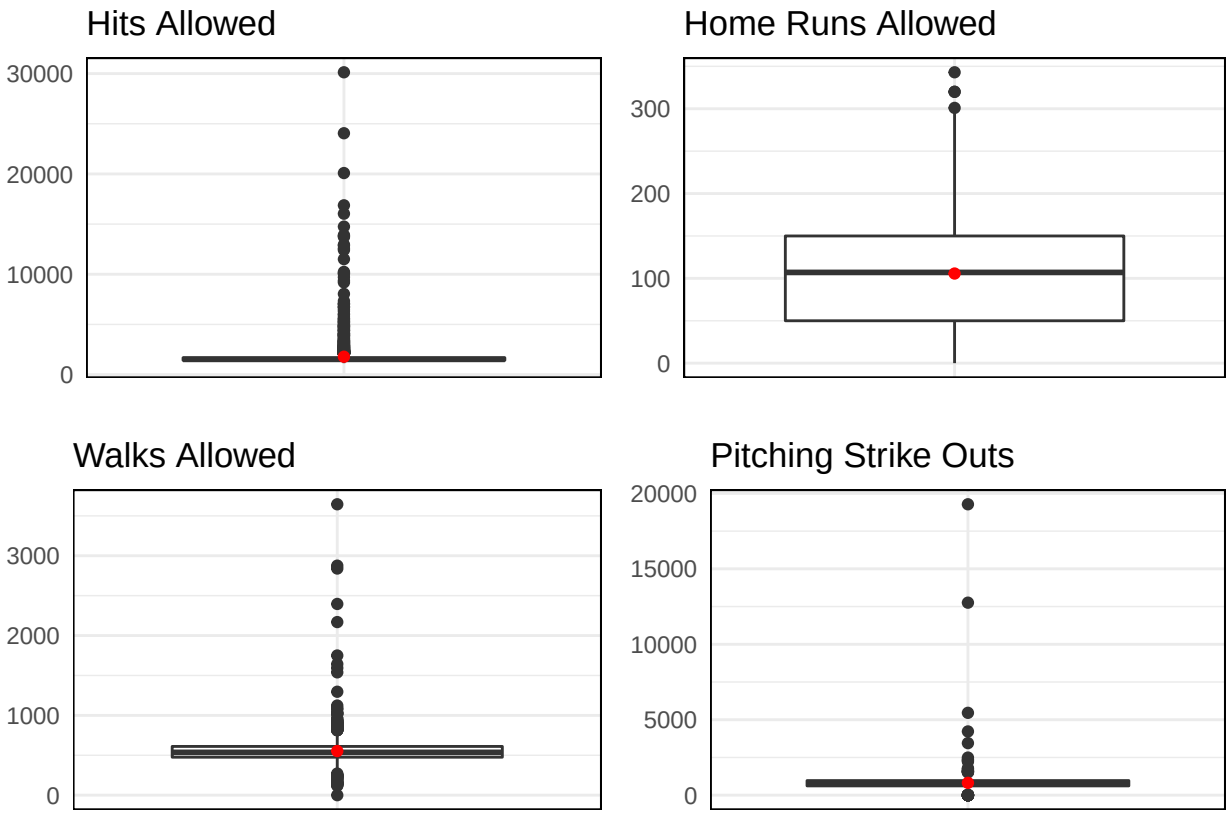
The box plots of the batting variables show many outliers in Hits and Triples. The spread on Home Runs is large.



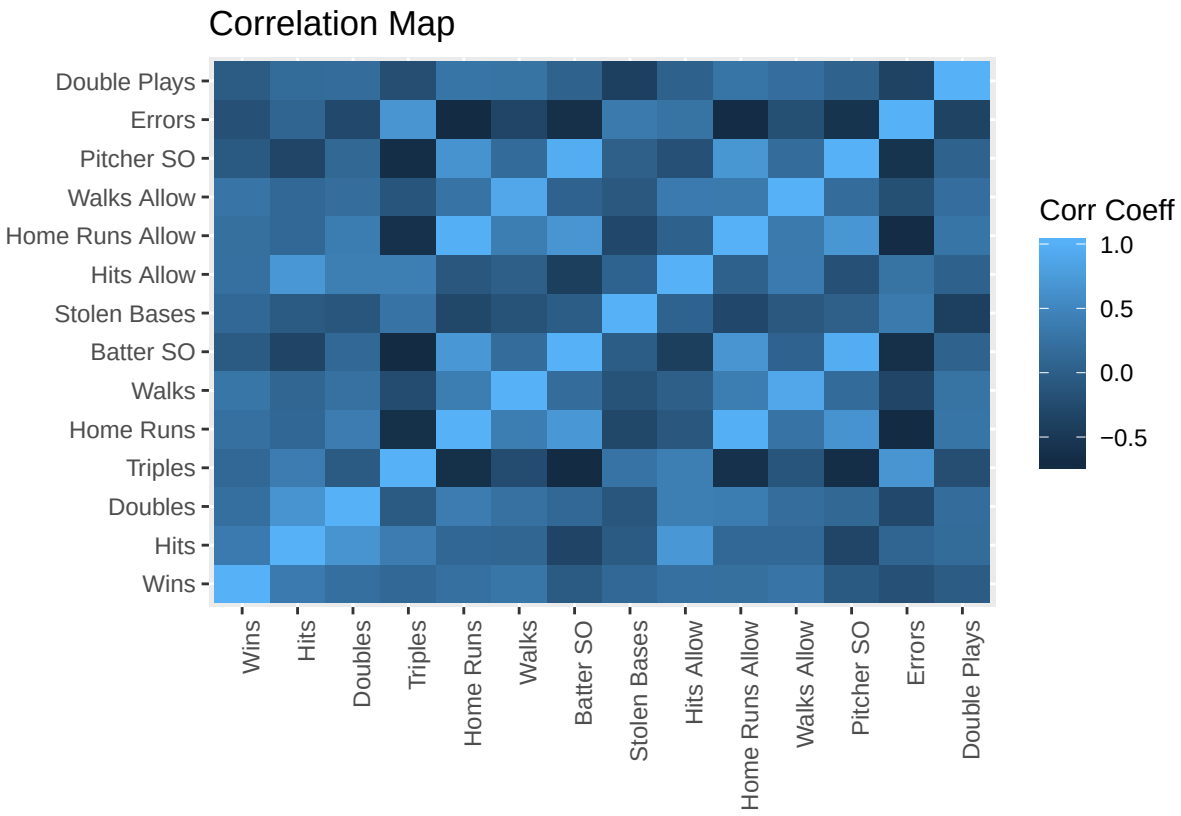
Further, Walks, Stolen Bases, and Errors also have quite a few outliers.



The pitching statistics box plots reveal many outliers in Hits Allowed, Walks Allowed, and Pitching Strike Outs. As was seen in Home Runs, the Home Runs Allowed has a large spread.



The correlation matrix below provides some insight into the data. Wins has the highest positive correlation with Hits and Walks, and negative correlation with Errors. In addition, the batting variables that have corresponding pitching variables are highly correlated, i.e. Walks is highly positively correlated with Walks Allowed, Strike Outs is highly positively correlated with Pitcher Strike Outs, etc. Other interesting correlations found are: HOme Runs and Errors are negatively correlated, Triples and Batter Strike Outs are negatively correlated, and Home Runs and Batter Strike Outs are positively correlated.



Fit regression model to variables with few missing values. get rid of missning values.

```
fit <- lm(formula = Wins ~ ., data = train_data[c(2:9,12:16)][complete.cases(train_data[c(2:9,12:16)]),
summary(fit) # show result
```

```
##
## Call:
## lm(formula = Wins ~ ., data = train_data[c(2:9, 12:16)][complete.cases(train_data[c(2:9,
##     12:16)]), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.960  -7.929   0.237   7.578  50.199
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       18.6512751  5.1545536   3.618 0.000304 ***
## Hits               0.0388959  0.0038004  10.235  < 2e-16 ***
## Doubles           -0.0475523  0.0090949  -5.228 1.88e-07 ***
## Triples            0.0784471  0.0174928   4.485 7.72e-06 ***
## `Home Runs`        0.0326904  0.0290518   1.125 0.260619
## Walks              0.0279988  0.0066135   4.234 2.40e-05 ***
## `Batter SO`       -0.0054466  0.0053050  -1.027 0.304695
## `Stolen Bases`     0.0588649  0.0042986  13.694  < 2e-16 ***
```

```
## `Hits Allow`       0.0024753  0.0004045   6.119 1.13e-09 ***
## `Home Runs Allow`  0.0457063  0.0262823   1.739 0.082177 .
## `Walks Allow`     -0.0047439  0.0052439  -0.905 0.365764
## `Pitcher SO`      -0.0066348  0.0047258  -1.404 0.160493
## Errors            -0.0495831  0.0032314 -15.344  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.62 on 2030 degrees of freedom
## Multiple R-squared:  0.3621, Adjusted R-squared:  0.3583
## F-statistic: 96.01 on 12 and 2030 DF,  p-value: < 2.2e-16
```

Check to see how many rows were retained

```
nrow(train_data[c(2:9,12:16)][complete.cases(train_data[c(2:9,12:16)]),])
```

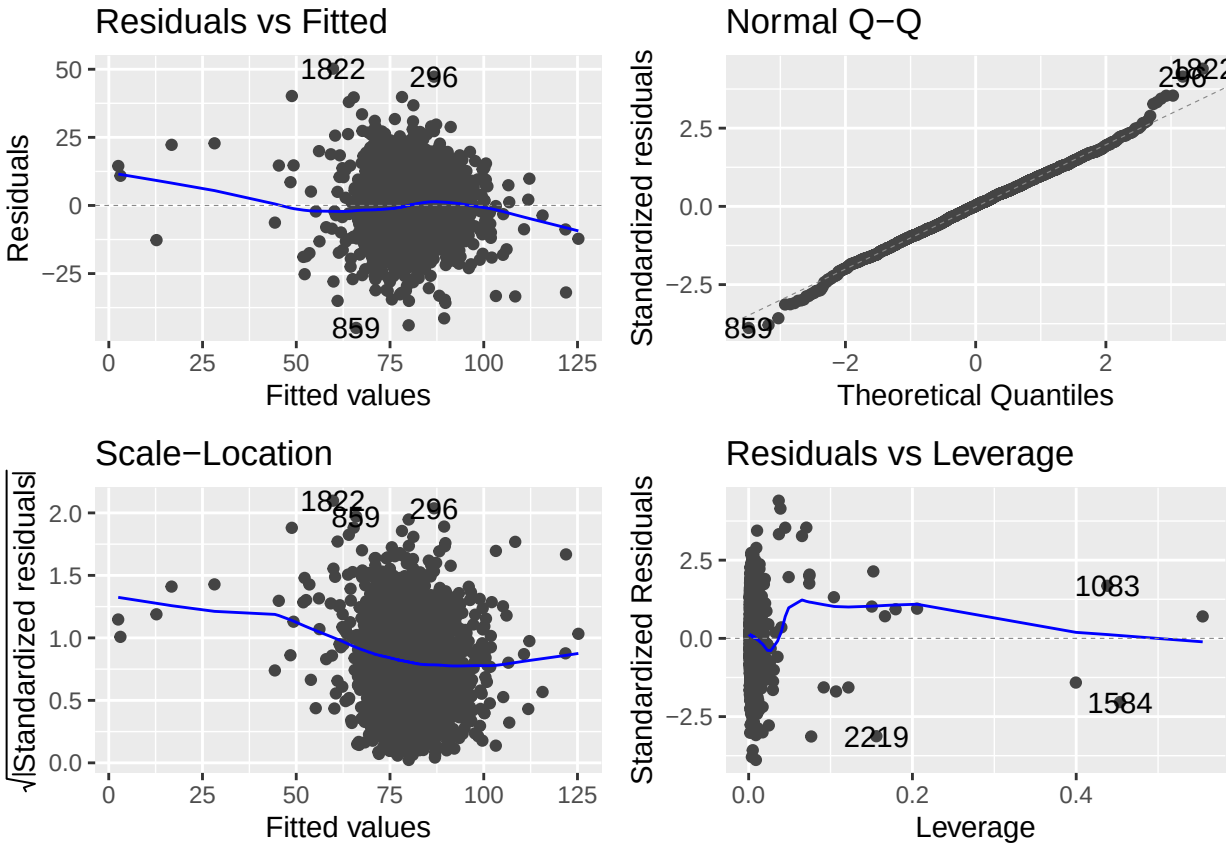
```
## [1] 2043
```

Analyze residuals

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 3.5.3
```

```
autoplot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
1822
296
859
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized Residuals
Leverage
1083
1584
2219