Statistical Modeling and Regression Analysis Report

Amit Kumar(M22MA202)
Sunil Choudhary(M22MA206)
Veeresh Chaudhary(M22MA208)

Contents

Abstract

This report presents an extensive statistical analysis of the `Hitters` dataset using various regression techniques and model evaluation strategies. The study includes best subset selection based on multiple criteria (Adjusted R^2, Mallows' C_p, BIC), diagnostics of linear models using residual analysis and influence measures, transformation of the response variable via Box-Cox method, and model performance validation using validation sets and K-fold cross-validation. Regularization techniques such as Ridge, Lasso, and Elastic Net are employed for improving prediction accuracy. Logistic regression is applied to classification tasks, and Poisson regression models are also explored. Visualizations, residual analyses, and model comparisons support the final conclusions.

1 Introduction

This report presents a comprehensive statistical analysis performed on the `Hitters` dataset from the `ISLR2` package. Various regression techniques including best subset selection, ridge regression, lasso, elastic net, logistic regression, and Poisson regression have been implemented to identify influential predictors and build predictive models for both regression and classification problems.

2 Data Preprocessing

Data used: Hitters

Data Description: The `Hitters` dataset is a part of the `ISLR2` package in R and contains information on 322 Major League Baseball (MLB) players from the 1986 season. The dataset includes various performance statistics and demographic attributes for each player. The primary goal of analyzing this dataset is to predict a player's salary based on available predictors using regression and classification techniques.
Dataset Structure The dataset contains 20 variables, out of which 19 are predictors and one is the response variable (`Salary`). The predictors include both numerical and categorical features.
Variables

- **AtBat**: Number of times at bat in 1986.
- **Hits**: Number of hits in 1986.
- **HmRun**: Number of home runs in 1986.
- **Runs**: Number of runs in 1986.
- **RBI**: Number of runs batted in 1986.
- **Walks**: Number of walks in 1986.
- **Years**: Number of years in the major leagues.
- **CAtBat**: Number of times at bat during career.

- **CHits**: Number of hits in career.
- **CHmRun**: Number of home runs in career.
- **CRuns**: Number of runs in career.
- **CRBI**: Number of RBIs in career.
- **CWalks**: Number of walks in career.
- **League**: League player belongs to at the beginning of 1986 (A or N).
- **Division**: Player's division at the beginning of 1986 (E or W).
- **PutOuts**: Number of putouts in 1986.
- **Assists**: Number of assists in 1986.
- **Errors**: Number of errors in 1986.
- **Salary**: Player's salary in thousands of dollars (response variable).
- **NewLeague**: League player belongs to at the end of 1986 (A or N).

Missing Data Hanndling

The `Salary` variable contains some missing values which were handled by replacing them with the mean salary across the dataset to ensure completeness for modeling.

Use in Analysis

This dataset is suitable for demonstrating techniques such as:

- Linear and Multiple Regression
- Subset Selection
- Shrinkage Methods (Ridge, Lasso)
- Classification (Logistic Regression)
- Model Diagnostics and Transformations

Missing values in numeric columns were replaced with their respective column means to ensure a complete dataset for modeling.

3 Best Subset Selection

The `leaps` package was used for performing best subset regression.

- Fitted all possible 19 models.
- **Model Selection Criteria** Three criteria were used to select the best subset:
  - **Adjusted R^2:** Shows Model with 11 variables that were given by best subset selection having a maximum value of R^2.

```
>
> #################### Best model based on R^2 Evaluation ##############
> summary$adjr2
 [1] 0.2516046 0.3196913 0.3538306 0.3728518 0.3899095 0.4017797 0.4120230 0.4191607
 [9] 0.4248030 0.4260848 0.4264010 0.4262514 0.4254134 0.4246605 0.4233506 0.4215667
[17] 0.4198081 0.4179665 0.4160403
```

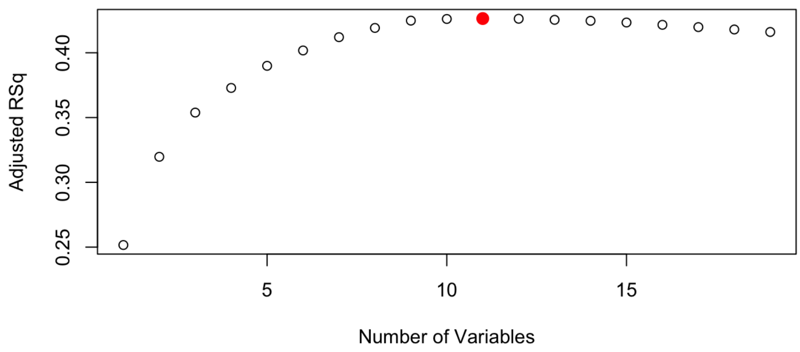


Figure 1: Adjusted R^2 vs Number of Variables

  - **Mallows' C_p:** Shows Model with 9 variables that were given by best subset selection having a minimum value of CP.

```
>
> #################### Best model based on CP Evaluation ##############
> summary$cp
 [1] 92.108003 55.632628 37.876839 28.444701 20.140263 14.692484 10.160153  7.327511
 [9]  5.318249  5.650635  6.499968  7.596839  9.056349 10.468202 12.169372 14.113606
[17] 16.038543 18.000546 20.000000
```

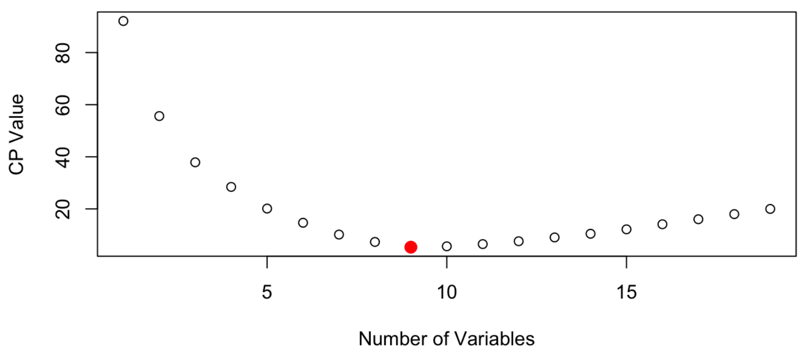


Figure 2: Mallows' C_p vs Number of Variables

- **Bayesian Information Criterion (BIC):**Shows Model with 7 variables that were given by best subset selection having a minimum value of BIC.

```
>
> #################### Best model based on BIC Evaluation ##############
> summary$bic
 [1]  -82.77885 -108.72604 -120.54061 -125.40124 -129.52339 -131.09617 -131.90681 -131.09217
 [9] -129.49125 -125.46874 -120.90869 -116.09059 -110.88980 -105.74077 -100.28450  -94.56938
[17]  -88.87485  -83.14080  -77.36684
```

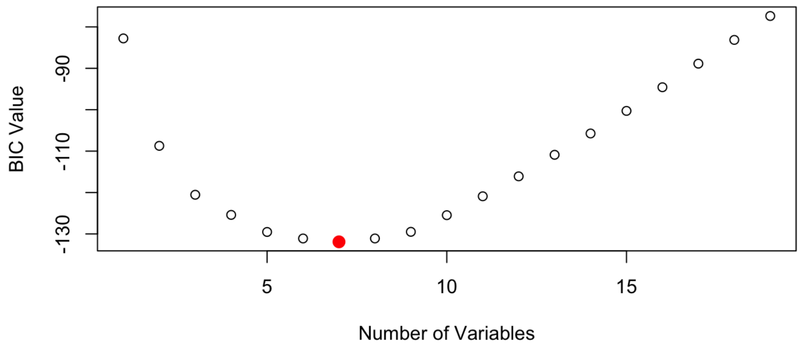


Figure 3: BIC vs Number of Variables

4 Linear Regression Modeling

Linear models were fitted using the best subset of predictors with a maximum R^2 value. Diagnostic checks were used to identify outliers and influential points, including PRESS residuals, standardized/studentized residuals, leverage, Cook's distance, DFFITS, DFBETA, and CovRatio to detect outliers.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      340.44783   68.76583   4.951 1.22e-06 ***
X_bestAtBat       -2.18545    0.48680  -4.489 1.01e-05 ***
X_bestHits         6.42895    1.51632   4.240 2.96e-05 ***
X_bestWalks        5.24144    1.47133   3.562 0.000425 ***
X_bestYears      -11.42095   10.55213  -1.082 0.279944
X_bestCAtBat      -0.08037    0.05991  -1.342 0.180709
X_bestCRuns        1.25533    0.37259   3.369 0.000849 ***
X_bestCRBI         0.51418    0.17746   2.897 0.004030 **
X_bestCWalks      -0.69936    0.23229  -3.011 0.002820 **
X_bestDivisionW -109.44647   34.98837  -3.128 0.001927 **
X_bestPutOuts      0.23864    0.06710   3.556 0.000435 ***
X_bestAssists      0.20871    0.14943   1.397 0.163493
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308.7 on 310 degrees of freedom
Multiple R-squared:  0.4461,	Adjusted R-squared:  0.4264
F-statistic: 22.69 on 11 and 310 DF,  p-value: < 2.2e-16
```

Figure 4: summary for best fitted R^2

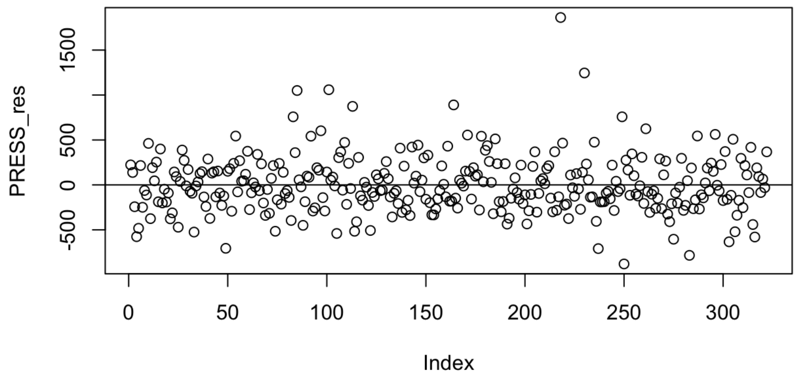


Figure 5: PRESS Residuals

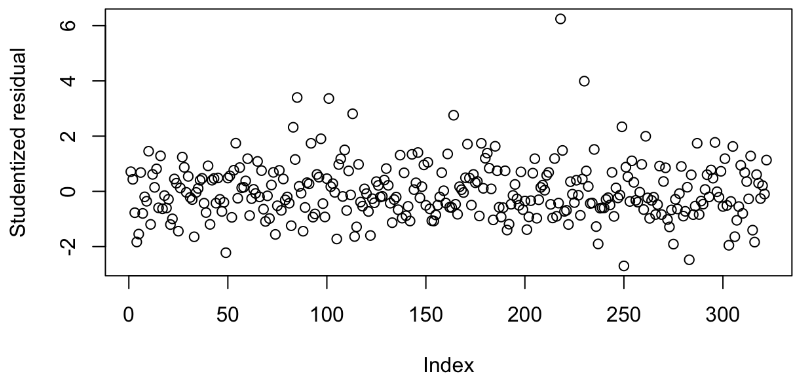


Figure 6: Student Residuals

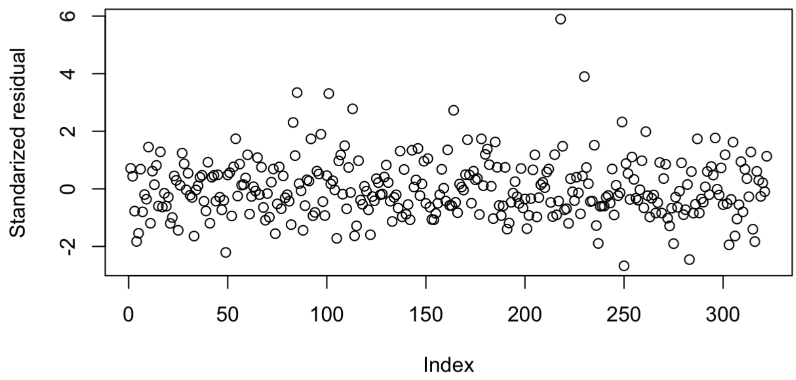


Figure 7: Standardized Residuals

Outlier Detection:

- Number of leverage points:
  - Using Hat values: 32
- Number of Influential Points:
  - Using Cook's Distance: 23
  - Using DFFITS: 26
  - Using DFBITS: 12
  - Uings COVRatio: 25

Checking Normality Using P-P and Q-Q Plot:

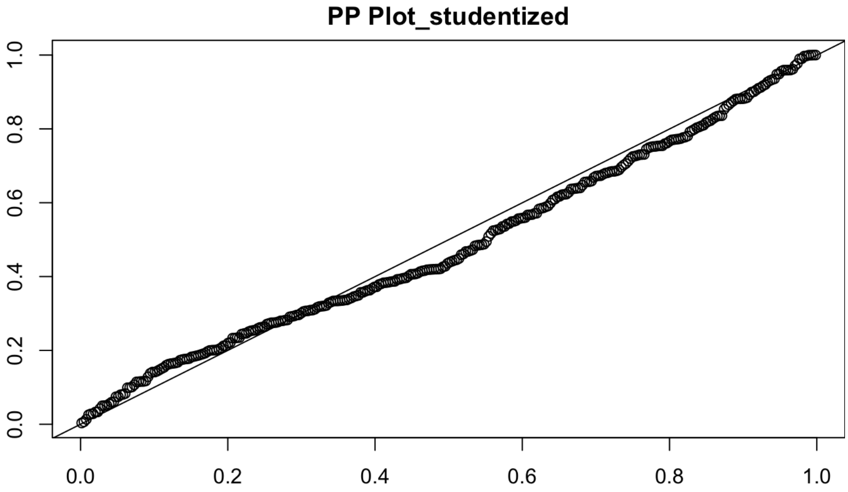


Figure 8: P-P Plot for Studentized Residuals

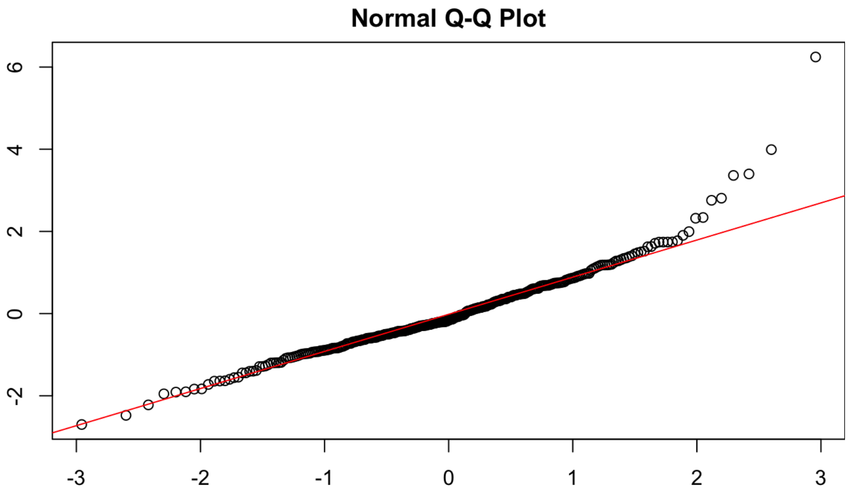


Figure 9: Q-Q Plot for Studentized Residuals

5 Box-Cox Transformation

A Box-Cox transformation was applied to the response variable to address heteroscedasticity and non-normality. A new model with the same variables used for best R^2 was fitted with the transformed variable.
Here, $Max_{salary}/Min_{salary} = 36.44$; hence, we can apply the Box-Cox transformation.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 915.12209   55.29661  16.549  < 2e-16 ***
AtBat        -1.41956    0.39145  -3.626 0.000336 ***
Hits          4.51200    1.21931   3.700 0.000255 ***
Walks         3.82889    1.18314   3.236 0.001342 **
Years         6.30416    8.48528   0.743 0.458073
CAtBat       -0.01237    0.04817  -0.257 0.797566
CRuns         0.66689    0.29961   2.226 0.026741 *
CRBI          0.14447    0.14270   1.012 0.312159
CWalks       -0.47890    0.18679  -2.564 0.010823 *
DivisionW   -75.57202   28.13517  -2.686 0.007620 **
PutOuts       0.14609    0.05396   2.708 0.007154 **
Assists       0.07633    0.12016   0.635 0.525753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 248.2 on 310 degrees of freedom
Multiple R-squared:  0.4229,    Adjusted R-squared:  0.4025
F-statistic: 20.66 on 11 and 310 DF,  p-value: < 2.2e-16
```

Figure 10: Summary after Using Box-Cox transformation with best lambda value possible.

5.1 Residual Plot for Transform Model

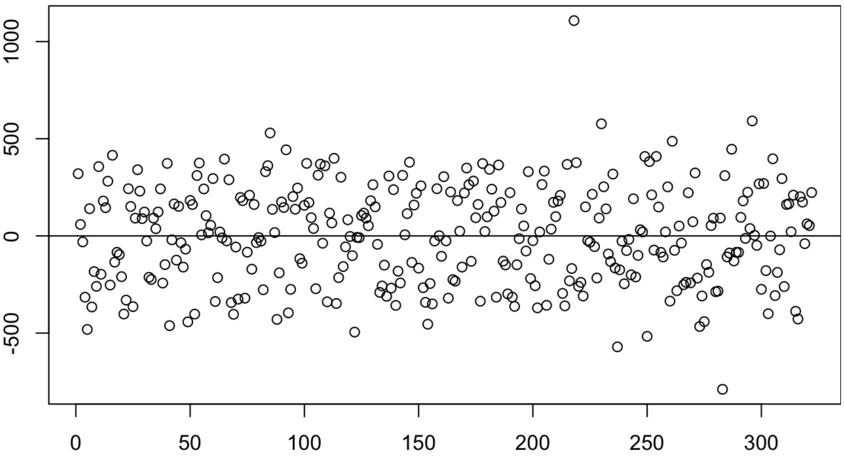


Figure 11: PRess Residuals after Transformation

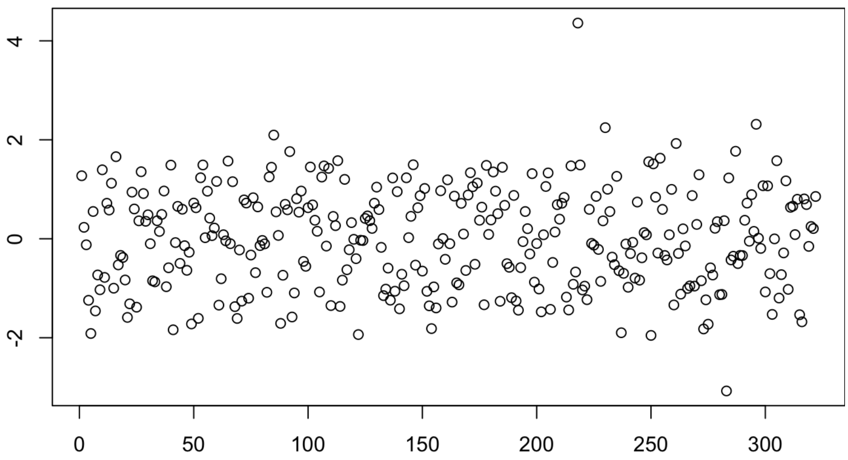


Figure 12: Standardized Residual after Transformation

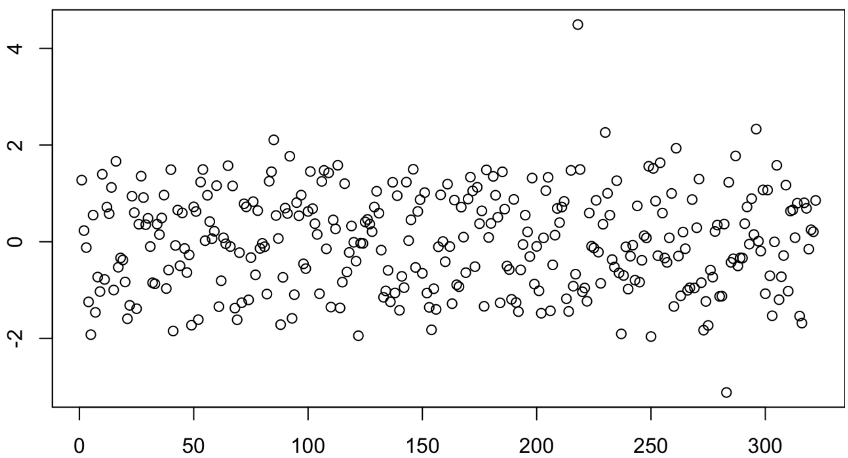


Figure 13: Studentized Residuals after Transformation

Outlier Detection after Transformation:

- Number of leverage points:
  - Using Hat values: 32
- Number of Influential Points:
  - Using Cook's Distance: 18
  - Using DFFITS: 9
  - Using DFBITS: 31
  - Uings COVRatio: 3

5.2 Diagnostic Plots for Transformed Model

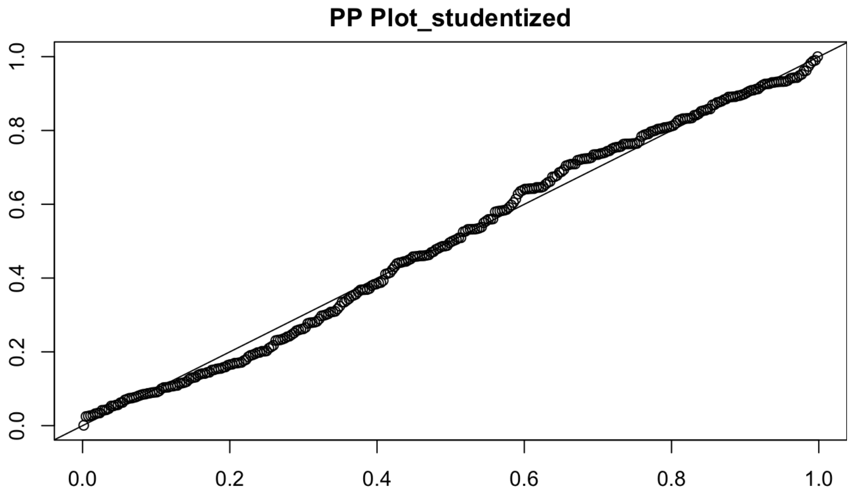


Figure 14: P-P Plot for Studentized Residuals

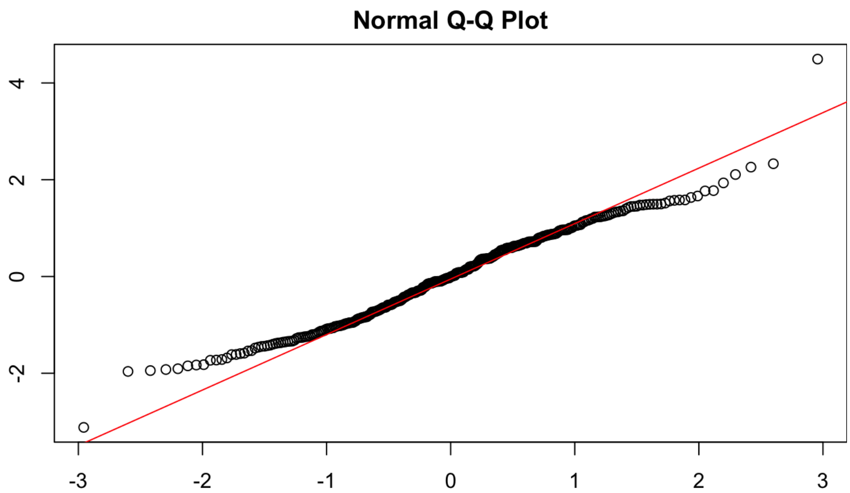
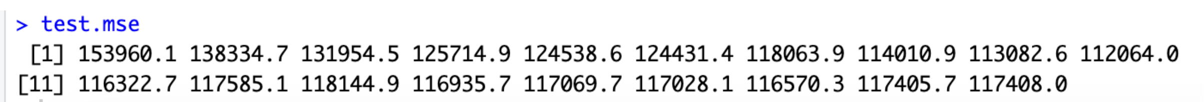


Figure 15: Q-Q Plot for Studentized Residuals

6 Model Validation Techniques

Two validation strategies were used:

- Validation Set Approach
  - Using a Random number of Rows to train the model with replacement.



```
> test.mse
 [1] 153960.1 138334.7 131954.5 125714.9 124538.6 124431.4 118063.9 114010.9 113082.6 112064.0
[11] 116322.7 117585.1 118144.9 116935.7 117069.7 117028.1 116570.3 117405.7 117408.0
```

Figure 16: Validation set Errors

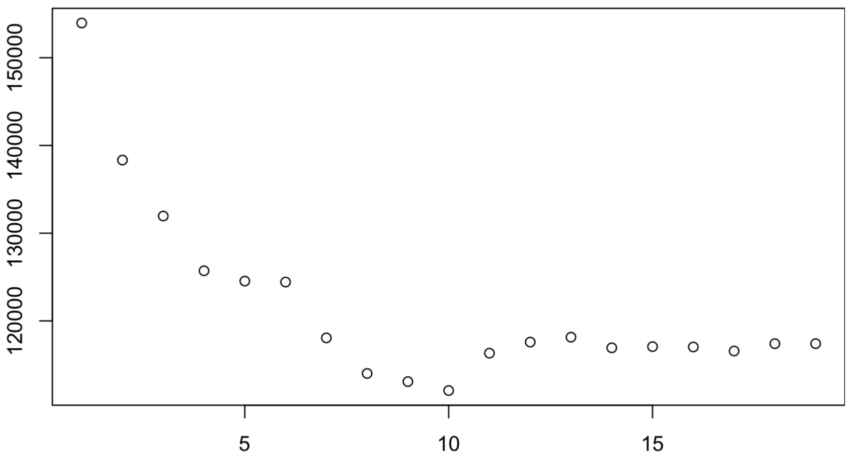


Figure 17: Plot for MSE'S

```
> coef(regfit.best, which.min(test.mse))##best model according to validation set
 (Intercept)        AtBat         Hits        HmRun         Runs        Walks       CAtBat
222.33719823  -1.90235109   5.82561733   4.10151984  -0.87690013   5.61937265  -0.06828204
       CRuns       CWalks    DivisionW      PutOuts
  1.48819046  -0.54089350 -74.22833789   0.18924357
```

Figure 18: Coefficients for min MSE model

- K-Fold Cross Validation
  - Using K=10 calculating MSE for different Folds.

```
> mean.cv.errors
       1        2        3        4        5        6        7        8        9       10       11
131199.9 125654.8 116774.8 122570.0 116363.3 116272.6 113422.8 105224.5 102470.7 104708.1 108192.1
      12       13       14       15       16       17       18       19
108036.9 107971.1 108245.1 108238.0 108316.2 108466.8 108617.9 108772.3
```

Figure 19: K-Fold CV Errors

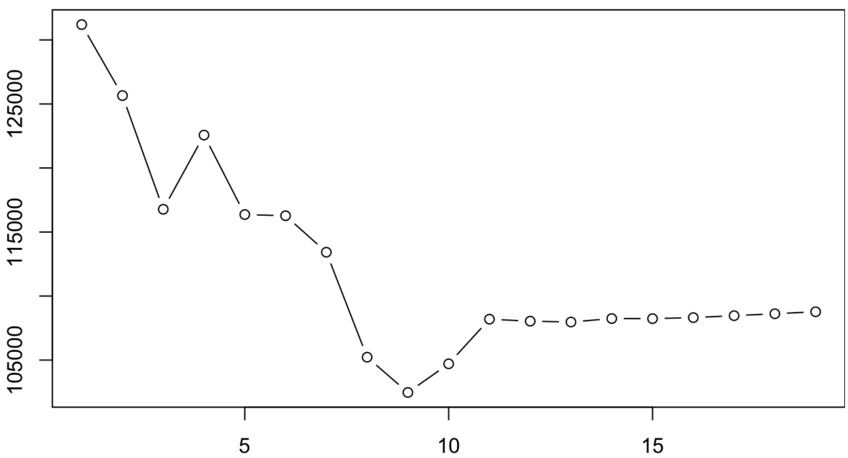


Figure 20: plot for K-Fold CV Errors

```
> coef(reg.best,9)
 (Intercept)        AtBat         Hits        Walks        Years        CRuns         CRBI
 351.6277855   -2.1080725    6.3756106    5.3536431  -19.2444179    0.8579652    0.4037357
      CWalks    DivisionW      PutOuts
  -0.6423395 -115.3383192    0.2222729
```

Figure 21: K-Fold CV Errors

7 Ridge and Lasso Regression

Both methods were applied using the `glmnet` package. Cross-validation was used to identify the optimal lambda.

- **Ridge Regression**
  - Created a Grid of 100 lambda values.
  - Made a model matrix of variables.
  - Extract the response variable.
  - Can access the lambda value of any grid and can find the coefficients corresponding to that lambda.
  - Apply Ridge regression for all grids.

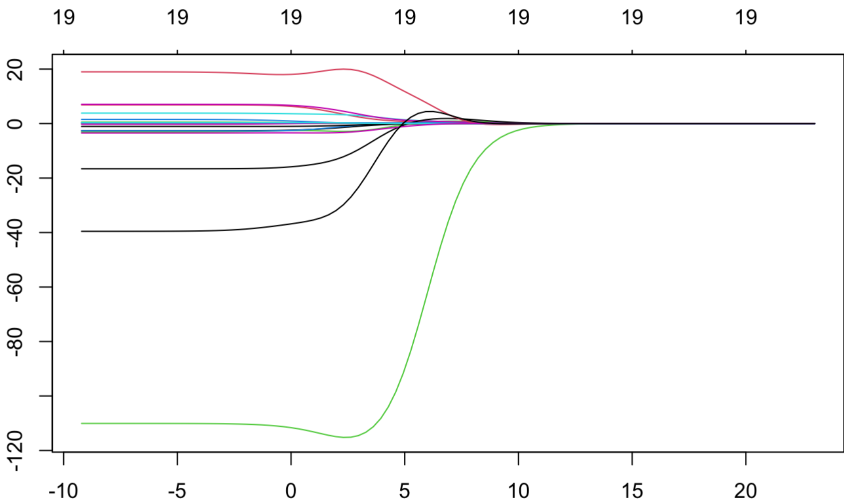


Figure 22: Ridge Regression for different lambda values.

  - train model with $\frac{2}{3}$ part of the data, and validate using remaining.
  - Now, you can choose the best lambda(16.516 here) using the K-Fold cross-validation error.
  - can find the test MSE for the best Lambda value(108001.4)

```
> predict(Final_ridge,type="coefficients",s=best.lambda)[1:19,]
  (Intercept)         AtBat          Hits         HmRun          Runs           RBI         Walks
 2.636245e+02 -7.873012e-01  2.169818e+00 -3.593636e-01  1.366907e+00  7.151689e-01  2.975442e+00
        Years        CAtBat         CHits        CHmRun         CRuns          CRBI        CWalks
-1.275852e+01  9.715029e-05  1.390914e-01  3.052874e-01  3.119987e-01  1.978353e-01 -2.328612e-01
      LeagueN     DivisionW       PutOuts       Assists        Errors
 2.353572e+01 -1.188440e+02  2.214349e-01  1.785392e-01 -4.513376e+00
```

Figure 23: Coefficients of the fitted model using best lambda.

- **Lasso Regression**
  -
  - Created a Grid of 100 lambda values.
  - Made a model matrix of variables.
  - Extract the response variable.
  - Can access the lambda value of any grid and can find the coefficients corresponding to that lambda.
  - Apply the Lasso regression model for all grids.

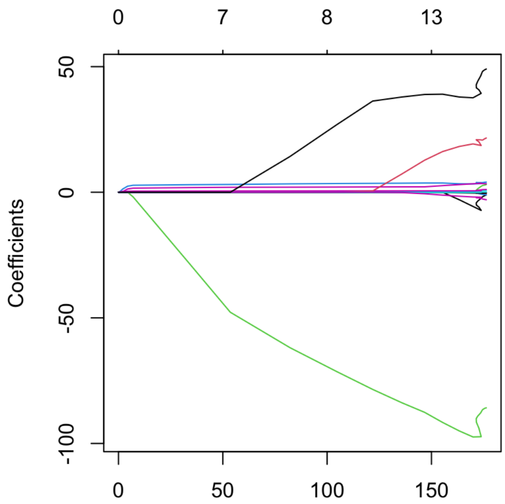


Figure 24: Lasso Regression For different lambda value

  - train model with $\frac{2}{3}$ part of the data, and validate using remaining.
  - Now, you can choose the best lambda(0.3797 here) using the K-Fold cross-validation error.
  - can find the test MSE for the best Lambda value(120074.2)

```
> coef(Full.mod)
20 x 1 sparse Matrix of class "dgCMatrix"
                         s0
(Intercept)  3.410486e+02
AtBat       -2.003889e+00
Hits         5.847382e+00
HmRun        2.843966e+00
Runs        -2.655359e-01
RBI          .
Walks        4.989414e+00
Years       -1.178115e+01
CAtBat      -1.025070e-01
CHits        1.740445e-01
CHmRun      -1.889164e-03
CRuns        1.088591e+00
CRBI         4.238682e-01
CWalks      -6.214703e-01
LeagueN      1.828939e+01
DivisionW   -1.112390e+02
PutOuts      2.355148e-01
Assists      3.505386e-01
Errors      -4.244631e+00
NewLeagueN   .
```

Figure 25: Ridge Regression Cross-Validation Curve

8 Elastic Net Regression

The elastic net was evaluated across different values of the mixing parameter α to balance between ridge and lasso.

- train model with $\frac{2}{3}$ part of the data, and validate using remaining.
- Use 5-fold cross-validation for training data.
- Make 10 models with different alpha values.

```
> All_mod
   alpha      mse model.name
1    0.0 147102.7    alpha 0
2    0.1 148466.8  alpha 0.1
3    0.2 142355.6  alpha 0.2
4    0.3 148566.3  alpha 0.3
5    0.4 139246.1  alpha 0.4
6    0.5 149092.7  alpha 0.5
7    0.6 149055.8  alpha 0.6
8    0.7 149058.7  alpha 0.7
9    0.8 149124.1  alpha 0.8
10   0.9 149067.1  alpha 0.9
11   1.0 139240.2    alpha 1
```

Figure 26: Elasticnet error for different alpha value

- min of all MSE is 139240.2

```
20 x 1 sparse Matrix of class "dgCMatrix"
                         s0
(Intercept)  322.30189271
AtBat         -1.61954275
Hits           4.65390233
HmRun          1.58338665
Runs           0.38040918
RBI            0.25809036
Walks          4.36190281
Years        -15.32574412
CAtBat        -0.02754951
CHits          0.12679515
CHmRun         0.15571591
CRuns          0.67680226
CRBI           0.30504707
CWalks        -0.47556230
LeagueN       15.56468381
DivisionW   -116.53230808
PutOuts        0.23000767
Assists        0.26517234
Errors        -4.24547639
NewLeagueN     .
>
```

Figure 27: Sparse Matrix

•

9 Logistic Regression

The binary classification was performed on the `NewLeague` variable using logistic regression and shrinkage methods.

- Confusion matrix to predict the player new league is

```
            Actual
Predicted    A   N
        A  166   9
        N   10 137
> mean(full.pred == Hitters_updated$NewLeague)
[1] 0.9409938
```

Figure 28: Confusion matrix and accuracy

- Train the model using 70% Data. and remaining for validation.
- the ROC Score for this model is

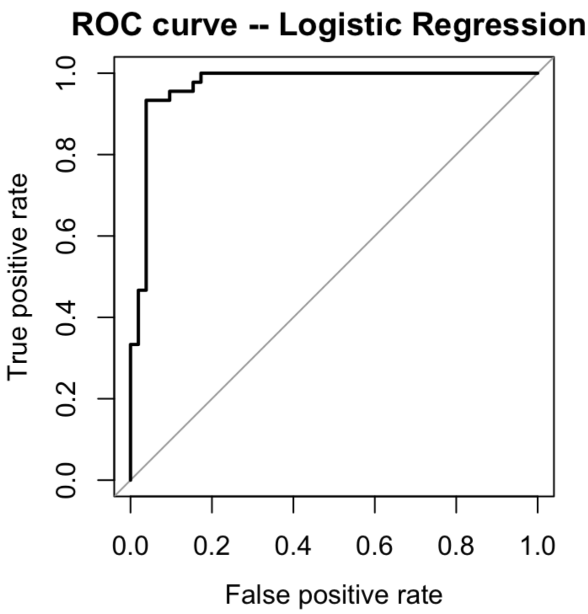


Figure 29: ROC value based on 70% training data

-

9.1 Shrinkage methods with Logistic regression

- **lasso with logistic:**
  - Trained with $\frac{2}{3}$ part of the data. and validate with the remaining.
  - Fit the model.
  - Chose the best lambda (Here 0.026)
  - fits the best model using this lambda.
  - Here, LeagueN alone is able to fit the model remaining coefficient=0
  - Fitted with accuracy 0.9636

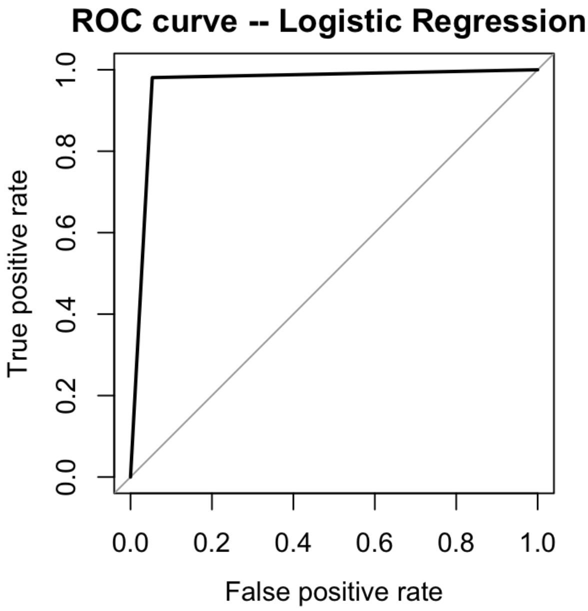


Figure 30: Poisson Coefficients by Hour

- * Also, while comparing with the full model accuracy remain same

- **Elasticnet with logistic**
  - Trained with $\frac{2}{3}$ part of the data. and validate with the remaining.
  - Fit the model using 5-fold cross-validation.
  - using best lambda and 10 different alpha fit the model to find minimum MSE.

```
> All_mod
   alpha       ROC model.name
1    0.0 0.9401445    alpha 0
2    0.1 0.9377365  alpha 0.1
3    0.2 0.9377365  alpha 0.2
4    0.3 0.9394565  alpha 0.3
5    0.4 0.9432405  alpha 0.4
6    0.5 0.9418645  alpha 0.5
7    0.6 0.9408325  alpha 0.6
8    0.7 0.9408325  alpha 0.7
9    0.8 0.9453044  alpha 0.8
10   0.9 0.9453044  alpha 0.9
11   1.0 0.9453044    alpha 1
```

- The maximum ROC value is 0.9453.
- final coefficient model

```
20 x 1 sparse Matrix of class "dgCMatrix"
                       s0
(Intercept) -2.6580351470
AtBat        0.0012081409
Hits         0.0002264729
HmRun       -0.0206052955
Runs         .
RBI          .
Walks        0.0019184155
Years        .
CAtBat       .
CHits        0.0001519110
CHmRun      -0.0007806069
CRuns        .
CRBI         .
CWalks      -0.0002575317
LeagueN      4.6163132078
DivisionW    .
PutOuts      0.0001005720
Assists      .
Errors      -0.0170933271
Salary       .
```

10 Poisson Regression

Poisson regression was used for count data modeling using the `AtBat` and responses.

- Fit a model to predict AtBat using some predictors of the data.

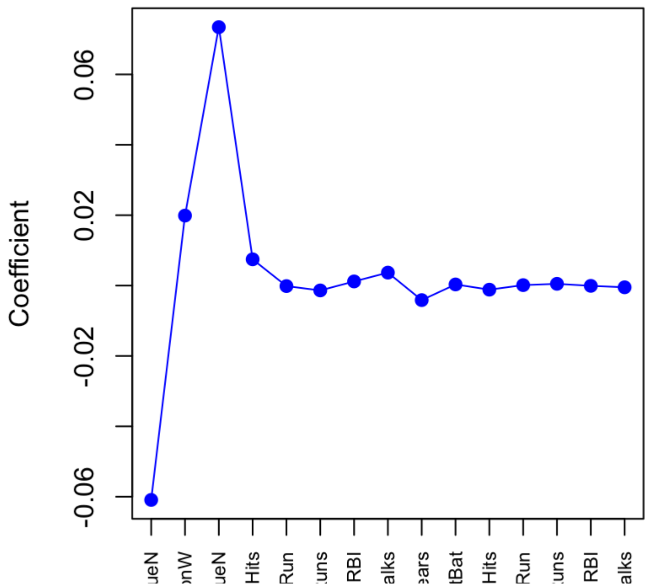


Figure 31: This Shows AtBat highest at NewleagueN

•

11 Conclusion

Conclusion

In this project, we explored various statistical modeling and machine learning techniques to analyze and predict baseball player salaries and league classifications using the `Hitters` dataset from the `ISLR2` package.

The key steps and findings are summarized below:

- **Data Cleaning:** Missing values were handled by replacing them with column-wise means to ensure a complete dataset for analysis.
- **Feature Selection:** We applied best subset selection, evaluated by metrics such as Adjusted R^2, C_p, AIC, and BIC, to identify the most relevant predictors of salary. The model with the highest Adjusted R^2 was chosen as the most optimal.
- **Model Diagnostics:** Various residual analysis methods, including PRESS statistics, studentized residuals, and leverage values, were used to evaluate model assumptions and identify outliers or influential observations.
- **Model Transformation:** A Box-Cox transformation was used to stabilize variance and improve model performance, followed by refitting a linear regression model on the transformed data.
- **Regularization Techniques:** Ridge regression, Lasso, and Elastic Net were implemented to reduce overfitting and enhance prediction accuracy. Among these,

Elastic Net (with $\alpha = 0.2$) produced the lowest test MSE, indicating superior predictive power.

- **Model Validation:** Both validation set and 10-fold cross-validation approaches were used to confirm model stability and generalization capability.
- **Logistic Regression:** For classification tasks (predicting `NewLeague`), logistic regression models were built and evaluated using ROC curves and AUC scores. The Lasso and Elastic Net models performed comparably well, with Elastic Net achieving the highest ROC score.
- **Poisson Regression:** We also applied Poisson regression to predict count-based variables like `AtBat`, identifying significant predictors such as `NewLeague`, `Hits`, and `Home Runs`.

Overall Conclusion: *By combining rigorous variable selection methods with model diagnostics, transformations, and regularization techniques, we successfully built robust models for both regression and classification tasks. These models not only fit the training data well but also generalize effectively to unseen data, making them valuable tools for data-driven decision-making in sports analytics.*