INTRODUCTION

Understanding the underlying driving force behind "why sportsmen make what they make" is critical as it allows sports clubs to optimise their operation, as accurate offering of salaries will help clubs to ensure that they secure talented players in the most cost-efficient manner. This report is to analyse, and to draw inference about the factors affecting the salaries of major league baseball players' in 1987 in the United States. In order to allow for more flexibility while maintaining interpretability, we will be focusing on fitting the data to various form of generalised linear models. Ultimately, we will draw comparison between the candidate model and interpret the result.

DATA COLLECTION AND AGGREGATION

The relevant data set is assessed at: http://stat-computing.org/dataexpo/1988.html. The 1988 Exposition's data was collected by Lorraine Denby, which contains information about 1986 salaries and statistics of North American Major League Baseball players. The data is downloaded and saved as csv file, as there were also a set of published corrections to be made to the original dataset. With necessary manipulations, the information and statistics about the hitter and pitcher players, as well as the teams in both major baseball leagues are prepared in three separate spreadsheets, namely **pitcher**, **hitter** and **team**. It is worthwhile to note that the observations in the pitcher and hitter dataset are mutually exclusive, and the two data set share minimal number of predictors. Therefore, we will investigate the factor affecting the response variable, **salary1987** for hitter and pitcher in parallel to each other.

DATA EXPLORATION

MISSING VALUES

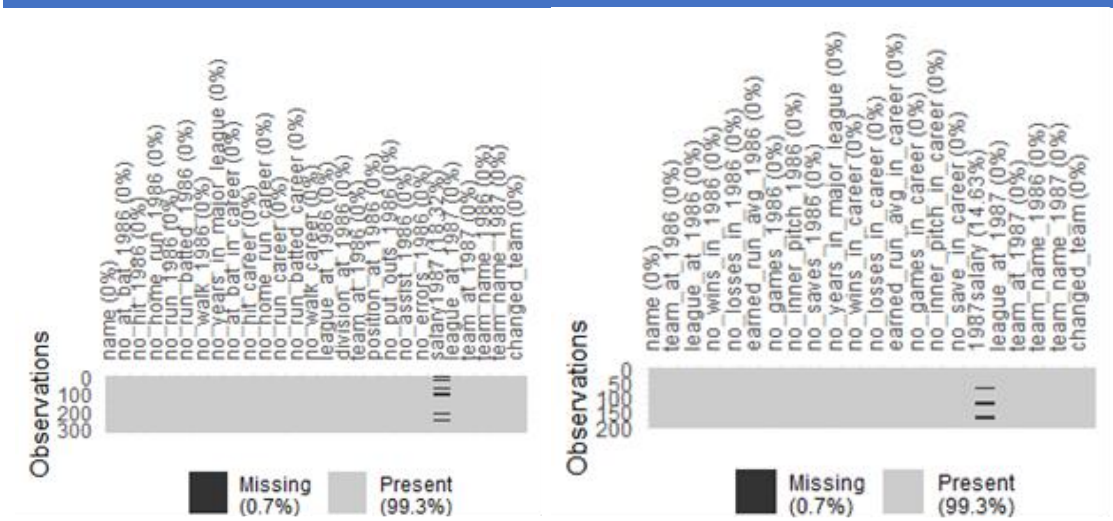

Figure 1: Left – missing value in hitter; Right – missing values in pitcher dataset

The first step we've taken is to understand the structure and distribution of missing values in the data set. As seen in Figure 1, missing values exist only in response variable that we are interested in modelling, instead of the design matrix, deletion method is taken to avoid introducing bias.

DISTRIBUTION OF RESPONSE VARIABLE

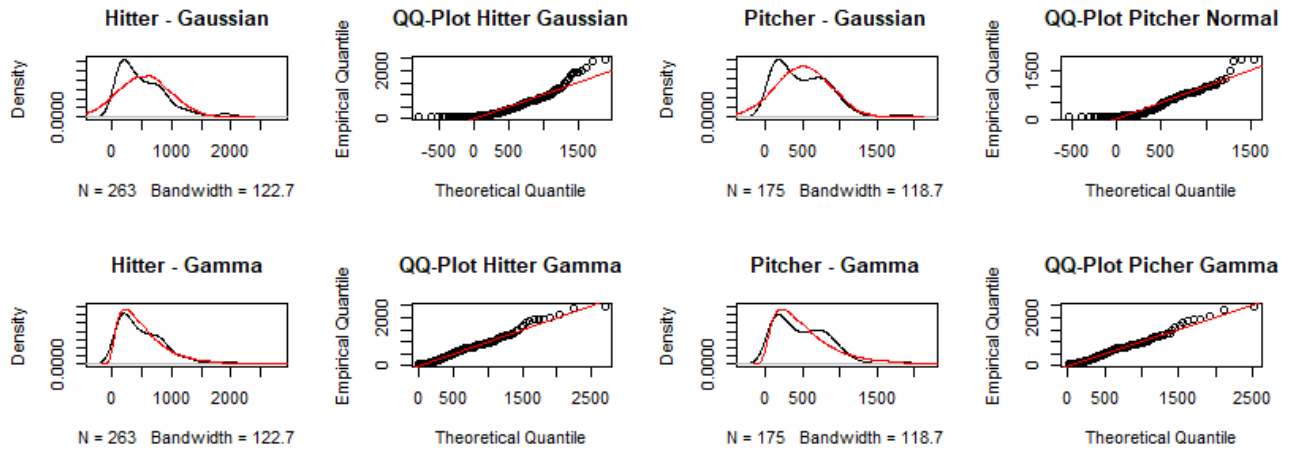

Figure 2 : Density plots and qq-plots of hitter and pitcher against fitted gaussian distribution and gamma distribution respectively

	Hitter - Gaussian	Pitcher – Gaussian	Hitter – Gamma	Pitcher - Gamma
AD Test	0.0094	0.0133	0.1816	0.0509
CvM Test	0.0157	0.0379	0.2617	0.0759

Table 1 : Goodness of fit tests

As seen in Figure 2, firstly the response is right-skewed and positively distributed in both hitters and pitchers. Such that we may wish to consider the use of link functions that will ensure the positively of fitted response, or alternative families of GLM. As we expected, the empirically gamma distribution fits both hitter and pitcher salary data better. Furthermore, at 5% significance, AD test and CvM test results align to reject the null hypothesis that the response are normally distributed; and fails to reject the null hypothesis that the response is gamma distributed. Therefore, we will explore gamma family GLM in addition to transformation of gaussian family.

VARIABLE CORRELATION

HITTER DATA

By plotting a correlation plot amongst all numeric variables (see Figure3), we noticed that there are three clusters of predictors that have high correlations pairwise internally, these groups can be roughly divided into personal performance in 1986 (top left), person performance in career (centre) and team status (bottom right). As there are very little correlation inter-group, analysis should be conducted individually for each cluster.

Firstly, we notice that among the centre cluster of career performance, almost all predictors are heavily correlated with the variable number of years in major league. As such, we divided every statistic about a player's career by the number of years in major league and we notice that the first two clusters forms a larger cluster and all individual performance related predictors are somewhat correlated (see Figure 3). It is important to note that, from this point, all careers related statistics refers to the annual average statistics in the player's career.
Furthermore, strong correlation pairwise between number of runs, number of hits and number at bats; across both 1986 performance and career performance clusters. As the number of runs relates most closely with scoring and hence winning games, we will keep number of runs only.

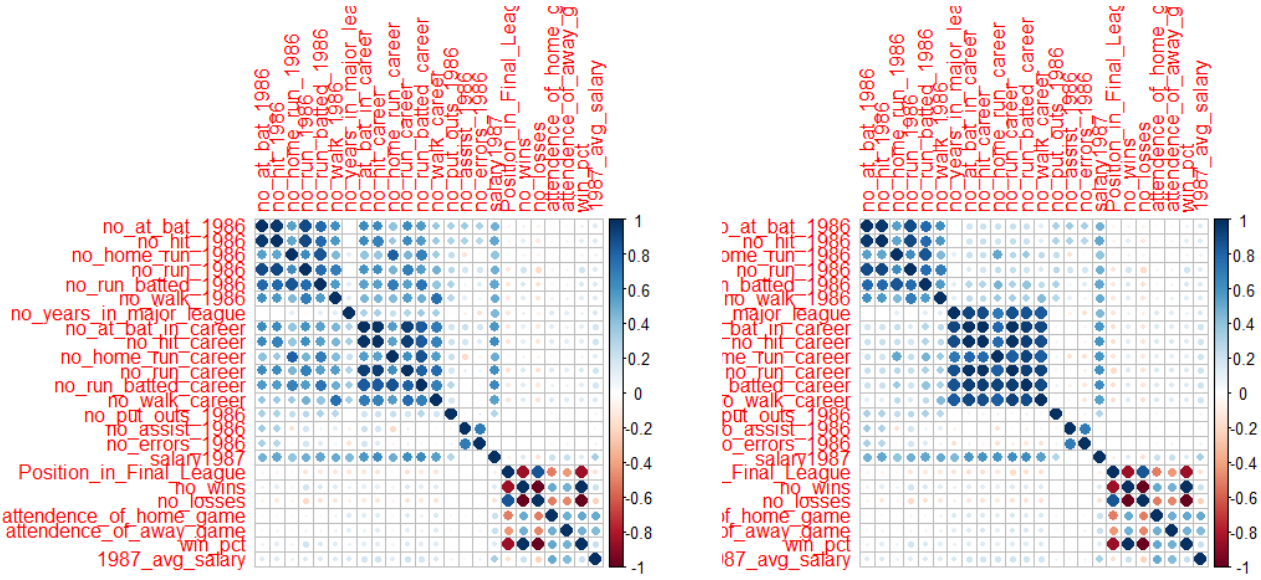

Figure 3 : Left - correlation plot before finding average career statistics; Right - correlation plot after finding career averages

Finally, amongst the bottom right cluster, win percentage is strongly correlated with number of games won positively, and with number of games lost negatively. Therefore, it is reasonable to remove the latter two predictors from the model. Furthermore, a clear connection between win percentage and ranking of team by league, by division can be also noted (see Figure 4); and intuitively, by default rankings in sports league are decided primarily based on win rates. Hence it is also safe to remove the ranking predictor.

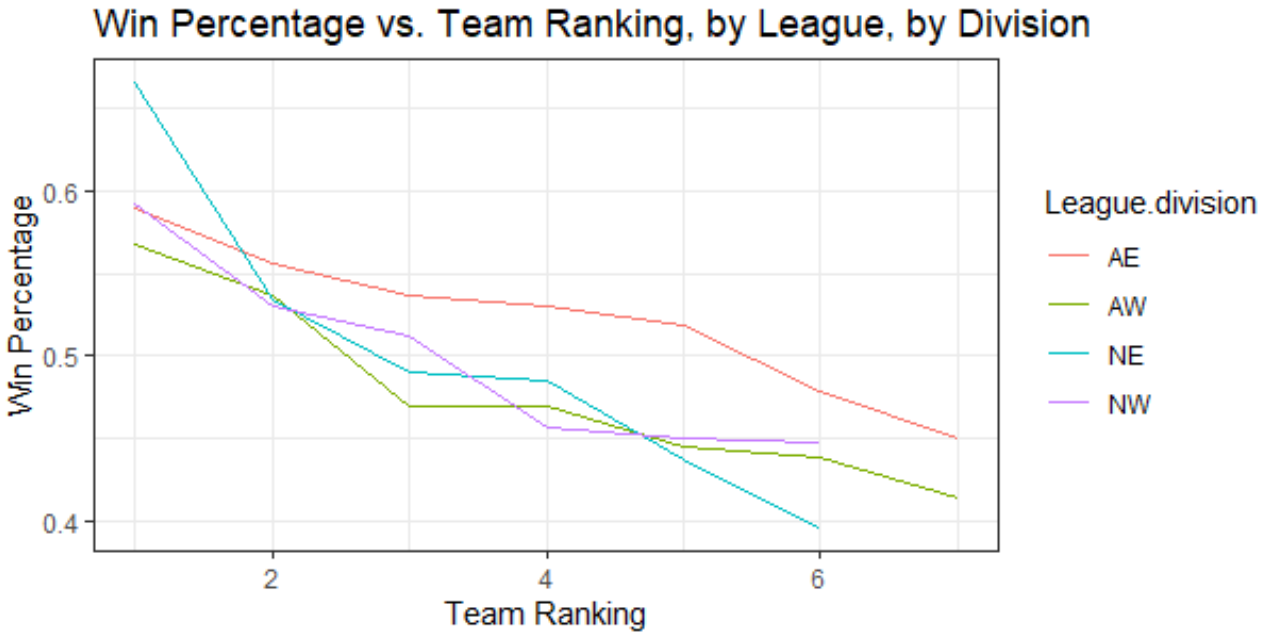

Figure 4: Win Percentage vs. Ranking

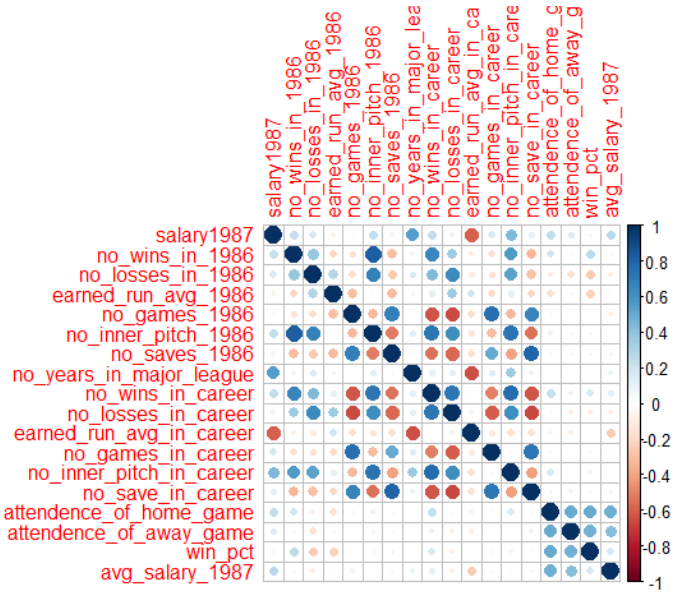

Figure 5: Correlation for pitchers

PITCHER DATA

Similar observation has been made on the pitcher data set that number of years in leagues is highly correlated with careers basis statistics, and we have divided each of these variables by the number of years in league to obtain annual average.

OTHER OBSERVATION AND CORRESPONDING DATA MANIPULATION

TEAMS

It is observed that both major leagues have teams in Chicago and New York. This may lead to issues in merging data, therefore new variables are created by catenating the variables League and team to produce unique team codes. Furthermore, a new categorical predictor has been generated to indicate whether the player has changed team between 1986 and 1987, for both hitter and pitcher.

POSITION

A variable indicating the player's position defensively can be found in the hitter dataset. Upon failing to recognise any significant pattern (see Figure 6), we have created some new variables to capture the information carried in this categorical predictor. Namely we explored whether the player has one defensive position only, whether the player is dedicated hitter and whether the player's position is primarily responsible for defending the bases.

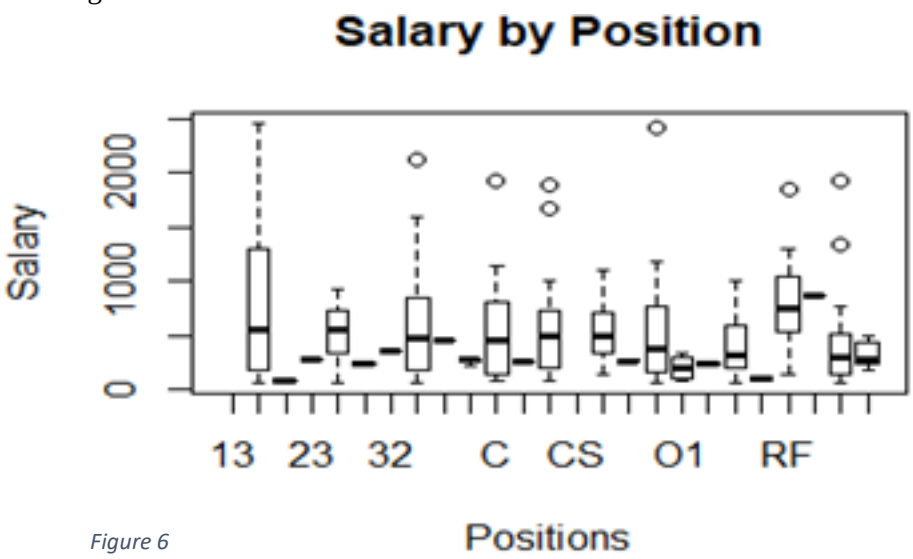

Figure 6

Model Choosing and Fitting (hitters)

As discussed previously in the Data Exploration Section, we observe that the distribution of response variable, salary1987, is positive and right skewed, recall the exploratory analysis in the preliminary data exploration we have found that one cannot reject the null hypothesis that the response is gamma distributed by both Kolmogorov–Smirnov test and Cramér–von Mises Test. Given this assumption of the data generating process, it is reasonable to explore the possibility of fitting the response variable to selected predictors with a gamma family GLM (generalised linear model).

Noted the assumptions for GLM are listed below:

- The data response variable vectors are independently distributed. Errors need to be independent
- The data response variable vectors are independently distributed. Errors need to be independent
- Random component: Response variable are independent and belongs to the exponential family, $E[Y] = \mu$.
- Systematic Component: Covariates are combined into a linear predictor, η.
- Link function g(.) between random and systematic component, $\eta_i = g(\mu_i)$.

GLM WITH GAUSSIAN FAMILY AND IDENTITY LINK

All experiments are performed using the R tool. GLM with gaussian family and identity link, which generally referred as linear regression, is fitted to hitter data to generate the full model with AIC 3749.7. [Appendix 1] The summary output suggests only 6 predictors are considered significant, therefore, to eliminate unnecessary affect from insignificant features, Backward selection method is used to select the optimal model with lowest AIC value. For hitters' salaries in 1987, the optimal model after selection contains 13 predictors with AIC 3741.4 which has no significant difference with full model, and now 10 of the features considered significant.

According to the result of ANOVA Chi-square test in R for the nested models [Appendix], p-value is 0.9046 with deviance increase from 20,576,958 in full model to 20,709,503 in the sub-model, which indicate the sub-model provides better fit than full model.

DIAGNOSTIC

The R^2 statistic measures how much the model has improved by predictors and is calculated according to $R^2 = 1 - \frac{D}{D_0}$, where D denote the model deviance and D_0 denote the null deviance. Comparing to the null model, the R^2 is equal to 0.61 which means the sub-model is capable in explaining 61% of null deviance.

From the residual plot (figure 7), there is a pattern of residuals and cluster of observations are concentrated on the left tail. The variance of residuals is clearly a non-constant and the mean of residuals is also not constant at zero, which indicates the violation of assumptions. The Q-Q plot shows severe deviations and some outliers on both tails. In conclusion, the generalized linear model with gaussian family and identity link may not provide an appropriate fit to the hitters' salaries in 1987.

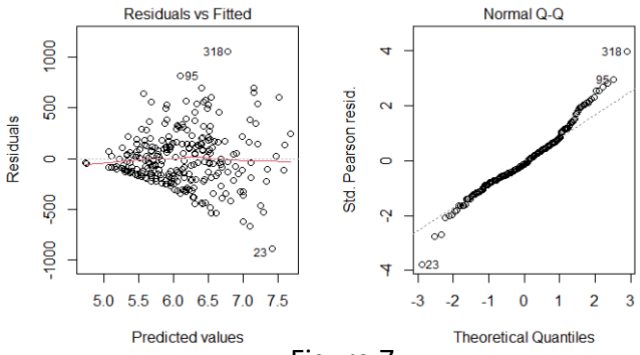

Figure 7

GLM WITH GAUSSIAN FAMILY AND LOG LINK

As stated in the preliminary analysis, we now fit the GLM with gaussian family and log link to the hitters' salaries. Similar procedure as GLM was gaussian family, full model presents that among 20 predictors, only 9 features are significant, whereas under backward selection method starting from model with all features, the

result in [Appendix 2] suggests 12 predictors are kept, also the AIC has drop from 3727.7 in full model to 3721.1 which has no significant decrease from the full model.

Using ANOVA Chi-square test in [Appendix 2], p-value is 0.5361 with a deviance reduction from backward model to full model of 392,463, which states the failure of rejecting null hypothesis. From table 2, the R^2 statistic for log link model is equal to 0.636 which indicates 63% of deviance is explained by the predictors.

Table 2	Degree of Freedom	Deviance
Null Model	262	53,096,433
Backward Model	250	19,316,564

DIAGNOSTIC

The residual plot (figure8) shows a clear pattern of the residuals. The mean of residuals is constant at zero which satisfy the assumption, but the variance of residuals is increasing with the predicted values. From the Q-Q plot, there is a slight deviation on the left tail and strong deviations on the right tail, which means the model may fit the salaries of hitters in 1987 well for the lower range.

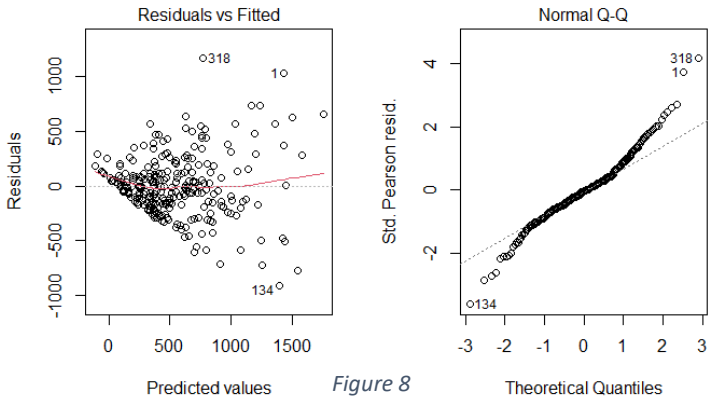

Figure 8

Comparing with the generalized linear model with identity link, the generalized linear model with log link provides a fit with lower AIC and less deviations on the low range of salaries. In addition, the mean of residuals is constant and the cluster of observations in residual plot is shifted to the right and not concentrated on the left part that much.

GLM WITH GAMMA FAMILY AND LOG LINK

With a similar argument in the gaussian with log link section, we will consider the use of log link to fit the Gamma GLM with Log link. By naïvely fitting the model with a formula that incorporates all predictors in the cleaned dataset, we obtain a full model with 20 predictors, amongst which only 6 are significant by Wald's test at 5% significance level [Appendix 3].

We then employed stepwise backward selection with AIC criterion to find an appropriate nested model. By doing so we reduce the size of the model to 10 predictors, where 7 are significant and the AIC is reduced from 3661.4 to 3647.1. [Appendix 3] To compare the full model and the nested model, we investigated the ANOVA with Chi-square test. [Appendix 3] The resultant p-value of 0.7967 indicates that the additional predictors in the full model does not reduce deviance significant when compared to the sub-model selected, therefore we will continue to discuss with sub-model.

DIAGNOSTIC

Table 3	Degree of Freedom	Deviance
Null Model	262	184.439
Backward Model	252	95.837

We will then diagnosis the chosen sub-model by examining the model's reduction in deviance relative to the null model, deviance residuals visualisation and variance inflation factors (VIF). From table 3 we observe that the sub-model is able to reduce the deviance by 88.6, in total explaining 49% of the null deviance.

Furthermore, from figure 9, we observe that there is no obvious pattern in the deviance residuals vs. linear predictor plot, which supports the gamma family assumption. However, in the QQ-plot (see figure 10) of standardised deviance residuals against standard normal distribution indicates that it fits poorly on one tail. Finally, the VIF have been computed for each of the

predictors in this model and we observe that there exist some variables with VIF > 10, indicating the existence multi-collinearity.

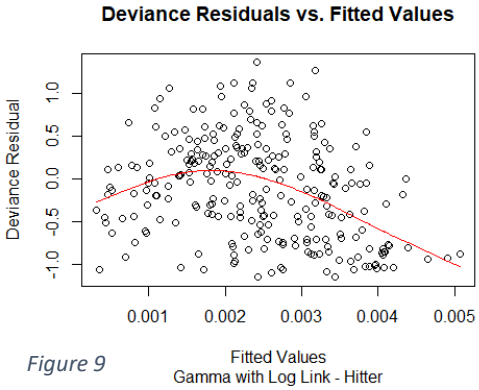

Figure 9

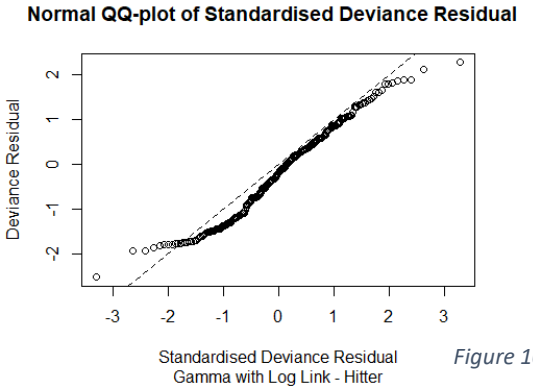

Figure 10

PCA

Considering both the severe correlation identified in the data exploration section, and VIF result in the previous model, we will also explore the possibility of utilising Principal Component Analysis (PCA) to eliminate multi-collinearity. Since there exist a cluster of predictors related to the performance statistics of the players that have very high correlation amongst themselves pairwise, we will apply PCA to only this subset of predictors.

Result of the PCA in figure 11 [Appendix 3] indicates that the first 4 principal components is sufficient to explain over 95% of the variances. Hence, we will replace all performance statistics in the original training dataset with the first 4 rotated principal components before proceeding to model fitting, such performance statistics including the number of at bat, the number home run or the number batted in both 1986 only and in the career.

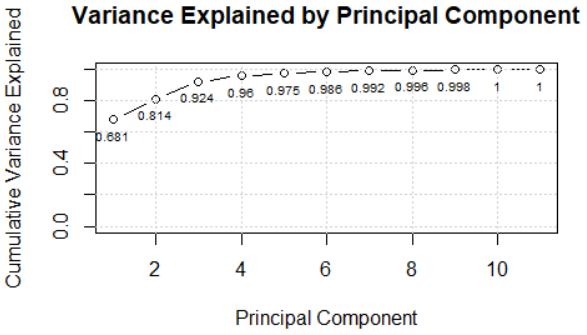

Figure 11

Upon conducting dimension reduction on the training dataset, we will again fit the Gamma family GLM with Log link with a procedure similar to above.
The full model consists of 21 predictors and 6 of which are significant; whereas the backward selected sub-model contains only 5 predictors and only one are insignificant. In comparison, the selected sub-model reduces the AIC from 3628.5 to 3616.6, and again, ANOVA with Chi-squared test indicates that the additional variables in the full models do not meaningfully reduce deviance, with a p-value of 0.7967. Hence the selected sub-model is preferred.

DIAGNOSTIC

Upon diagnosis of the deviance of the more optimal sub-model with partial PCA, we observe that the selected model is able to reduce the deviance from the null of 184.44 to 89.02, even lower than normal backward model deviance of 95.837, PCA sub-model effectively reducing null deviance by 52%. Considering this model shares same null deviance as the previous model with the same family and link, it can be argued that the model with partial PCA dimension reduction fits the data better.

Furthermore, it can be seen from the deviance residuals vs. fitted value plot of this model in figure 12 that there is almost no pattern in the residuals at all. This supports our assumption about data generating process and the application of gamma regression with dimension reduction. However, the Normal QQ-plot in figure 13 of standardised deviance residuals indicates that the standardised deviance residuals is systematically left

skewed. Lastly, VIF are computed for this model and all values are small, indication that there is minimal multicollinearity.

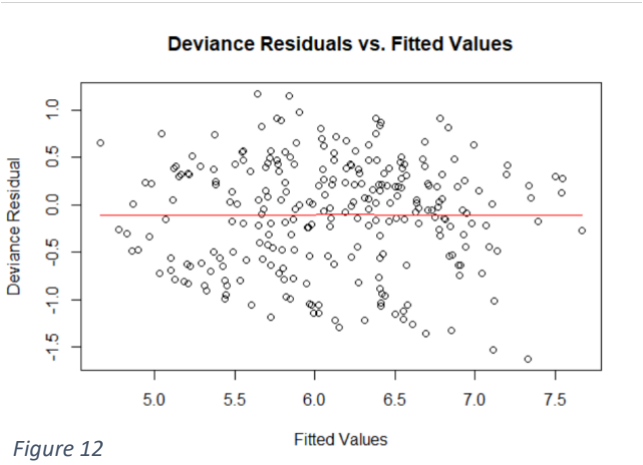

Figure 12

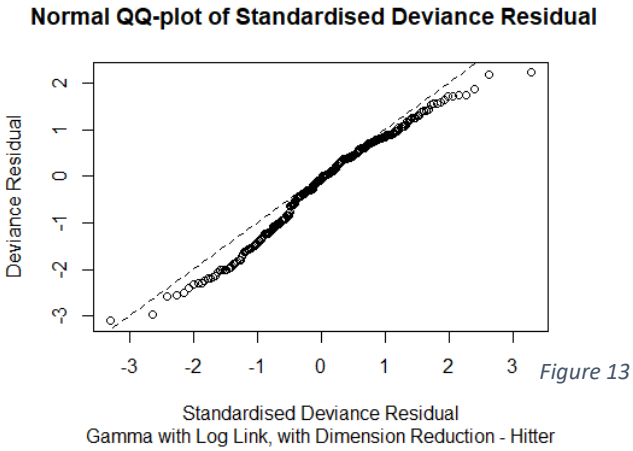

Figure 13

Model Choosing and Fitting (Pitchers)

After finishing the exploration and modelling analysis for hitter salary in 1987 in baseball team, now we will look at how the pitcher in baseball team be affected in 1987. As stated in preliminary data exploration, some variable has been altered to better suit for further analysis, including taking average performance statistics in career, also taking winning percentage instead of number of winning and losing game for team.

Implied by pitcher salary distribution analysis in preliminary exploration, it is reasonable to believe pitcher salary in 1987 follows either Gaussian Distribution with Log scale, or Gamma Distribution. These two distributions compromise the heavy left skewed and non-negative response issues for pitcher salary, therefore GLM is used as selected model for fitting.

GLM WITH GAUSSIAN FAMILY AND LOG LINK

GLM with Gamma log link will be fitted. Based on R computation, 10 predictors out of 20 presents a significant nature in predicting pitcher salary in 1987 [Appendix 2.1]. As such low proportion of significant predictors, Stepwise Selection for features is used based on the principle of minimisation of AIC. Referred to [Appendix 2.1], only 7 predictors were selected, with the residual deviance further decrease by 1.18.
Under sequential test based on chi-squared distribution [Appendix 2.1] by ANOVA, the p-value of 0.84 is far from the significant level, reveals it is statistical viable to eliminate half of the predictors with increase of small proportion of deviance. Therefore, forward model of Gamma GLM with log link will be discussed onward.

Table 4

	Degree of Freedom	Deviance
Null Model	174	114.371
Purposed Model	154	42.234
Forward Model	167	43.412

DIAGNOSITIC

From the above table 14, the purposed model has explained 38% deviance of the null deviance. As discussing the correspondence of constant error rate assumption, deviance residual vs fitted value plot is required to check such behaviours. Despite a slight decrease trend as fitted value increase, apparently most points randomly scatted on the plot. Reinforced by figure 15, forward Gamma GLM have the standardised deviance

residual fit perfect for normal distribution, suggests it has followed the assumption for normally distributed error rate.

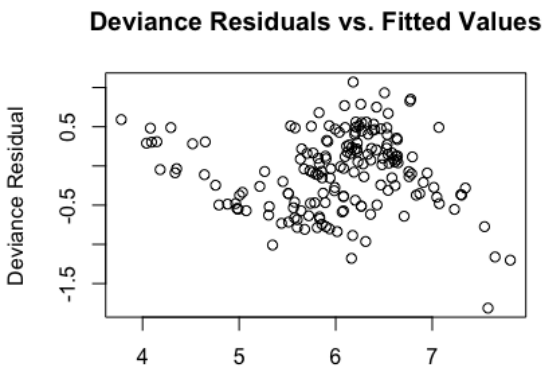

Figure 14

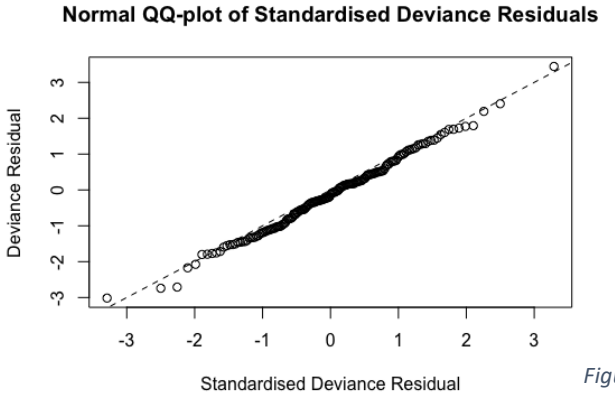

Figure 15

Model Assessment

One method of Model Assessment is to split the dataset into training set and test set in the first place and consider the test MSE as the criterion for accuracy of the model. This is because fitting data the training set inevitably causes overfitting to some degree and using test data unseen to the model in the training process can effectively expose the shortfalls should the model over fit. However, in our case, given the small size of the dataset, it is unfavourable to split the data even further. Thus alternative ways to simulate the test error, such as cross validation can be taken. One can either consider the use of Leave One Out Cross Validation (LOOCV) or K-fold Cross validation. Given that we are utilising small dataset, LOOCV may be preferred as it utilises as much data points as possible, at the same time it is not prone to variation.

In the assessment of hitter model, we are able to obtain the following result:

	Gaussian Identity link	Gaussian log link
MSE	94331.23	91619.30
Adjusted CV estimate	94299.48	91579.97

Such result implies both MSE and adjusted CV estimate is lower for gaussian log link GLM, on the other hand, we a minimum value of MSE in the Gaussian log link GLM, hence we can assess the performance of these two models, concludes Gaussian Log link had a better performance and vice versa.

Furthermore, to assess the accuracy the estimate, simulation method can be taken as opposed to reading the standard errors from summary() as summary function tend to have optimistic bias. This can be done by resampling the dataset a large number of times via bootstrapping and obtain coefficient estimate with each re-generated sample, and finally constructing confidence interval and compute standard errors empirically from the generated coefficients. The more the standard errors of estimates deteriorate compared to the output give in summary function, the more likely that there may be problems with the model's assumption or accuracy.

For example, for the bootstrap of Stepwise Gamma GLM, the result of 1000 bootstrap [Appendix 2.2] suggests estimator is consistent with the predictor estimate from the summary output, also we notice the difference of standard error from bootstrap and standard error from summary output generally not exceed 10%. This implies a good correspondence between bootstrap estimate and standard estimate in Gamma GLM, implies that this model is not over optimistic about the performance.

Limitation

Limitation can be discussed in two section, first section will be the limitation of the model selected, where the second section will be based on the diagnostic method.

The limitation of model selected, in our case will be GLM with various of family and link, also in the hitter salary analysis, we apply Principal Component Analysis (PCA), therefore the limitation can be including:

- Strongly rely on assumptions. GLM assumes the independence of variables, but the variables in data cannot be completely independent. For example, the factor of team winning percentage is highly correlated with the other feature of personal wining rate, as team with stronger economical background might be more attractive to those well-performance hitters/pitchers, which makes these two features positive correlated. The high correlated covariates can make the model unstable and less valuable.

- The response variable may not belong to an exponential distributed family. In our assumption of data generating process, exponential distributed family is the basis of building GLM, however we are perfectly not confidence with the coincidence of response and gamma or gaussian distribution, which this will leads to the unexpected error and excessive residual.

- Less interpretability due to the different link functions. Unlike linear model, the use of various link function did allow better adaption to different response, however it is hard to intuitively interpret how features affecting response and to which extent.

- PCA makes the independent variables less interpretable. The algorithm of PCA is to compress high dimension feature to principal component, therefore some independent variable becomes part of the linear combination of principal component, which extremely difficult to tell how that particular variable influences the response.

- PCA also leads to data missing, as in our case, we select the 4 first principal component which explain 95% of variance, however we still miss out the remaining 5% explanation, and that is the trade-off between high dimension and dimension reduction.

The limitation for diagnostic method, in our cases cross validation and bootstrap can be including:

- Highly time and cost consuming. Use of such method require continued looping to randomly choose training and testing set, which definitely cannot be done manually, even though with the help of computing, it is still costly for processing large dimension dataset.

TECHNICAL APPENDIX 1.1 (HITTERS)

GLM Gaussian family with Log link full model

```
glm(formula = salary1987 ~ ., data = hitter_data_mod1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-905.8  -159.0    -9.3   132.0  1130.0

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.71e+02   2.24e+02   -2.10   0.0365 *
league_at_1987N          1.11e+01   3.86e+01    0.29   0.7735
no_home_run_1986        -8.28e+00   5.04e+00   -1.64   0.1014
no_run_1986              2.57e+00   1.65e+00    1.56   0.1206
no_run_batted_1986       2.91e+00   2.00e+00    1.45   0.1472
no_walk_1986             3.80e+00   1.58e+00    2.41   0.0166 *
no_years_in_major_league -5.79e+00   9.85e+00   -0.59   0.5570
no_home_run_career        5.66e-01   8.21e-01    0.69   0.4908
no_run_career             6.38e-01   2.74e-01    2.33   0.0209 *
no_run_batted_career      3.41e-01   3.58e-01    0.95   0.3412
no_walk_career           -5.02e-01   2.46e-01   -2.04   0.0426 *
no_put_outs_1986          2.26e-01   7.09e-02    3.18   0.0017 **
no_assist_1986            5.74e-02   1.92e-01    0.30   0.7657
no_errors_1986           -5.22e+00   3.99e+00   -1.31   0.1921
changed_teamTRUE          1.02e+02   4.96e+01    2.06   0.0400 *
attendence_of_home_game   2.93e-05   4.46e-05    0.66   0.5126
attendence_of_away_game   1.75e-04   1.35e-04    1.29   0.1985
win_pct                  -5.60e+02   3.76e+02   -1.49   0.1373
avg_salary_1987           5.72e-01   2.01e-01    2.85   0.0047 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 84332)

    Null deviance: 53096433  on 262  degrees of freedom
Residual deviance: 20576958  on 244  degrees of freedom
```

GLM Gaussian family with Log link Backward model

```
glm(formula = salary1987 ~ no_home_run_1986 + no_run_1986 + no_run_batted_1986 +
    no_walk_1986 + no_home_run_career + no_run_career + no_walk_career +
    no_put_outs_1986 + no_errors_1986 + changed_team + attendence_of_away_game +
    win_pct + avg_salary_1987, data = hitter_data_mod1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-907.67  -159.77   -10.63   119.80  1165.33

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.547e+02  2.042e+02  -2.717 0.007058 **
no_home_run_1986        -1.039e+01  4.503e+00  -2.307 0.021899 *
no_run_1986              2.297e+00  1.510e+00   1.521 0.129537
no_run_batted_1986       3.766e+00  1.763e+00   2.136 0.033672 *
no_walk_1986             3.822e+00  1.550e+00   2.465 0.014379 *
no_home_run_career       1.182e+00  4.911e-01   2.406 0.016836 *
no_run_career            7.722e-01  1.867e-01   4.137 4.81e-05 ***
no_walk_career          -5.304e-01  2.404e-01  -2.206 0.028273 *
no_put_outs_1986         2.378e-01  6.822e-02   3.485 0.000581 ***
no_errors_1986          -4.191e+00  2.934e+00  -1.429 0.154397
changed_teamTRUE         1.058e+02  4.875e+01   2.171 0.030894 *
attendence_of_away_game  2.099e-04  1.293e-04   1.623 0.105955
win_pct                 -4.831e+02  3.421e+02  -1.412 0.159140
avg_salary_1987          6.033e-01  1.825e-01   3.306 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 83170.69)

    Null deviance: 53096433  on 262  degrees of freedom
Residual deviance: 20709503  on 249  degrees of freedom
AIC: 3741.4
```

ANOVA for sequential test to full and sub-model

```
> anova(full.mod1.gau.hit,step.mod1.gau.hit,test="Chisq")
Analysis of Deviance Table

Model 1: salary1987 ~ league_at_1987 + no_home_run_1986 + no_run_1986 +
    no_run_batted_1986 + no_walk_1986 + no_years_in_major_league +
    no_home_run_career + no_run_career + no_run_batted_career +
    no_walk_career + no_put_outs_1986 + no_assist_1986 + no_errors_1986 +
    changed_team + attendence_of_home_game + attendence_of_away_game +
    win_pct + avg_salary_1987
Model 2: salary1987 ~ no_home_run_1986 + no_run_1986 + no_run_batted_1986 +
    no_walk_1986 + no_home_run_career + no_run_career + no_walk_career +
    no_put_outs_1986 + no_errors_1986 + changed_team + attendence_of_away_game +
    win_pct + avg_salary_1987
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       244   20576958
2       249   20709503 -5  -132545   0.9046
>
```

TECHNICAL APPENDIX 1.2 (HITTERS)

GLM with Gaussian Family Log link full model

```
Call:
glm(formula = salary1987 ~ ., family = gaussian(link = "log"),
    data = hitter_data_mod1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-733.85  -170.96   -32.48   138.54  1031.55

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.196e+00  3.037e-01  17.111  < 2e-16 ***
league_at_1987N          8.443e-02  5.710e-02   1.479 0.140549
no_home_run_1986        -8.780e-03  6.747e-03  -1.301 0.194387
no_run_1986              5.133e-03  2.470e-03   2.078 0.038768 *
no_run_batted_1986       4.472e-03  2.872e-03   1.557 0.120743
no_walk_1986             8.363e-03  2.431e-03   3.441 0.000682 ***
no_years_in_major_league -2.449e-02  1.773e-02  -1.381 0.168427
no_home_run_career       -5.107e-04  9.535e-04  -0.536 0.592692
no_run_career             8.996e-04  3.794e-04   2.371 0.018495 *
no_run_batted_career      7.779e-04  4.648e-04   1.674 0.095482 .
no_walk_career           -6.519e-04  3.111e-04  -2.096 0.037156 *
no_put_outs_1986          2.431e-04  7.876e-05   3.086 0.002259 **
no_assist_1986           -5.689e-05  2.631e-04  -0.216 0.828987
no_errors_1986           -5.425e-03  6.421e-03  -0.845 0.398946
changed_teamTRUE          2.260e-01  1.048e-01   2.156 0.032063 *
attendence_of_home_game   9.982e-08  7.357e-08   1.357 0.176058
attendence_of_away_game  -1.004e-07  2.023e-07  -0.496 0.620104
win_pct                  -1.883e+00  6.488e-01  -2.902 0.004052 **
avg_salary_1987           1.604e-03  4.004e-04   4.006 8.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 77558.35)
```

GLM with Gaussian Family Log Link backward model

```
glm(formula = salary1987 ~ league_at_1987 + no_home_run_1986 +
    no_run_1986 + no_run_batted_1986 + no_walk_1986 + no_run_career +
    no_run_batted_career + no_walk_career + no_put_outs_1986 +
    changed_team + win_pct + avg_salary_1987, family = gaussian(link = "log"),
    data = hitter_data_mod1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-884.91  -165.22   -38.08   144.44  1057.37

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.905e+00  2.534e-01  19.355  < 2e-16 ***
league_at_1987N        9.641e-02  5.605e-02   1.720 0.086638 .
no_home_run_1986      -9.351e-03  5.700e-03  -1.641 0.102144
no_run_1986            6.241e-03  2.384e-03   2.617 0.009400 **
no_run_batted_1986     3.661e-03  2.679e-03   1.367 0.172968
no_walk_1986           8.729e-03  2.315e-03   3.771 0.000203 ***
no_run_career          6.639e-04  3.191e-04   2.081 0.038494 *
no_run_batted_career   6.260e-04  2.578e-04   2.428 0.015878 *
no_walk_career        -6.560e-04  2.630e-04  -2.494 0.013279 *
no_put_outs_1986       2.543e-04  7.097e-05   3.583 0.000408 ***
changed_teamTRUE       2.376e-01  1.019e-01   2.331 0.020531 *
win_pct               -1.698e+00  4.871e-01  -3.486 0.000579 ***
avg_salary_1987        1.659e-03  3.014e-04   5.504 9.17e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 77268.48)

    Null deviance: 53096433  on 262  degrees of freedom
Residual deviance: 19316564  on 250  degrees of freedom
AIC: 3721.1
```

ANOVA for sequential test to full and sub-model

```
> anova(full.mod1.gau.log.hit,step.mod1.gau.log.hit,test="Chisq")
Analysis of Deviance Table

Model 1: salary1987 ~ league_at_1987 + no_home_run_1986 + no_run_1986 +
    no_run_batted_1986 + no_walk_1986 + no_years_in_major_league +
    no_home_run_career + no_run_career + no_run_batted_career +
    no_walk_career + no_put_outs_1986 + no_assist_1986 + no_errors_1986 +
    changed_team + attendence_of_home_game + attendence_of_away_game +
    win_pct + avg_salary_1987
Model 2: salary1987 ~ league_at_1987 + no_home_run_1986 + no_run_1986 +
    no_run_batted_1986 + no_walk_1986 + no_run_career + no_run_batted_career +
    no_walk_career + no_put_outs_1986 + changed_team + win_pct +
    avg_salary_1987
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       244   18924101
2       250   19316564 -6  -392463   0.5361
```

TECHNICAL APPENDIX 1.3 (HITTERS)

GLM with Gamma Family Log link full model

```
glm(formula = salary1987 ~ ., family = Gamma(), data = hitter_data_mod2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2058  -0.5903  -0.1035   0.3160   1.4045

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.834e-03  6.983e-04   5.490 1.02e-07 ***
league_at_1987N         -2.121e-04  1.267e-04  -1.674  0.09548 .
no_home_run_1986         1.823e-06  1.495e-05   0.122  0.90308
no_run_1986             -1.361e-05  5.739e-06  -2.371  0.01851 *
no_run_batted_1986      -3.285e-06  6.538e-06  -0.502  0.61583
no_walk_1986            -1.584e-05  5.796e-06  -2.733  0.00673 **
no_years_in_major_league -8.105e-05  3.346e-05  -2.422  0.01616 *
no_home_run_career       1.564e-06  2.011e-06   0.778  0.43758
no_run_career           -1.571e-06  8.662e-07  -1.814  0.07097 .
no_run_batted_career    -3.007e-07  9.802e-07  -0.307  0.75929
no_walk_career           1.495e-06  7.027e-07   2.127  0.03442 *
no_put_outs_1986        -2.811e-07  2.143e-07  -1.312  0.19091
no_assist_1986          -1.135e-07  6.484e-07  -0.175  0.86113
no_errors_1986           7.049e-06  1.503e-05   0.469  0.63950
changed_teamTRUE        -1.153e-04  2.234e-04  -0.516  0.60612
attendence_of_home_game -2.012e-10  1.668e-10  -1.207  0.22877
attendence_of_away_game  7.428e-10  4.426e-10   1.678  0.09465 .
win_pct                  2.960e-03  1.485e-03   1.993  0.04742 *
avg_salary_1987         -2.745e-06  8.673e-07  -3.166  0.00175 **
one_position_onlyTRUE   -2.745e-04  2.259e-04  -1.215  0.22554
dedicated_hitterTRUE     2.736e-04  3.332e-04   0.821  0.41230
on_baseTRUE              1.588e-04  1.720e-04   0.923  0.35681
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GLM with Gamma Family Log Link backward model

```
Call:
glm(formula = salary1987 ~ league_at_1987 + no_run_1986 + no_walk_1986 +
    no_years_in_major_league + no_run_career + no_walk_career +
    no_put_outs_1986 + attendence_of_away_game + win_pct + avg_salary_1987,
    family = Gamma(), data = hitter_data_mod2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1542  -0.6400  -0.1094   0.3049   1.3639

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.044e-03  5.659e-04   7.145 9.64e-12 ***
league_at_1987N         -2.330e-04  1.213e-04  -1.921 0.055889 .
no_run_1986             -1.342e-05  4.004e-06  -3.351 0.000927 ***
no_walk_1986            -1.850e-05  5.133e-06  -3.604 0.000378 ***
no_years_in_major_league -9.221e-05  3.072e-05  -3.001 0.002957 **
no_run_career           -1.624e-06  5.523e-07  -2.941 0.003573 **
no_walk_career           1.903e-06  5.401e-07   3.524 0.000505 ***
no_put_outs_1986        -3.220e-07  1.497e-07  -2.151 0.032449 *
attendence_of_away_game  6.667e-10  4.085e-10   1.632 0.103943
win_pct                  1.855e-03  1.249e-03   1.486 0.138657
avg_salary_1987         -3.078e-06  7.539e-07  -4.082 6.00e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.363158)

    Null deviance: 184.439  on 262  degrees of freedom
Residual deviance:  95.837  on 252  degrees of freedom
AIC: 3647.1
```

ANOVA for sequential test to full and sub-model

```
> anova(full.mod2.gam.log.hit,step.mod2.gam.log.hit,test="Chisq")
Analysis of Deviance Table

Model 1: salary1987 ~ league_at_1987 + no_home_run_1986 + no_run_1986 +
    no_run_batted_1986 + no_walk_1986 + no_years_in_major_league +
    no_home_run_career + no_run_career + no_run_batted_career +
    no_walk_career + no_put_outs_1986 + no_assist_1986 + no_errors_1986 +
    changed_team + attendence_of_home_game + attendence_of_away_game +
    win_pct + avg_salary_1987 + one_position_only + dedicated_hitter +
    on_base
Model 2: salary1987 ~ league_at_1987 + no_run_1986 + no_walk_1986 + no_years_in_major_lea
gue +
    no_run_career + no_walk_career + no_put_outs_1986 + attendence_of_away_game +
    win_pct + avg_salary_1987
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       241     93.216
2       252     95.837 -11  -2.6209   0.7921
```

PCA analysis

```
> perform.pca <-prcomp(scale(cor_performance_hitter),center = TRUE)
> summary(perform.pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6     PC7     PC8
Standard deviation     2.7362 1.2115 1.0978 0.63523 0.40867 0.3366 0.26191 0.20341
Proportion of Variance 0.6806 0.1334 0.1096 0.03668 0.01518 0.0103 0.00624 0.00376
Cumulative Proportion  0.6806 0.8141 0.9236 0.96030 0.97548 0.9858 0.99202 0.99578
                           PC9    PC10    PC11
Standard deviation     0.16719 0.11574 0.07092
Proportion of Variance 0.00254 0.00122 0.00046
Cumulative Proportion  0.99832 0.99954 1.00000
```

TECHNICAL APPENDIX 2.1 (PITCHERS)

GLM with Gamma family and log link full model.

```
glm(formula = salary1987 ~ ., family = Gamma(link = "log"), data = pitcher_clean2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6003  -0.3296  -0.0781   0.2274   1.4786

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.368e+00  5.627e-01   9.539  < 2e-16 ***
DivisionW                 9.996e-03  7.559e-02   0.132  0.89497
no_wins_in_1986           1.616e-02  1.847e-02   0.875  0.38303
no_losses_in_1986        -1.106e-02  1.792e-02  -0.617  0.53816
earned_run_avg_1986       8.624e-02  4.979e-02   1.732  0.08529 .
no_games_1986            -1.401e-02  5.054e-03  -2.771  0.00626 **
no_inner_pitch_1986      -7.499e-04  1.857e-03  -0.404  0.68686
no_saves_1986             1.901e-02  8.919e-03   2.132  0.03460 *
no_years_in_major_league  2.257e-02  1.168e-02   1.933  0.05507 .
no_wins_in_career         4.282e-01  8.264e-01   0.518  0.60511
no_losses_in_career      -6.017e-01  7.675e-01  -0.784  0.43423
earned_run_avg_in_career -5.025e-01  5.067e-02  -9.917  < 2e-16 ***
no_games_in_career        1.694e-02  8.597e-03   1.971  0.05053 .
no_inner_pitch_in_career  4.906e-03  2.150e-03   2.282  0.02388 *
no_save_in_career        -3.950e-03  1.781e-02  -0.222  0.82473
changed_teamTRUE          2.783e-01  1.173e-01   2.373  0.01888 *
attendence_of_home_game   1.688e-07  9.168e-08   1.841  0.06750 .
attendence_of_away_game   8.665e-08  2.500e-07   0.347  0.72936
win_pct                  -1.490e+00  7.619e-01  -1.955  0.05234 .
avg_salary_1987           3.126e-04  4.165e-04   0.750  0.45417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.1999308)
```

GLM Gamma family and log link Forward model

```
Call:
glm(formula = salary1987 ~ earned_run_avg_in_career + no_inner_pitch_in_career +
    no_save_in_career + no_games_1986 + changed_team + avg_salary_1987 +
    no_games_in_career + no_saves_1986 + no_years_in_major_league,
    family = Gamma(link = "log"), data = pitcher_clean2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.51807  -0.36193  -0.06949   0.21179   1.61232

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.1230780  0.2913187  17.586  < 2e-16 ***
earned_run_avg_in_career -0.4808868  0.0468258 -10.270  < 2e-16 ***
no_inner_pitch_in_career  0.0050764  0.0009568   5.306 3.57e-07 ***
no_save_in_career         0.0058685  0.0170676   0.344  0.73140
no_games_1986            -0.0146627  0.0046604  -3.146  0.00196 **
changed_teamTRUE          0.2782257  0.1138905   2.443  0.01562 *
avg_salary_1987           0.0008247  0.0003372   2.446  0.01551 *
no_games_in_career        0.0164682  0.0063754   2.583  0.01066 *
no_saves_1986             0.0149307  0.0079450   1.879  0.06197 .
no_years_in_major_league  0.0204747  0.0111392   1.838  0.06785 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.206426)

    Null deviance: 114.371  on 174  degrees of freedom
Residual deviance:  32.249  on 165  degrees of freedom
AIC: 2290.8
```

ANOVA for sequential test to full and sub-model

```
> anova(pitcher_gamma,pitcher_step,test="Chisq")
Analysis of Deviance Table

Model 1: salary1987 ~ Division + no_wins_in_1986 + no_losses_in_1986 +
    earned_run_avg_1986 + no_games_1986 + no_inner_pitch_1986 +
    no_saves_1986 + no_years_in_major_league + no_wins_in_career +
    no_losses_in_career + earned_run_avg_in_career + no_games_in_career +
    no_inner_pitch_in_career + no_save_in_career + changed_team +
    attendence_of_home_game + attendence_of_away_game + win_pct +
    avg_salary_1987
Model 2: salary1987 ~ earned_run_avg_in_career + no_inner_pitch_in_career +
    no_save_in_career + no_games_1986 + changed_team + avg_salary_1987 +
    no_games_in_career + no_saves_1986 + no_years_in_major_league
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       155     30.272
2       165     32.249 -10  -1.9774   0.4502
```

TECHNICAL APPENDIX 2.2 (PITCHERS)

Bootstrap result for estimator

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = pitcher_clean2, statistic = boot.gam_pit, R = 1000)


Bootstrap Statistics :
          original        bias     std. error
t1*   6.7022978240 -2.559716e-02 0.4545450852
t2*  -1.7874398999  2.732009e-02 0.2290504836
t3*   1.4234957217  4.078210e-02 0.7335934148
t4*  -0.0741110145  2.304275e-03 0.0149074773
t5*   0.0170940316 -2.562072e-04 0.0064679930
t6*   0.4005144775  2.053952e-02 0.1394661342
t7*  -0.0105657380  3.411869e-05 0.0040917931
t8*  -0.7900045418 -3.877393e-02 0.5484109411
t9*   0.0003894952  1.963360e-05 0.0003590631
t10*  0.0191182044 -2.945612e-03 0.0124733687
t11*  0.0095759417  3.608403e-04 0.0036606904
t12*  0.0652948895 -5.646831e-04 0.0349423634
t13*  0.0026402204  4.475344e-05 0.0016463441
t14* -0.7913154818 -1.066251e-02 0.5912487605
```

Estimator result from standard summary output

```
Call:
glm(formula = salary1987 ~ earned_run_avg_in_career + no_inner_pitch_in_career +
    no_save_in_career + no_games_1986 + changed_team + avg_salary_1987 +
    no_games_in_career + no_saves_1986 + no_years_in_major_league,
    family = Gamma(link = "log"), data = pitcher_clean2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.51807  -0.36193  -0.06949   0.21179   1.61232

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.1230780  0.2913187  17.586  < 2e-16 ***
earned_run_avg_in_career -0.4808868  0.0468258 -10.270  < 2e-16 ***
no_inner_pitch_in_career  0.0050764  0.0009568   5.306 3.57e-07 ***
no_save_in_career         0.0058685  0.0170676   0.344  0.73140
no_games_1986            -0.0146627  0.0046604  -3.146  0.00196 **
changed_teamTRUE          0.2782257  0.1138905   2.443  0.01562 *
avg_salary_1987           0.0008247  0.0003372   2.446  0.01551 *
no_games_in_career        0.0164682  0.0063754   2.583  0.01066 *
no_saves_1986             0.0149307  0.0079450   1.879  0.06197 .
no_years_in_major_league  0.0204747  0.0111392   1.838  0.06785 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.206426)

    Null deviance: 114.371  on 174  degrees of freedom
Residual deviance:  32.249  on 165  degrees of freedom
AIC: 2290.8
```

CODE

```
library(openxlsx)
library(dplyr)
library(stats)
library(MASS)
library(goftest)
library(visdat)
library(ggcorrplot)
library(ggplot2)
library(car)
library(VIM)
library(mgcv)
library(gam)
library(car)
library(corrplot)
library(boot)
install.packages("DAAG")
library(DAAG)
################ Load Data ################

hitter <- read.xlsx(xlsxFile = "baseball_cleaned(3)(1).xlsx",sheet = 1)
pitcher <- read.xlsx(xlsxFile = "baseball_cleaned(3)(1).xlsx",sheet = 2)
team <- read.xlsx(xlsxFile = "baseball_cleaned(3)(1).xlsx",sheet = 3)

################ Preliminary Manipulation ################

## create variable for win percentage for each team
team$win_pct <- team$no_wins/(team$no_losses+team$no_wins)

## drop the dollar unit average salary
team<-team[,!(colnames(team)=="1987_average_salary")]

## identify team by both city and the league,
## for all three dataframe
## and for both 1986 and 1987
hitter$team_name_1986 <- paste(hitter$league_at_1986,".",hitter$team_at_1986,sep = "")
hitter$team_name_1987 <- paste(hitter$league_at_1987,".",hitter$team_at_1987,sep = "")
pitcher$team_name_1986 <- paste(pitcher$league_at_1986,".",pitcher$team_at_1986,sep = "")
pitcher$team_name_1987 <- paste(pitcher$league_at_1987,".",pitcher$team_at_1987,sep = "")
team$team_name <- paste(team$League,".",team$team,sep = "")

## creat indicator variable for whether player changed team
hitter<-hitter%>%mutate(changed_team=team_name_1986==team_name_1987)
pitcher <- pitcher%>%mutate(changed_team=team_name_1986==team_name_1987)

## because all variable except average salary in teams dataset
## refers to performance in 1986, and average salary is 1987,
```

```
## break it into two datasets: team_1986_status, team_1987_salary

team_1986_status <- team[,c(11,1:3,5:8,10)]
team_1987_salary <- team[,c(11,9)]


## merge hitter and pitcher data frame each with team, by 1986 team
merged_hitter <- merge(hitter,team_1986_status,
                  by.x = "team_name_1986", by.y= "team_name")
merged_hitter <- merge(merged_hitter,team_1987_salary,
                  by.x = "team_name_1987", by.y= "team_name")
merged_pitcher <- merge(pitcher,team_1986_status,
                   by.x ="team_name_1986",by = "team_name")
merged_pitcher <- merge(merged_pitcher,team_1987_salary,
                   by.x ="team_name_1987",by = "team_name")

################ Characters Columns -> factors  ################

#### step 1: identify character columns
is_char <- lapply(merged_hitter,typeof)=="character"
#### step 2: separate the data set by character and numeric
numerical_hitter <- merged_hitter[,!is_char]
categorical_hitter <- merged_hitter[,is_char]
#### step 3: apply transformation to all character columns
categorical_hitter <- as.data.frame(lapply(categorical_hitter,as.factor))
#### step 4: merge back together
merged_hitter <- cbind(categorical_hitter,numerical_hitter)

#### repeat for pitcher
is_char_pitcher <- lapply(merged_pitcher,typeof)=="character"
numerical_pitcher <- merged_pitcher[,!is_char_pitcher]
categorical_pitcher <- merged_pitcher[,is_char_pitcher]
merged_pitcher<-cbind(categorical_pitcher,numerical_pitcher)

################## distribution of response ####################

set.seed(1)
summary(hitter$salary1987)

#### density plot
par(mfrow=c(1,1))
plot(density(na.omit(hitter$salary1987)),ylim=c(0,0.0014))

fit_hitter_gamma<-fitdistr(na.omit(hitter$salary1987), densfun = "gamma")
rand.gamma <- rgamma(10000,
                  shape = fit_hitter_gamma$estimate[1],
                  rate = fit_hitter_gamma$estimate[2])
```

```
lines(density(rand.gamma),col="red")

#### qq plot
qqplot(x=qgamma(ppoints(hitter$salary1987),shape = fit_hitter_gamma$estimate[1],
           rate = fit_hitter_gamma$estimate[2]) ,y=hitter$salary1987)
abline(a=0,b=1,col="red")

#### goodness of fit
cvm.test(na.omit(hitter$salary1987),"pgamma",shape = fit_hitter_gamma$estimate[1],
       rate = fit_hitter_gamma$estimate[2])
ad.test(na.omit(hitter$salary1987),"pgamma",shape = fit_hitter_gamma$estimate[1],
      rate = fit_hitter_gamma$estimate[2])

## repeat the above on pitcher

plot(density(na.omit(pitcher$`1987salary`)),ylim=c(0,0.0014))

fit_pitcher_gamma<-fitdistr(na.omit(pitcher$`1987salary`), densfun = "gamma")
set.seed(1)
rand.gamma_pitcher <- rgamma(10000,
                     shape = fit_pitcher_gamma$estimate[1],
                     rate = fit_pitcher_gamma$estimate[2])
lines(density(rand.gamma),col="red")

qqplot(x=qgamma(ppoints(pitcher$`1987salary`),shape = fit_hitter_gamma$estimate[1],
           rate = fit_hitter_gamma$estimate[2]) ,y=hitter$salary1987)
abline(a=0,b=1,col="red")

cvm.test(na.omit(pitcher$`1987salary`),"pgamma",shape = fit_pitcher_gamma$estimate[1],
       rate = fit_pitcher_gamma$estimate[2])
ad.test(na.omit(pitcher$`1987salary`),"pgamma",shape = fit_pitcher_gamma$estimate[1],
      rate = fit_pitcher_gamma$estimate[2])

####
## therefore, gamma provides good fit to hitter data, goodness of fit
## test indicate that we cannot reject the null hypthesis that hitter
## salary is of gamma distribution, for pitcher data, although gamma
## distribution fits the data's empirical density less well,by cvm test,
## we cannot reject the null hypothesis that it is gamma distributed at
## 5% significance level
##
## conclusion: use GAMMA family to fit glm
####

################## missing data ####################

vis_miss(hitter)+theme(axis.text.x = element_text(angle = 90))
```

```
vis_miss(pitcher)+theme(axis.text.x = element_text(angle = 90))

## however should not impute??
hitter.no.na <- na.omit(merged_hitter)
pitcher.no.na <- na.omit(merged_pitcher)

######### move response to the first column #######
pos_y_hit <- which(colnames(hitter.no.na)=="salary1987")
salary1987.h <-hitter.no.na[,pos_y_hit]
x_h <- hitter.no.na[,-pos_y_hit]
hitter_data0<-cbind(salary1987.h,x_h)
colnames(hitter_data0)[1]<-"salary1987"

pos_y_pit <- which(colnames(pitcher.no.na)=="1987salary")
salary1987.p <- pitcher.no.na[,pos_y_pit] # changes the name of salary column btw
x_p <- pitcher.no.na[,-pos_y_pit]
pitcher_data0 <- cbind(salary1987.p,x_p)
colnames(pitcher_data0)[1]<-"salary1987"

# also the last column
colnames(hitter_data0)[which(colnames(hitter_data0)=="1987_avg_salary")]<-"avg_salary_1987"
colnames(pitcher_data0)[which(colnames(pitcher_data0)=="1987_avg_salary")]<-
"avg_salary_1987"


######### data exploration ################

## correlation

str(numerical_hitter)
colnames(numerical_hitter)[c(17:18)]
cor_data_hitter <- numerical_hitter[,c(1:17,19:25)]
cor_data_hitter<-na.omit(cor_data_hitter)
corrplot(corr = cor(cor_data_hitter))
# conlusion:
# at first glance: three clusters of predictors:
#   1) personal performance in 1986
#   2) personal performance cumulatively in career
#   3) performance of team player with in 1986
# however note the strong correlation between years in
# league and all careers stats
# try: find average stat in career instead

cor_data_hitter[,8:13] <-cor_data_hitter[,8:13]/cor_data_hitter[,7]
corrplot(corr = cor(cor_data_hitter))
# conclusion: personal performance forms a big cluster with relative
# higher correlations between them pairwise
```

```
# try use principal component analysis to reduce dimension only amongst
# performance statistics

# note the high vif as well (with )
colnames(cor_data_hitter)
vif_hitter <- cor_data_hitter[,c(17,1:16)]
model <- glm(salary1987~.,data = vif_hitter)
vif_hit<-vif(model)
par(mar=c(12,4,6,1))
b_h <- barplot(vif_hit,las=3,ylim = c(0,140))
text(b_h,vif_hit+10,round(vif_hit,1),cex=0.8)
abline(h=10,lty=2,col="red")
dev.off()
summary(model)
# no_walk_1986, no_put_outs_career, no_assist_1986, no_errors_1986
# do not have high VIF

colnames(cor_data_hitter)
cor_performance_hitter <- cor_data_hitter[,c(1:5,8:13)]
corrplot(cor(cor_performance_hitter))

#########################
## from here, we have two choices:
##    1) remove some variables based on assumption:
##         + easier to interpret
##         - risk of erorr in assumption
##    2) use partial PCA to reduce dimension but only amongst
##       performance data
##         + less likely to commit assumption bias
##         - difficult to interpret
#########################

perform.pca <-prcomp(scale(cor_performance_hitter),center = TRUE)
summary(perform.pca)
perform.var <-perform.pca$sdev^2
perform.var.total <- sum(perform.var)
(cumulative.var <- cumsum(perform.var/perform.var.total))
plot(cumulative.var,type="b",ylim=c(0,1))
grid()
## from this graph, we see that first four principal component explains
## 95% of the variance, take first four principal component
dev.off()
biplot(perform.pca)
## two clear cluster of performance measure career vs. 1986
## this shows  the need to not eradicate any one of these
## two groups of predictors entirely
```

```
head((perform.pca$x))

## team performance and ranking
table(team$no_wins+team$no_losses)
ranking_data <- team[,c(1:3,10)]
ranking_data$League.division <- paste(ranking_data$League,ranking_data$Division,sep = "")

ggplot(data = ranking_data)+
  geom_line(aes(Position_in_Final_League,win_pct,
            color=League.division))+theme_bw()
cor(ranking_data$win_pct,ranking_data$Position_in_Final_League)
## given that every team played about the same number of games
## remove no_wins and no_loss and keep win rate only
## also note by default ranking of teams is decided by win rates,
## and the clear relationship, therefore drop Position_in_Final_League


########################################
########################################
##########          ##############
##########  Modelling  ##############
##########          ##############
########################################
########################################

########################################
############### GLM ###################
########################################

########## Model 1:
# remove some covariates by correlation, including position

# prepare data
colnames(hitter_data0)

drop_hitter1 <- c("team_name_1986","team_name_1987","name","team_at_1986",
            "team_at_1987","division_at_1986","league_at_1986","League",
            "Division","no_wins","no_losses","Position_in_Final_League",
            "no_at_bat_1986","no_hit_1986",
            "no_at_bat_in_career","no_hit_career","position_at_1986")
hitter_data_mod1 <- hitter_data0[,!(colnames(hitter_data0)%in%drop_hitter1)]
str(hitter_data_mod1)

# gaussian
full.mod1.gau.hit <- glm(salary1987~.,data=hitter_data_mod1)
(summary1 <- summary(full.mod1.gau.hit))
step.mod1.gau.hit <- step(full.mod1.gau.hit,direction = "backward",K=2)
```

```
(summary2 <- summary(step.mod1.gau.hit))
anova(full.mod1.gau.hit,step.mod1.gau.hit,test="Chisq")

par(mfrow=c(2,2))
plot(full.mod1.gau.hit)
plot(step.mod1.gau.hit)
dev.off()

R.sq1 <- 1-full.mod1.gau.hit$deviance/full.mod1.gau.hit$null.deviance
R.sq2 <- 1-step.mod1.gau.hit$deviance/step.mod1.gau.hit$null.deviance

scaled.deviance1 <- summary1$deviance /summary1$dispersion
scaled.deviance2 <- summary2$deviance /summary2$dispersion

set.seed(1)
cv.glm(hitter_data_mod1,full.mod1.gau.hit,K=10)[3]
set.seed(1)
cv.glm(hitter_data_mod1,step.mod1.gau.hit,K=10)[3]

# gaussian with log link
full.mod1.gau.log.hit <- glm(salary1987~.,family = gaussian(link = "log"),data=hitter_data_mod1)
(summary3 <- summary(full.mod1.gau.log.hit))
step.mod1.gau.log.hit <- step(full.mod1.gau.log.hit,direction = "backward",K=2)
(summary4 <- summary(step.mod1.gau.log.hit))
anova(full.mod1.gau.log.hit,step.mod1.gau.log.hit,test="Chisq")

par(mfrow=c(2,2))
plot(full.mod1.gau.log.hit)
plot(step.mod1.gau.log.hit)[1]
dev.off()

R.sq3 <- 1-full.mod1.gau.log.hit$deviance/full.mod1.gau.log.hit$null.deviance
R.sq4 <- 1-step.mod1.gau.log.hit$deviance/full.mod1.gau.log.hit$null.deviance

(scaled.deviance3<-summary3$deviance/summary3$dispersion)
(scaled.deviance4<-summary4$deviance/summary4$dispersion)

cv.glm(hitter_data_mod1,full.mod1.gau.log.hit,K=10)[3]
cv.glm(hitter_data_mod1,step.mod1.gau.log.hit,K=10)[3]

## gamma log

full.mod1.gam.log.hit <- glm(salary1987~.,family = Gamma(link = "log"),data=hitter_data_mod1)
summary5 <- summary(full.mod1.gam.log.hit)
step.mod1.gam.log.hit <- step(full.mod1.gam.log.hit,direction = "backward",K=2)
summary6 <- summary(step.mod1.gam.log.hit)
```

```
par(mfrow=c(2,2))
plot(full.mod1.gam.log.hit)
plot(step.mod1.gam.log.hit)
dev.off()

R.sq5 <- 1-full.mod1.gam.log.hit$deviance/full.mod1.gam.log.hit$null.deviance
R.sq6 <- 1-step.mod1.gam.log.hit$deviance/step.mod1.gam.log.hit$null.deviance

(scaled.deviance5<-summary5$deviance/summary5$dispersion)
(scaled.deviance6<-summary6$deviance/summary6$dispersion)

set.seed(1)
cv.glm(hitter_data_mod1,full.mod1.gam.log.hit,K=10)[3]
set.seed(1)
cv.glm(hitter_data_mod1,step.mod1.gam.log.hit,K=10)[3]

## gamma inverse
full.mod1.gam.inv.hit <- glm(salary1987~.,family = Gamma(),data=hitter_data_mod1)
summary7 <- summary(full.mod1.gam.inv.hit)
step.mod1.gam.inv.hit <- step(full.mod1.gam.inv.hit,direction = "backward",K=2)
summary8 <- summary(step.mod1.gam.log.hit)

par(mfrow=c(2,2))
plot(full.mod1.gam.inv.hit)
plot(step.mod1.gam.inv.hit)
dev.off()

R.sq7 <- 1-full.mod1.gam.inv.hit$deviance/full.mod1.gam.inv.hit$null.deviance
R.sq8 <- 1-step.mod1.gam.inv.hit$deviance/step.mod1.gam.log.hit$null.deviance

(scaled.deviance7<-summary7$deviance/summary7$dispersion)
(scaled.deviance8<-summary8$deviance/summary8$dispersion)

set.seed(1)
cv.glm(hitter_data_mod1,full.mod1.gam.inv.hit,K=10)[3]
set.seed(1)
cv.glm(hitter_data_mod1,step.mod1.gam.inv.hit,K=10)[3]


########## Model 2:
# manipulate position, remove some covariates by correlation
hitter_data_mod2 <- hitter_data0
one_position <- c("1B","2B","3B","SS","RF","LF","C","DH","OF")

hitter_data_mod2$one_position_only <- hitter_data_mod2$position_at_1986%in%one_position
hitter_data_mod2$dedicated_hitter <-
hitter_data_mod2$position_at_1986%in%c("DH","CD","OD")
```

```
hitter_data_mod2$on_base <- hitter_data_mod2$position_at_1986%in%c("13","1B","23",
                                                   "2B","2S","32","3B",
                                                   "3S")
#c("13","1B","1O","23",
#  "2B","2S","32","3B",
#  "3O","3S","C","CD",
#  "CF","CS")

boxplot(hitter_data_mod2$salary1987~hitter_data_mod2$position_at_1986)

str(hitter_data0)

plot(density(hitter_data_mod2$salary1987[!hitter_data_mod2$on_base]),col="red")
lines(density(hitter_data_mod2$salary1987[hitter_data_mod2$on_base]))



drop_hitter2 <- c("team_name_1986","team_name_1987","name","team_at_1986",
            "team_at_1987","division_at_1986","league_at_1986","League",
            "Division","no_wins","no_losses","Position_in_Final_League",
            "no_at_bat_1986","no_hit_1986",
            "no_at_bat_in_career","no_hit_career","position_at_1986")
hitter_data_mod2 <- hitter_data_mod2[,!(colnames(hitter_data_mod2)%in%drop_hitter2)]
str(hitter_data_mod2)



# gaussian
full.mod2.gau.hit <- glm(salary1987~.,data=hitter_data_mod2)
(summary2_1 <- summary(full.mod2.gau.hit))
step.mod2.gau.hit <- step(full.mod2.gau.hit,direction = "backward",K=2)
(summary2_2 <- summary(step.mod2.gau.hit))

par(mfrow=c(2,2))
plot(full.mod1.gau.hit)
plot(step.mod1.gau.hit)
dev.off()

# note that the only difference between model 2 stepwise gaussian and mdoel 1 is Dedicated hitter
variable
anova(step.mod1.gau.hit,step.mod2.gau.hit,test = "F")

R.sq2_1 <- 1-full.mod2.gau.hit$deviance/full.mod1.gau.hit$null.deviance
R.sq2_2 <- 1-step.mod2.gau.hit$deviance/step.mod1.gau.hit$null.deviance

scaled.deviance2_1 <- summary2_1$deviance /summary2_1$dispersion
scaled.deviance2_2 <- summary2_2$deviance /summary2_2$dispersion

set.seed(1)
```

```
cv.glm(hitter_data_mod1,full.mod1.gau.hit,K=10)[3]
set.seed(1)
cv.glm(hitter_data_mod1,step.mod1.gau.hit,K=10)[3]

# gaussian with log link
full.mod2.gau.log.hit <- glm(salary1987~.,family = gaussian(link = "log"),data=hitter_data_mod2)
(summary2_3 <- summary(full.mod2.gau.log.hit))
step.mod2.gau.log.hit <- step(full.mod2.gau.log.hit,direction = "backward",K=2)
(summary2_4 <- summary(step.mod2.gau.log.hit))

anova(step.mod1.gau.log.hit,step.mod2.gau.log.hit,test="F")

par(mfrow=c(2,2))
plot(full.mod2.gau.log.hit)
plot(step.mod1.gau.log.hit)
dev.off()

R.sq2_3 <- 1-full.mod2.gau.log.hit$deviance/full.mod2.gau.log.hit$null.deviance
R.sq2_4 <- 1-step.mod2.gau.log.hit$deviance/full.mod2.gau.log.hit$null.deviance

(scaled.deviance2_3<-summary2_3$deviance/summary2_3$dispersion)
(scaled.deviance2_4<-summary2_4$deviance/summary2_4$dispersion)

cv.glm(hitter_data_mod2,full.mod1.gau.log.hit,K=10)[3]
cv.glm(hitter_data_mod2,step.mod1.gau.log.hit,K=10)[3]

# gamma log
full.mod2.gam.log.hit <- glm(salary1987~.,family = Gamma(),data=hitter_data_mod2)
step.mod2.gam.log.hit <- step(full.mod2.gam.log.hit,direction = "backward",K=2)
summary6 <- summary(step.mod2.gam.log.hit)
anova(full.mod2.gam.log.hit,step.mod2.gam.log.hit,test="Chisq")

plot(step.mod2.gam.log.hit)
cv.glm(hitter_data_mod2,step.mod2.gam.log.hit,K=10)[3]

########## Model 3:
# manipulate position, partial PCA dimension reduction

colnames(hitter_data_mod2)

trans.performance <- perform.pca$x[,1:4]

hitter_data_mod3 <- cbind(hitter_data_mod2[,c(1,2,15:22)],trans.performance)


# gaussian
```

```
mod3_1 <- glm(salary1987~., family = gaussian(link = "log"), data=hitter_data_mod3)
summary(mod3_1)

mod3_2 <- step(mod3_1,direction = "backward")
summary(mod3_2)

plot(mod3_2)
cv.glm(hitter_data_mod3,mod3_2,K=10)[3]

# gamma

mod3_3 <- glm(salary1987~., family = Gamma(link = "log"), data=hitter_data_mod3)
summary(mod3_3)
mod3_4 <- step(mod3_3,direction = "ba")
cv.glm(hitter_data_mod3,mod3_4,K=10)[3]
plot(mod3_4)


##### candidate

# model 2, gaussian, log link
# model 3, gamma log link

### bootstrap for gaussian
formula(step.mod2.gau.log.hit)
boot.gau<-function(dataset,rows.used){
  return(coef(glm(salary1987 ~ league_at_1987 + no_run_1986 + no_walk_1986 + no_run_career +
                    no_run_batted_career + no_walk_career + no_put_outs_1986 +
                    changed_team + attendence_of_home_game + win_pct + avg_salary_1987 +
                    dedicated_hitter, family = gaussian(link = "log"),
                  data=dataset,
                  subset = rows.used)))
}
boot(hitter_data_mod2,boot.gau,R=1000)
summary(step.mod2.gau.log.hit)

### bootstrap for gamma
formula(mod3_4)
boot.gam<-function(dataset,rows.used){
  return(coef(glm(salary1987 ~ avg_salary_1987 + one_position_only + PC1 + PC2 +
                    PC3, family = Gamma(link = "log"),
                  data=dataset,
                  subset = rows.used)))
}
boot(hitter_data_mod3,boot.gam,R=1000)
summary(mod3_4)
```

```
############################
############################
### pitcher manipulation
####### for variable in terms of career, divided it by the # of year in major league
####### for variable wins and loss in career, divided it by the # of games in career
####### for variable games in career, divided it by # of year in major league (after modify wins and loss in career)
pitcher_data0$earned_run_avg_in_career<-
pitcher_data0$earned_run_avg_in_career/pitcher_data0$no_years_in_major_league
pitcher_data0$no_inner_pitch_in_career<-
pitcher_data0$no_inner_pitch_in_career/pitcher_data0$no_years_in_major_league
pitcher_data0$no_save_in_career<-
pitcher_data0$no_save_in_career/pitcher_data0$no_years_in_major_league
pitcher_data0$no_wins_in_career<-
pitcher_data0$no_wins_in_career/pitcher_data0$no_games_in_career
pitcher_data0$no_losses_in_career<-
pitcher_data0$no_losses_in_career/pitcher_data0$no_games_in_career
pitcher_data0$no_games_in_career<-
pitcher_data0$no_games_in_career/pitcher_data0$no_years_in_major_league

####### deleting variable team name as change team indicator presented, also delete name,
####### team ranking 1986, team # of win and loss (win.pct exist),
pitcher_clean2<-pitcher_data0[,-c(2:9,25:27)]

####### correlation plot
cor_pitcher<-cor(pitcher_clean2[,-c(2,16)])
corrplot(cor_pitcher)


################
######
################
###### fitting model
###inverse gamma
pitcher_gamma<-glm(salary1987~., data=pitcher_clean2,family=Gamma(link="log"))
###fit stepwise
pitcher_gamma_null<-glm(salary1987~1,data=pitcher_clean2,family=Gamma(link="log"))
pitcher_step<-stepAIC(pitcher_gamma_null,formula(pitcher_gamma),direction = "forward",k=2)
summary(pitcher_gamma)
summary(pitcher_step)
anova(pitcher_gamma,pitcher_step,test="Chisq")
plot(pitcher_gamma)

###deviance residual plot
dev.res_p <- residuals(pitcher_step,type = "deviance")
```

```
plot(predict(pitcher_step),dev.res_p, xlab="Fitted Values",
    ylab="Deviance Residual", main="Deviance Residuals vs. Fitted Values")
smooth.fit_p <- smooth.spline(predict(pitcher_step),dev.res_p)
lines(smooth.fit_p,type = "l",col="red")
###qq plot
qqplot(x=qnorm(ppoints(1000),mean = 0,sd=1),y=rstandard(pitcher_step),
      xlab="Standardised Deviance Residual", ylab="Deviance Residual", main="Normal QQ-plot of
Standardised Deviance Residuals")
abline(a=0,b=1,lty=2)

###fit gaussian log
pitcher_gau<-glm(salary1987~., data=pitcher_clean2,family=gaussian(link="log"))
summary(pitcher_gau)
plot(pitcher_gau)

###fit stepwise
pitcher_gau_null<-glm(salary1987~1,data=pitcher_clean2,family=gaussian(link="log"))
pitcher_step_gau<-stepAIC(pitcher_gau_null,formula(pitcher_gau),direction = "forward",k=2)
summary(pitcher_step_gau)
anova(pitcher_gau,pitcher_step_gau,test="Chisq")


plot(density(predict(pitcher_gau,newdata=pitcher_clean2,type="response")))
lines(density(pitcher_clean2$salary1987))
###deviance residual
dev.res_p <- residuals(mod3_4,type = "deviance")
plot(predict(mod3_4),dev.res_p, xlab="Fitted Values",
    ylab="Deviance Residual", main="Deviance Residuals vs. Fitted Values")
smooth.fit_p <- smooth.spline(predict(mod3_4),dev.res_p)
lines(smooth.fit_p,type = "l",col="red")
###qq plot
qqplot(x=qnorm(ppoints(1000),mean = 0,sd=1),y=rstandard(pitcher_step_gau),
      xlab="Standardised Deviance Residual", ylab="Deviance Residual", main="Normal QQ-plot of
Standardised Deviance Residuals")
abline(a=0,b=1,lty=2)

cv.glm(pitcher_clean2,pitcher_step_gau)[3]


###bootstrap for  gamma GAM
formula(pitcher_step)
boot.gam_pit<-function(dataset,rows.used){
  return(coef(glm(salary1987 ~ earned_run_avg_in_career + no_years_in_major_league +
                attendence_of_home_game + no_inner_pitch_in_career + changed_team +
                avg_salary_1987 + win_pct + no_saves_1986 + no_wins_in_1986 +
                no_wins_in_career + no_losses_in_career,
data=pitcher_clean2,family=Gamma(link="log"),
```

```
            subset = rows.used)))
}
boot(pitcher_clean2,boot.gam_pit,R=1000)
summary(pitcher_step)

###bootstrap for gaussian log
formula(pitcher_step_gau)
boot.gam_pit<-function(dataset,rows.used){
  return(coef(glm(salary1987 ~ earned_run_avg_in_career + no_wins_in_career +
no_years_in_major_league +
              no_games_in_career + changed_team + no_games_1986 + win_pct +
              avg_salary_1987 + no_wins_in_1986 + no_saves_1986 + earned_run_avg_1986 +
              no_inner_pitch_in_career + no_losses_in_career,
data=pitcher_clean2,family=gaussian(link="log"),
            subset = rows.used)))
}
boot(pitcher_clean2,boot.gam_pit,R=1000)
summary(pitcher_gau)

###### vif
colnames(pitcher_clean2)
vif_pitcher <- pitcher_clean2[-c(1,16)]
model_p <- glm(salary1987~.,data = vif_pitcher)
vif_pit<-vif(model_p)
par(mar=c(12,4,6,1))
b_h_p <- barplot(vif_pit,las=3,ylim = c(0,30))
text(b_h,vif_pit+10,round(vif_pit,1),cex=0.8)
abline(h=10,lty=2,col="red")
```