UNIVERSITY OF NEW SOUTH WALES
SCHOOL OF MATHEMATICS AND STATISTICS

MATH3821 Statistical Modelling and Computing
Term Two 2020

Assignment Two

Given: Friday 17th July 2020 **Due date: Sunday 2nd August 2020**

INSTRUCTIONS: This assignment is to be done **collaboratively** by a group of **5 students**. The same mark will be given for the report to each student within the group, unless I have good reasons to believe that somebody did not do anything.

You will need to produce and submit a report of your work in PDF format. This report will not contain more than 10 pages, excluding the Appendix that should contain your computing codes. The report is due 11:59 pm, Sunday 2nd August. The first page of this PDF should be **this page**. Only one of the five students should submit the PDF file on Moodle, with the names of the other students in the group clearly indicated in the document.

Name	Student No	Signature	Date
Eugenia Zhuohuan Cao	z5165757		02/08/20
Jinhao Huang	z5207964		02/08/20
Nigel Lin	z5160614		02/08/20
Ellen Wang	z5209394		02/08/20
Pai Yu	z5206136		02/08/20

MATH3821 Assignment 2

Eugenia Zhuohuan Cao, Jinhao Huang, Nigel Lin, Ellen Wang, Pai Yu

02/08/20

Baseball Data Set

For this report we will be assessing the baseball data set, which addresses the 1986 salaries and performances of North American Major League Baseball players. The data set was obtained from the ASA 1988 Data Exposition, with the corrections and revisions incorporated. Our main objective is to determine the key performance drivers of baseball hitters and how they contribute to their salaries.

For the purpose of our analysis, we will primarily be focusing on numerical variables and will be excluding most qualitative variables such as hitter's name, league, division and team, unless we find valid reasons to include them.

Linear Model

We begin by examining the correlation coefficients of all the predictors individually against the response variable, Salary. The correlation coefficient is a value that measures how strong a relationship is between two variables. This coefficient lies between two values 1 (strong positive relationship) and -1 (strong negative relationship) where 0 indicates that there is no relationship at all. The table below shows a side by side comparison of a linear model versus a log-linear model for all our quantitative variables:

Variable	Correlation Coefficient	Correlation Coefficient (Log)
AB_1986	0.4699	**0.4761**
H_1986	0.5144	**0.5145**
HR_1986	**0.3941**	0.3639
R_1986	**0.4881**	0.4777
RBI_1986	**0.5179**	0.4872
BB_1986	**0.5049**	0.4715
Years	0.4467	**0.5707**
AB_career	0.5756	**0.6451**
H_career	0.5963	**0.6534**
HR_career	**0.5812**	0.5465
R_career	0.6161	**0.6533**
RBI_career	0.6210	**0.6335**
W_career	0.5468	**0.5743**
Put_outs_1986	**0.2995**	0.2238
Assists_1986	0.0330	**0.0631**
Errors_1986	-0.0042	**-0.0174**

A log transformation seems to give us higher correlations with Salary. We assess the summary of a log-linear model and determine which variables are significant under a 5% significance level.

```
Call:
lm(formula = log(Salary) ~ ., data = baseball, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0394 -0.3053  0.0000  0.2815  1.1308 

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.4648451  0.7405503   4.679 5.71e-06 ***
AB_1986        -0.0044007  0.0012158  -3.620 0.000385 ***
H_1986          0.0186288  0.0042294   4.405 1.83e-05 ***
HR_1986         0.0123347  0.0128832   0.957 0.339660    
R_1986         -0.0040963  0.0058675  -0.698 0.486011    
RBI_1986       -0.0018696  0.0051120  -0.366 0.715010    
BB_1986         0.0151912  0.0033301   4.562 9.45e-06 ***
Years           0.0750785  0.0243599   3.082 0.002385 ** 
AB_career       0.0003752  0.0002532   1.482 0.140216    
H_career       -0.0015586  0.0012856  -1.212 0.226992    
HR_career       0.0004229  0.0032386   0.131 0.896257    
R_career        0.0026746  0.0014519   1.842 0.067134 .  
RBI_career     -0.0003171  0.0013326  -0.238 0.812191    
W_career       -0.0019199  0.0006134  -3.130 0.002044 ** 
League_1986N   -0.7388627  0.5683023  -1.300 0.195249    
Div_1986W       0.0658330  0.3822772   0.172 0.863467    
Team_1986Bal.  -2.4538961  1.1882045  -2.065 0.040360 *  
Team_1986Bos.  -2.7341553  1.3031030  -2.098 0.037308 *  

Position_19863O  1.1793063  0.8366105   1.410 0.160405    
Position_19863S  0.7663942  0.4900179   1.564 0.119601    
Position_1986C   0.5892994  0.2139356   2.755 0.006491 ** 
Position_1986CD  0.3236938  0.6343331   0.510 0.610484    
Position_1986CF  0.3883224  0.3112416   1.248 0.213805    
Position_1986DH  0.5253804  0.3918509   1.341 0.181715    
Position_1986DO  0.3944031  0.4939433   0.798 0.425663    
Position_1986LF  0.5796125  0.3247032   1.785 0.075965 .  
Position_1986O1  0.6906195  0.3831693   1.802 0.073186 .  
Position_1986OD  0.7867757  0.6032320   1.304 0.193836    
Position_1986OF  0.6372662  0.3075439   2.072 0.039705 *  
Position_1986OS  0.1185044  0.5115302   0.232 0.817065    
Position_1986RF  0.7473509  0.3349007   2.232 0.026899 *  
Position_1986S3  2.3438785  0.7185304   3.262 0.001327 ** 
Position_1986SS  1.1915888  0.3983721   2.991 0.003176 ** 
Position_1986UT  0.7950514  0.3491243   2.277 0.023966 *  
Put_outs_1986    0.0009024  0.0003724   2.423 0.016382 *  
Assists_1986    -0.0009406  0.0008490  -1.108 0.269447    
Errors_1986     -0.0116363  0.0092183  -1.262 0.208501    
League_1987N     1.0318537  0.5354157   1.927 0.055556 .  
Team_1987Bal.    2.9045039  0.9953546   2.918 0.003979 ** 
Team_1987Bos.    2.8073593  1.1593472   2.421 0.016466 *  
Team_1987Cal.    3.4411240  1.1430301   3.011 0.002989 ** 
Team_1987Chi.    1.6496229  0.5932194   2.781 0.006010 ** 
Team_1987Cin.    0.2778396  0.6635466   0.419 0.675929    
Team_1987Cle.    1.9236286  1.1416059   1.685 0.093747 .  
Team_1987Det.    1.4688275  0.6587529   2.230 0.027024 *  
```

```
Team_1986Cal.   -3.4098350  1.1613479  -2.936 0.003765 ** 
Team_1986Chi.   -1.3138272  0.6698423  -1.961 0.051401 .  
Team_1986Cin.   -0.7390633  0.6447573  -1.146 0.253232    
Team_1986Cle.   -1.4162692  1.3330540  -1.062 0.289490    
Team_1986Det.   -1.5151019  0.8720434  -1.737 0.084053 .  
Team_1986Hou.   -0.2120939  0.2705376  -0.784 0.434104    
Team_1986K.C.   -2.1451832  1.1034234  -1.944 0.053467 .  
Team_1986L.A.   -0.2947262  0.8767057  -0.336 0.737137    
Team_1986Mil.   -1.8736370  1.0122910  -1.851 0.065852 .  
Team_1986Min.   -2.6556639  1.1287785  -2.353 0.019737 *  
Team_1986Mon.   -2.1562229  0.8239636  -2.617 0.009641 ** 
Team_1986N.Y.   -1.4321605  0.7554873  -1.896 0.059632 .  
Team_1986Oak.   -2.0174258  1.1294560  -1.786 0.075780 .  
Team_1986Phi.   -0.5736677  0.6052089  -0.948 0.344481    
Team_1986Pit.   -0.0120866  0.6677717  -0.018 0.985579    
Team_1986S.D.   -1.6251522  0.4663458  -3.485 0.000621 ***
Team_1986S.F.   -1.3335062  0.6795821  -1.962 0.051302 .  
Team_1986Sea.   -2.1916523  1.1532944  -1.900 0.059014 .  
Team_1986St.L.   0.1812486  0.6665462   0.272 0.785999    
Team_1986Tex.   -1.3511888  0.8350000  -1.618 0.107402    
Team_1986Tor.   -0.5915663  1.2404747  -0.477 0.634031    
Position_19861O -0.1824636  0.5623103  -0.324 0.745951    
Position_198623  1.7343191  0.9221860   1.881 0.061661 .  
Position_19862B  1.0680215  0.3891872   2.744 0.006690 ** 
Position_19862S  1.7491259  0.6322149   2.767 0.006265 ** 
Position_198632  1.0115862  0.6481651   1.561 0.120383    
Position_19863B  0.9523186  0.3820287   2.493 0.013593 *  

Team_1987K.C.    1.8257016  1.0743079   1.699 0.090997 .  
Team_1987L.A.    0.5986775  0.8843842   0.677 0.499326    
Team_1987Mil.    1.9405140  0.7856580   2.470 0.014462 *  
Team_1987Min.    2.8265484  1.1066737   2.554 0.011489 *  
Team_1987Mon.    2.1076443  0.7369225   2.860 0.004745 ** 
Team_1987N.Y.    1.6633664  0.5497226   3.026 0.002849 ** 
Team_1987Oak.    2.2476301  1.0855246   2.071 0.039853 *  
Team_1987Phi.    0.6039423  0.4661302   1.296 0.196783    
Team_1987Pit.   -0.3341451  0.4870957  -0.686 0.493614    
Team_1987S.D.    1.3874263  0.4838814   2.867 0.004643 ** 
Team_1987S.F.    1.2714366  0.6556587   1.939 0.054070 .  
Team_1987Sea.    2.1670257  1.1286569   1.920 0.056465 .  
Team_1987St.L.   0.2137495  0.5278683   0.405 0.686018    
Team_1987Tex.    1.2232273  0.7887488   1.551 0.122725    
Team_1987Tor.    1.1013703  1.1144053   0.988 0.324353    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5002 on 177 degrees of freedom
Multiple R-squared:  0.784,	Adjusted R-squared:  0.6791 
F-statistic: 7.471 on 86 and 177 DF,  p-value: < 2.2e-16
```

Therefore, from a full model, we retain the intercept, AB_1986, H_1986, HR_1986, BB_1986, Years, W_Career, various Team_1986 variables, various Position_1986 variables, Put_outs_1986 and various Team_1987 variables. However, note that we would like to measure the qualitative variable in its entirety, rather than separated as displayed above. To do this, we used the ANOVA method to see if the qualitative variable is significant under the F statistic under the 5% significance level.

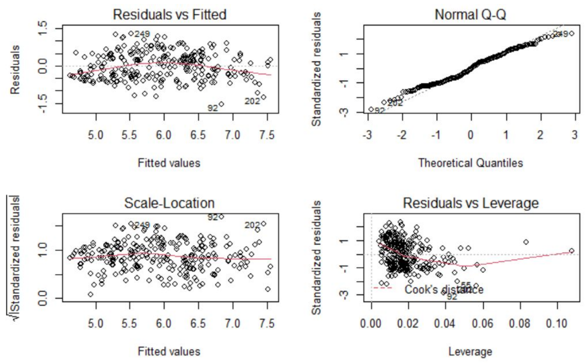
1. Comparing our current model with and without the Position_1986 factor. We see that the F-statistic is 1.4063 and is not significant. We therefore exclude the variable Position_1986 from our model.

```
Analysis of Variance Table

Model 1: log(Salary) ~ AB_1986 + H_1986 + HR_1986 + BB_1986 + Years + 
    W_career + Put_outs_1986 + factor(Team_1986) + factor(Team_1987)
Model 2: log(Salary) ~ AB_1986 + H_1986 + HR_1986 + BB_1986 + Years + 
    W_career + Put_outs_1986 + factor(Team_1986) + factor(Team_1987) + 
    factor(Position_1986)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    210 58.084                           
2    189 50.235 21    7.8495 1.4063 0.1191
```

2. Comparing our current model with and without the Team_1986 factor. We repeat this process and see that the F-statistic is 1.4073 and is not significant. We therefore exclude the variable Team_1986 from our model.

3. Comparing our current model with and without the Team_1987 factor. We see that the F-statistic is 1.7398 and is significant with a p-value of 0.02218. We therefore include the variable Team_1987 in our model.

Analysing residuals and diagnostics of our current linear model, we notice that there are some significant outliers skewing our data. We will remove these values - notably 202, 92 and 249 and re-run our summary.




```
Call:
lm(formula = log(Salary) ~ AB_1986 + H_1986 + BB_1986 + Years,
    data = baseball)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20710 -0.42674 -0.01388  0.38910  1.27902

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.0744265  0.1191081  34.208  < 2e-16 ***
AB_1986     -0.0019844  0.0008983  -2.209   0.0281 *
H_1986       0.0142313  0.0028208   5.045 8.73e-07 ***
BB_1986      0.0083843  0.0020067   4.178 4.07e-05 ***
Years        0.1055972  0.0072433  14.579  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5371 on 250 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6395
F-statistic: 113.6 on 4 and 250 DF,  p-value: < 2.2e-16
```

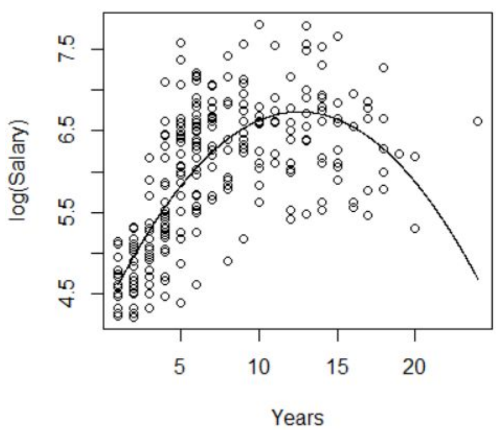
We are now left with our final model: log(Salary) ~ AB_1986+H_1986+BB_1986+Years. However, this model is intuitively flawed as Salary is only dependent on statistics that were generated in one year (1986) and the number of years (this is shown to be a better fit under a quadratic function) spent in major leagues. A more well-rounded approach considering hits and bats in the length of a career may be a better predictor.

Generalised Linear Model

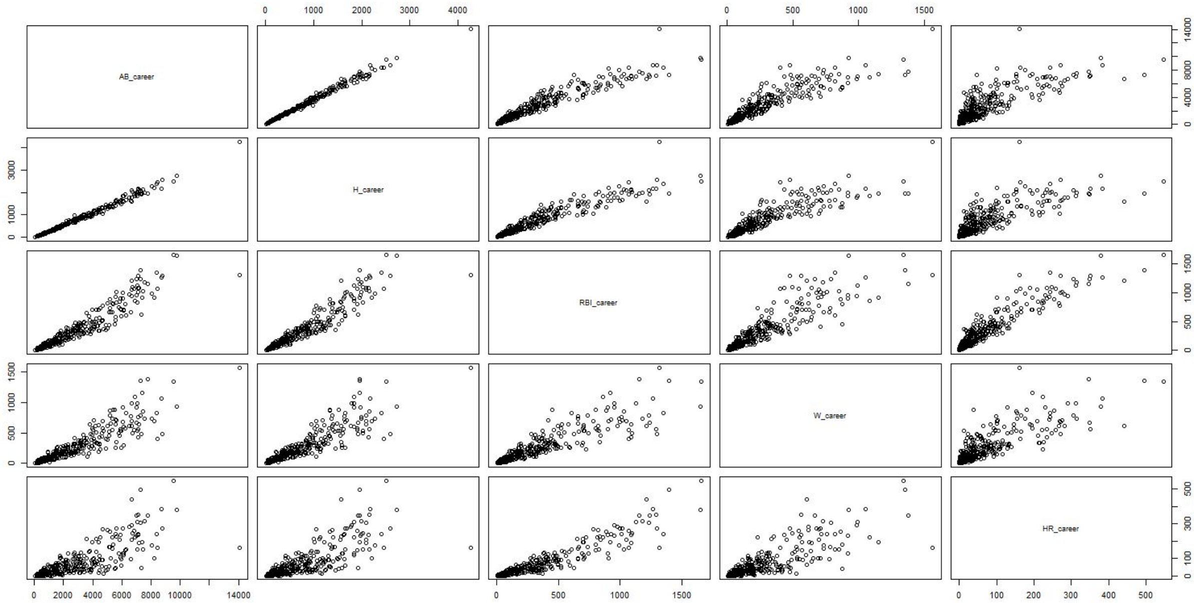
We considered the generalised linear model (GLM) next. We continued to assume that log(Salary) would be an accurate transformation as salaries were observed to have an exponential relationship with many of the predictors. The following assumptions were made about the GLM:

1. Independence of data points
2. Correct distribution of residuals
3. Correct specification of variance
4. Linear relationship between response and linear predictor

Next, we tested years against log(Salary) which did not appear to have a clear linear relationship. Instead, a parabolic relationship was observed and a polynomial for years was a better fit, potentially due to the notion that players peak in the middle of their careers as they experience an increase in salary.



We also observe that H_career and HR_1986 both have a significant outlier each, so we remove these to prevent overfitting of the model.



As seen above, there is a strong correlation between the AB (at bat) variables for ‘career’ (this is the same case with ‘1986’) and their respective performance variables (H, HR, RBI etc). This is consistent with our expectations as the more batting opportunity a player has, the more hits and runs they will score. Thus, we decided to divide the performance variables by AB to obtain proportion variables, thus gauging the true performance of baseball players. We can justify this method by looking at statistics that are valued highly by baseball statisticians, most notably BA or Batting Average, which calculates the proportion of Hits to At Bats. This is important as a player with more hits but proportionally more At Bats is deemed not to perform as well as a player with less hits but proportionally less At Bats.

With these new variables, we see H_1986_prop (hits) and BB_1986_prop (walks) are key predictors of salary as they are under the 0.05 significance level. Testing multicollinearity, both predictors have a Variation Inflation Factor (VIF) of 1.0027, suggesting that they are not correlated.

```
Call:
glm(formula = log(Salary) ~ H_1986_prop + HR_1986_prop + RBI_1986_prop +
    R_1986_prop + BB_1986_prop, data = baseball)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1514  -0.6252   0.1000   0.6282   1.5911

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.5126     0.4779   5.257 3.07e-07 ***
H_1986_prop    10.2392     2.1162   4.839 2.25e-06 ***
HR_1986_prop    2.3255     5.2315   0.445  0.65705
RBI_1986_prop   2.1626     2.3363   0.926  0.35548
R_1986_prop     0.1182     2.3480   0.050  0.95990
BB_1986_prop    3.6514     1.3772   2.651  0.00852 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.6493228)

    Null deviance: 205.09  on 263  degrees of freedom
Residual deviance: 167.53  on 258  degrees of freedom
  (60 observations deleted due to missingness)
AIC: 643.13

Number of Fisher Scoring iterations: 2
```

```
vif(test2)
H_1986_prop BB_1986_prop
   1.002708     1.002708
```

Similarly, we test the next variables to find that W_career (wins) and H_career_prop (hits in career) are both significant.

```
Call:
glm(formula = log(Salary) ~ poly(Years, 2, raw = TRUE) + H_career_prop +
    RBI_career_prop + W_career + HR_career_prop, data = baseball)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.23433  -0.33380   0.02025   0.27831   1.42391

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.8689146  0.3233279   2.687  0.00767 **
poly(Years, 2, raw = TRUE)1 0.3426267  0.0212369  16.134  < 2e-16 ***
poly(Years, 2, raw = TRUE)2 -0.0177674 0.0011192 -15.875  < 2e-16 ***
H_career_prop              12.7727431  1.3652615   9.356  < 2e-16 ***
RBI_career_prop             0.3258381  1.8531195   0.176  0.86056
W_career                    0.0015248  0.0001985   7.681 3.32e-13 ***
HR_career_prop              4.5439201  3.7356276   1.216  0.22496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1973225)

    Null deviance: 205.087  on 263  degrees of freedom
Residual deviance:  50.712  on 257  degrees of freedom
  (60 observations deleted due to missingness)
AIC: 329.66

Number of Fisher Scoring iterations: 2
```

Continuing from the linear model, we opt to exclude positions from our final model as they were not found to have a significant contribution towards salary.

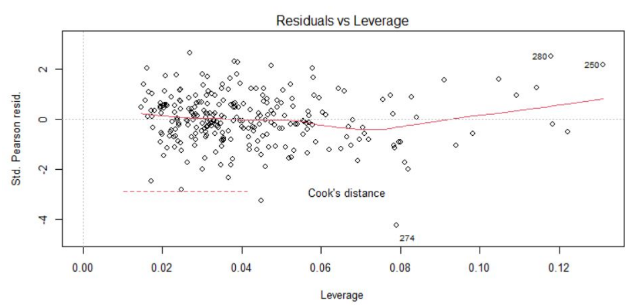
We then test the statistics which are not related to runs i.e. putouts, assists, errors. We observe all three variables have p-value < 0.05, however assists and errors have a moderately high VIF levels, indicating multicollinearity between the two. We keep them in the model temporarily but consider removing them later if proven to be insignificant. Similarly, we deem vision to be significant and include it in our final GLM.

```
Call:
glm(formula = log(Salary) ~ Put_outs_1986 + Assists_1986 + Errors_1986,
    data = baseball)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8383  -0.5832   0.1136   0.6440   1.9226

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.7763964  0.0999962  57.766  < 2e-16 ***
Put_outs_1986 0.0007742  0.0001912   4.049  6.8e-05 ***
Assists_1986  0.0011864  0.0005164   2.297   0.0224 *
Errors_1986  -0.0231291  0.0113766  -2.033   0.0431 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.7334625)

    Null deviance: 205.09  on 263  degrees of freedom
Residual deviance: 190.70  on 260  degrees of freedom
  (60 observations deleted due to missingness)
AIC: 673.33

Number of Fisher Scoring iterations: 2
```

```
> vif(test_6)
Put_outs_1986  Assists_1986   Errors_1986
     1.024056      2.010122      2.017950
```

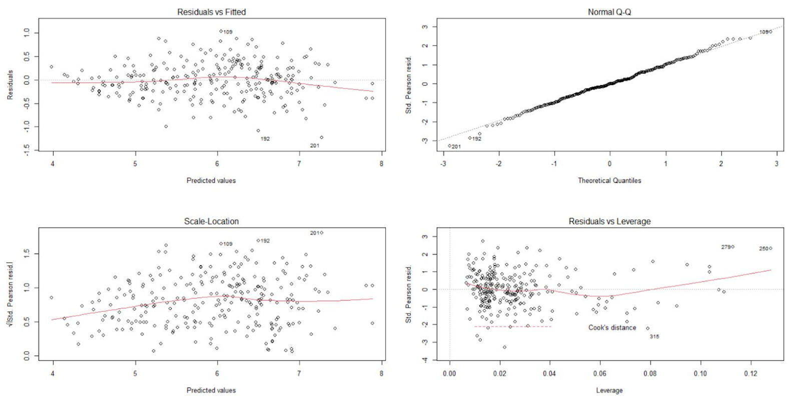
```
Call:
glm(formula = log(Salary) ~ Div_1986, data = baseball)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8446  -0.6851   0.1754   0.7276   1.7512

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.05671    0.07697  78.694   <2e-16 ***
Div_1986W   -0.22450    0.10803  -2.078   0.0387 *
---
```

In the final model, we observe that some variables are no longer useful predictors and have high multicollinearity. We remove these as they are comparably a less accurate fit. By analysing the diagnostic plots, it appears the 274th observation is skewing results and we have therefore chosen to remove it. Our finalised model is produced by using a stepwise function.




```
Call:
glm(formula = log(Salary) ~ BB_1986_prop + poly(Years, 2, raw = TRUE) +
    H_career_prop + I(W_career/Years) + Put_outs_1986, data = baseball)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.55276  -0.24922  -0.00173   0.25270   1.02355

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.5899338  0.3101374   5.127 5.82e-07 ***
BB_1986_prop               -3.0431204  0.7840000  -3.882 0.000132 ***
poly(Years, 2, raw = TRUE)1 0.4113333  0.0202068  20.356  < 2e-16 ***
poly(Years, 2, raw = TRUE)2 -0.0186314 0.0010781 -17.282  < 2e-16 ***
H_career_prop               8.5570363  1.2001359   7.130 1.02e-11 ***
I(W_career/Years)           0.0217887  0.0020518  10.619  < 2e-16 ***
Put_outs_1986               0.0003632  0.0000891   4.076 6.13e-05 ***
---
```

From our diagnostic plots, we see that the assumption of equal error variance holds as the line is considerably straight. The QQ plot indicates that our normality assumption is valid and the scale-location plot indicates there is homoscedasticity with equal randomly spread points along the predicted values. Finally, the last plot shows that all observations are within Cook's distance lines and therefore there are no influential outliers.



Generalised Additive Model

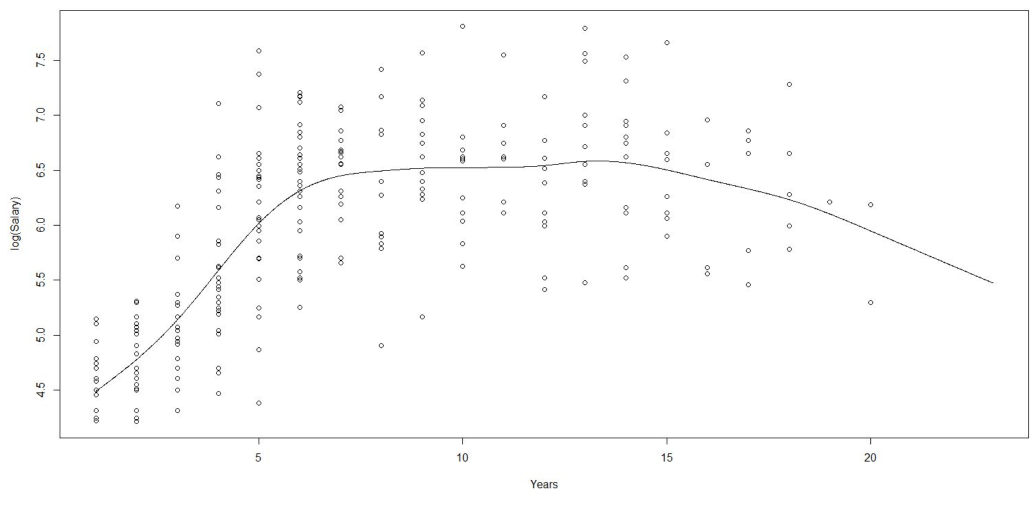
Continuing from our discoveries from the generalised linear model, we can see from the previous pairs plot that our predictive variables have a range of underlying patterns, some of which are nonlinear. We therefore explore the possibility of using a generalised additive model (GAM) to capture all the impacts of our variables through smoothing functions.

In a generalised additive model, we assume that

$$g(\mu_i) = \sum_{j=1}^{p} f_j(x_{ij})$$

Where the $f_j(.)$'s are a collection of smooth univariate functions and the responses are independent with a density or probability function from the exponential family. We also assume that the mean of errors is 0 and the variance is a constant of σ^2.

From the linear model, we note that years is a significant variable both intuitively and statistically to include in the model, however it does not appear to have a linear relationship with log(Salary). Therefore we choose to smooth the variable years in our GAM.



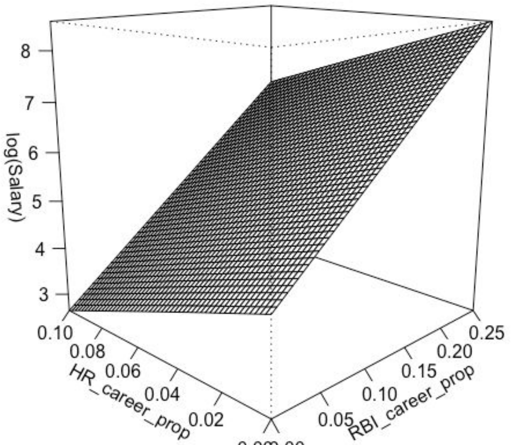
Furthermore, after testing the significance of smoothed variables recorded within the year 1986, we find that only H_1986_prop and BB_1986_prop are significant. For career-wide variables, years (as aforementioned), H_career_prop, RBI_career_prop and W_career appear to be significant. Similarly, we concluded that Put_outs_1986 was also a significant variable to include in our final model.

```
Family: gaussian
Link function: identity

Formula:
log(Salary) ~ s(H_1986_prop) + s(HR_1986_prop) + s(RBI_1986_prop) +
    s(R_1986_prop) + s(BB_1986_prop)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.94529    0.04936   120.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                  edf Ref.df      F  p-value
s(H_1986_prop)   1.000   1.00 22.325 3.74e-06 ***
s(HR_1986_prop)  1.867   2.36  0.951  0.33495
s(RBI_1986_prop) 1.000   1.00  0.637  0.42570
s(R_1986_prop)   1.000   1.00  0.029  0.86461
s(BB_1986_prop)  1.000   1.00  6.995  0.00867 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.179   Deviance explained = 19.7%
GCV = 0.65545  Scale est. = 0.63827   n = 262
```

```
Family: gaussian
Link function: identity

Formula:
log(Salary) ~ s(Years) + s(H_career_prop) + s(RBI_career_prop) +
    s(W_career) + s(HR_career_prop) + s(R_career_prop)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.94529    0.02106   282.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                     edf Ref.df      F  p-value
s(Years)           8.595  8.930 17.081  < 2e-16 ***
s(H_career_prop)   2.730  3.487 20.081 8.93e-13 ***
s(RBI_career_prop) 2.492  3.173  2.485   0.0585 .
s(W_career)        6.049  7.174 18.113  < 2e-16 ***
s(HR_career_prop)  2.598  3.295  1.311   0.2600
s(R_career_prop)   1.751  2.243  1.658   0.1936
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.851   Deviance explained = 86.4%
GCV = 0.12853  Scale est. = 0.11616   n = 262
```

Given the nature of the variables involved in the data set, we investigate the possibility of an interaction between the significant variables HR_career_prop and RBI_career_prop. Intuitively, we hypothesise that runs batted in by a hitter can depend on home runs since a home run allows the player to bat in the outstanding players on the bases.

From the two-dimensional scatter plot below we can confirm there is interaction between HR_career_prop and RBI_career_prop, however after attempting to test their significance in a GAM, it does not appear to have a significant contribution to the GAM.



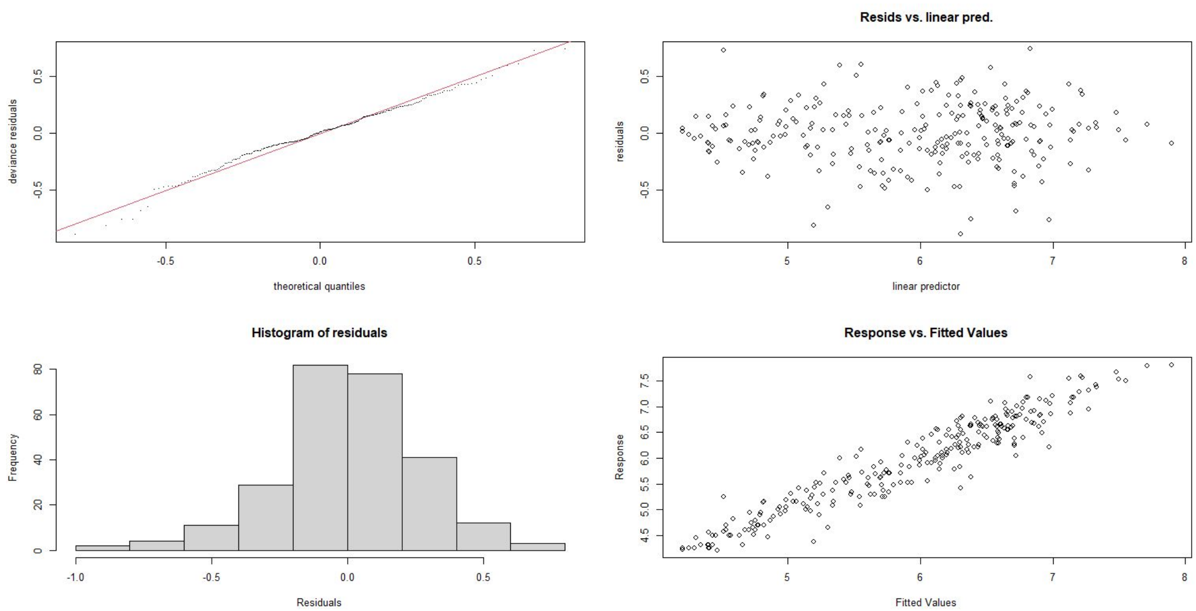

```
Family: gaussian
Link function: identity

Formula:
log(Salary) ~ s(H_1986_prop, k = -1) + s(BB_1986_prop, k = -1) +
    s(Years, k = -1) + s(H_career_prop, k = -1) + s(RBI_career_prop,
    k = -1) + s(W_career, k = -1) + s(I(RBI_career_prop * HR_career_prop)) +
    Put_outs_1986

Parametric coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.873e+00  3.145e-02 186.711   <2e-16 ***
Put_outs_1986 2.413e-04  8.207e-05   2.941   0.0036 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                         edf Ref.df      F  p-value
s(H_1986_prop)                         1.000  1.000  0.194    0.660
s(BB_1986_prop)                        4.047  5.036  6.539 9.71e-06 ***
s(Years)                               4.387  5.384 23.351  < 2e-16 ***
s(H_career_prop)                       2.079  2.675 12.307 1.06e-06 ***
s(RBI_career_prop)                     2.128  2.760  1.023    0.376
s(W_career)                            6.384  7.492 18.860  < 2e-16 ***
s(I(RBI_career_prop * HR_career_prop)) 4.861  5.952  0.897    0.488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.857   Deviance explained = 87.1%
GCV = 0.12393  Scale est. = 0.11131   n = 264
```

Given the limitations of interactions in additive models, we have chosen to exclude the interaction from our final GAM. However, we are also aware that the assumption of no interaction may be restrictive of our model.

Finally, after accounting for the insignificant variables, we arrive at our final GAM as shown below. We have used $k = -1$ such that the GAM automatically uses GCV to choose the optimal number of knots for our smoothed variable.

```
Family: gaussian
Link function: identity

Formula:
log(Salary) ~ s(BB_1986_prop, k = -1) + s(Years, k = -1) + s(H_career_prop,
    k = -1) + s(RBI_career_prop, k = -1) + s(W_career, k = -1) +
    s(Put_outs_1986, k = -1)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.94276    0.02026   293.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df      F  p-value
s(BB_1986_prop)    3.636  4.567  7.073 8.32e-06 ***
s(Years)           4.923  5.962 23.247  < 2e-16 ***
s(H_career_prop)   2.171  2.793 21.849 1.57e-11 ***
s(RBI_career_prop) 2.594  3.272  8.133 2.20e-05 ***
s(W_career)        6.528  7.610 20.001  < 2e-16 ***
s(Put_outs_1986)   2.256  2.790  5.416  0.00142 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.861   Deviance explained = 87.3%
GCV = 0.11878  Scale est. = 0.10838   n = 264
```

Our diagnostic plots show no significant violations of the mean-variance assumption and our residuals show no notable outliers. Our response plotted against fitted values appears to be linear indicating the final GAM is effective.

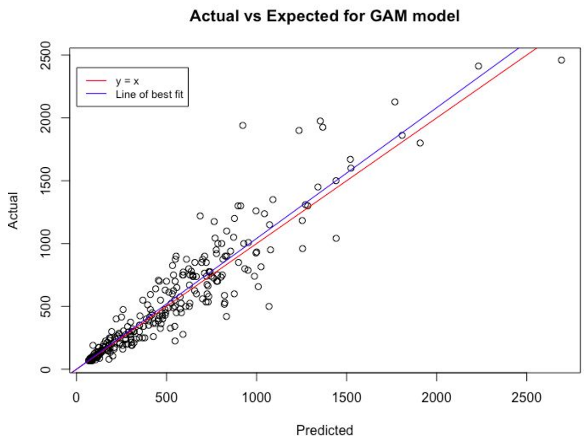


Conclusion

Our report sought to derive the key performance drivers of baseball hitters and how they contribute to their salaries through three different attempted models: linear model, generalised linear model, and generalised additive model.

Overall, we find that these variables were significant across at least two of the models: number of walks in 1986 (BB_1986), years spent in the major leagues (years), number of hits during their career (H_career), and put outs in 1986 (Put_outs_1986). Therefore, we deem these predictor variables to be important performance drivers that contribute to a baseball hitter’s salary in 1986.

When deciding on which model best determines the key performance drivers for a player's salary, we decided to look at predictive accuracy as well as which model provides the best inference. As the Year variable has a non-linear relationship with log(salary), we do not think it is wise to use a linear model for our statistical analysis.



When looking at the actual vs expected plots above, despite the line of best fit being more accurate for the Generalised Linear Model, we can see that the Generalised Additive Model is a better model for predicting salaries as there is overall less variation around the line. This is expected, as GAM allows for smoothing which reduces overall variance, while GLM does not allow for smoothing apart from polynomial treatment, thus resulting in greater variance in predictions.

However, it is important to consider that the simplicity of the GLM makes it easier to infer the impact of each significant performance driver on salary (such as the quadratic relationship of Years against Log(Salary)), while it is harder to infer any conclusions from the smoothing conducted by the GAM. Furthermore, using a GAM encourages overfitting to the current data, which may not be suitable when predicting future salaries.

Appendix

```
#Loading necessary packages
library(tidyverse)
library(magrittr)
library(readxl)
library(mgcv)
library(car)

#Reading data set
baseball <- read_excel("Baseball.xlsx", 1)
baseball$Salary %<>% as.numeric()

#Linear model

#Correlation coefficient values for linear model between quantitative variables
cor(baseball$AB_1986, baseball$Salary)
cor(baseball$H_1986, baseball$Salary)
cor(baseball$HR_1986, baseball$Salary)
cor(baseball$R_1986, baseball$Salary)
cor(baseball$RBI_1986, baseball$Salary)
cor(baseball$BB_1986, baseball$Salary)
cor(baseball$Years, baseball$Salary)
cor(baseball$AB_career, baseball$Salary)
cor(baseball$H_career, baseball$Salary)
cor(baseball$HR_career, baseball$Salary)
cor(baseball$R_career, baseball$Salary)
cor(baseball$RBI_career, baseball$Salary)
cor(baseball$W_career, baseball$Salary)
cor(baseball$Put_outs_1986, baseball$Salary)
cor(baseball$Assists_1986, baseball$Salary)
cor(baseball$Errors_1986, baseball$Salary)

#Correlation coefficient values for log-inear model between quantitative variables
cor(baseball$AB_1986, log(baseball$Salary))
cor(baseball$H_1986, log(baseball$Salary))
cor(baseball$HR_1986, log(baseball$Salary))
cor(baseball$R_1986, log(baseball$Salary))
cor(baseball$RBI_1986, log(baseball$Salary))
cor(baseball$BB_1986, log(baseball$Salary))
cor(baseball$Years, log(baseball$Salary))
cor(baseball$AB_career, log(baseball$Salary))
cor(baseball$H_career, log(baseball$Salary))
cor(baseball$HR_career, log(baseball$Salary))
cor(baseball$R_career, log(baseball$Salary))
cor(baseball$RBI_career, log(baseball$Salary))
cor(baseball$W_career, log(baseball$Salary))
cor(baseball$Put_outs_1986, log(baseball$Salary))
cor(baseball$Assists_1986, log(baseball$Salary))
cor(baseball$Errors_1986, log(baseball$Salary))

#Running our model and selecting significant variables
log.salary <- lm(log(Salary)~., data = baseball, na.action=na.omit)
```

```r
summary(log.salary)

#Checking position_1986 significance in ANOVA
anova(lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         +factor(Team_1986)+factor(Team_1987), data=baseball),
      lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         +factor(Team_1986)+factor(Team_1987)+factor(Position_1986), data=baseball))

#Checking significance of Team_1986 in ANOVA
anova(lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         +factor(Team_1987), data=baseball),
      lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         +factor(Team_1986)+factor(Team_1987), data=baseball))

#Checking significance of Team_1987 in ANOVA
anova(lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         , data=baseball),
      lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+Put_outs_1986
         +factor(Team_1987), data=baseball))

#Rerunning our model and removing some more variables
log.final_check <- lm(log(Salary)~AB_1986+H_1986+HR_1986+BB_1986+Years+W_career+
                        Put_outs_1986+factor(Team_1987), data=baseball)
summary(log.final_check)

#Final model
log.final <- lm(log(Salary)~AB_1986+H_1986+BB_1986+Years, data=baseball)

#Residuals analysis
par(mfrow=c(2,2))
plot(log.final)
```

```r
#Generalised linear model

#Removing qualitative variables
baseball %<>% select(-Name,
                     -League_1986,
                     -Team_1986,
                     -League_1987,
                     -Team_1987)

baseball %<>% select(Salary, Years, everything())

#Polynomial plot of years against log(salary)
plot(log(Salary) ~ Years, data = baseball)
years.glm <- glm(log(Salary) ~ poly(Years, 2, raw = TRUE), data = baseball)
test.years <- data.frame(Years = seq(min(baseball$Years), max(baseball$Years), 0.01))
predictedsalary <- predict(years.glm, test.years)
lines(test.years$Years, predictedsalary)

# Remove outliers
baseball %<>% filter(HR_1986 < 100, H_career < 4000)
```

```r
#Testing interaction variables
pairs(~ AB_career + H_career + RBI_career + W_career + HR_career, data = baseball)
```

```r
pairs(~ AB_1986 + H_1986 + HR_1986 + R_1986 + RBI_1986 + BB_1986, data = baseball)
```

```r
#### New variables ####
baseball %<>%
  mutate(H_1986_prop = H_1986/AB_1986,
         HR_1986_prop = HR_1986/AB_1986,
         R_1986_prop = R_1986/AB_1986,
         RBI_1986_prop = RBI_1986/AB_1986,
         BB_1986_prop = BB_1986/AB_1986,
         BA_1986 = H_1986_prop + BB_1986_prop,
         H_career_prop = H_career/AB_career,
         HR_career_prop = HR_career/AB_career,
         R_career_prop = R_career/AB_career,
         RBI_career_prop = RBI_career/AB_career)

#Testing new variables' significance
test1 <- glm(log(Salary) ~ H_1986_prop + HR_1986_prop + RBI_1986_prop + R_1986_prop
             + BB_1986_prop, data = baseball)
test2 <- glm(log(Salary) ~ H_1986_prop + BB_1986_prop,
            data = baseball)
summary(test1)
vif(test2)

## Test career variables
test_3 <- glm(log(Salary) ~ poly(Years, 2, raw = TRUE) + H_career_prop + RBI_career_prop +
                W_career + HR_career_prop,
              data = baseball)
summary(test_3)

#Other variables
test_6 <- glm(log(Salary) ~ Put_outs_1986 + Assists_1986 + Errors_1986, data = baseball)
summary(test_6)
vif(test_6)

#Removal of 274th observation
baseball <- baseball[-274,]

#Final model
glm.salary <- glm(log(Salary) ~
                    H_1986_prop +
                    BB_1986_prop +
                    poly(Years, 2, raw = TRUE) +
                    H_career_prop +
                    I(W_career/Years) +
                    Put_outs_1986 +
                    Assists_1986 +
                    Errors_1986 +
                    Div_1986,
                  data = baseball)
summary(glm.salary)
```

```r
par(mfrow=c(2,2))
plot(glm.salary)
```

```r
#Final stepwise function
glm.final <- step(glm.salary, direction = "both")
summary(glm.final)
glm.final <- glm(log(Salary) ~
                   BB_1986_prop +
                   poly(Years, 2, raw = TRUE) +
                   H_career_prop +
                   I(W_career/Years) +
                   Put_outs_1986,
                 data = baseball)
summary(glm.final)
par(mfrow=c(2,2))
plot(glm.final)
```

```r
#Generalised additive model

#Years
years.gcv = c()
for (i in c(3:20)) {
  years.gam <- gam(log(Salary) ~ s(Years, k = i), data = baseball)
  years.gcv[i-2] <- years.gam$gcv.ubre.dev
}
plot(years.gcv)
plot(log(Salary) ~ Years, data = baseball)
test.years <- data.frame(Years = seq(min(baseball$Years), max(baseball$Years), 0.01))
predictedsalary <- predict(years.gam, test.years)
lines(test.years$Years, predictedsalary)

## Test 1986 variables
test <- gam(log(Salary) ~
              s(H_1986_prop) +
              s(HR_1986_prop) +
              s(RBI_1986_prop) +
              s(R_1986_prop) +
              s(BB_1986_prop),
            data = baseball)

summary(test)

## Test career variables
test_2 <- gam(log(Salary) ~
              s(Years) +
              s(H_career_prop) +
              s(RBI_career_prop) +
              s(W_career) +
              s(HR_career_prop) +
              s(R_career_prop),
            data = baseball)

summary(test_2)
```

```r
## Other variables
test_3 <- gam(log(Salary) ~
                Put_outs_1986+
                s(Assists_1986) +
                s(Errors_1986),
              data = baseball)
summary(test_3)

test_3 <- gam(log(Salary) ~
                Div_1986,
              data = baseball)
summary(test_3)

#Testing interaction
grid <- list(RBI_career_prop = seq(from = 0, to = 0.25, length = 50),
             HR_career_prop = seq(from = 0, to = 0.1, length = 50))
baseball.gam <- gam(log(Salary) ~ s(RBI_career_prop) + s(HR_career_prop), data = baseball)
baseball.pr <- mgcv::predict.gam(baseball.gam, newdata = expand.grid(grid))
baseball.pr <- matrix(baseball.pr, nrow = 50, ncol = 50)
persp(grid$RBI_career_prop, grid$HR_career_prop, baseball.pr,
      xlab = "RBI_career_prop", ylab = "HR_career_prop",
      zlab = "log(Salary)", theta = -45, phi = 15, d = 2.0, tick = "detailed")

gam.salary <- gam(log(Salary) ~
                    s(H_1986_prop,k=-1) +
                    s(BB_1986_prop, k = -1) +
                    s(Years, k = -1) +
                    s(H_career_prop, k = -1) +
                    s(RBI_career_prop, k = -1) +
                    s(W_career, k = -1) +
                s(I(RBI_career_prop*HR_career_prop)),
                  data = baseball)
summary(gam.salary)

#Final GAM
gam.final <- gam(log(Salary) ~
                   s(BB_1986_prop,k=-1) +
                   s(Years, k = -1) +
                   s(H_career_prop, k = -1) +
                   s(RBI_career_prop, k = -1) +
                   s(W_career, k = -1)+
                   s(Put_outs_1986, k = -1),
                 data = baseball)
summary(gam.final)
gam.check(gam.final)
```

```r
plot(gam.final)
```

```r
#Conclusion plots
predicted.salary.final <- predict(glm.final, baseball)
plot(exp(predicted.salary.final), baseball$Salary,
     xlab = "Predicted",
     ylab = "Actual",
```

```
     main = "Actual vs Expected for GLM model")
x <- seq(0, 2500, 100)
y <- x
abline(lm(y ~ x), col = "red")
abline(lm(baseball$Salary ~ - 1 + exp(predicted.salary.final)), col = "blue")

legend(1, 2400, legend=c("y = x", "Line of best fit"),
       col=c("red", "blue"), lty=1, cex=0.8)

predicted <- predict(gam.final)
plot(predicted, log(baseball$Salary[which(!is.na(baseball$Salary))]))
```

```
plot(exp(predicted), baseball$Salary[which(!is.na(baseball$Salary))],
     xlab = "Predicted",
     ylab = "Actual",
     main = "Actual vs Expected for GAM model")
x <- seq(0, 2500, 100)
y <- x
abline(lm(y ~ x), col = "red")
abline(lm(baseball$Salary[which(!is.na(baseball$Salary))] ~ -1 + exp(predicted)),
       col = "blue")

legend(1, 2400, legend=c("y = x", "Line of best fit"),
       col=c("red", "blue"), lty=1, cex=0.8)
```