CS 624 Final

Chris Watkins

12/12/2017

Problem 1. (25 points) We are interested in comparing four characteristics (sepal length, sepal width, petal length and petal width) of three flower species Setosa, Versicolor and Virginica. We want to analyze the dataset "iris" available in the package MVN. Perform all necessary data analysis steps and write a section summarizing the findings.

First we want to verify that the data for each variable is normal before we proceed to any analysis.

```r
library(MVN)
```

```
## sROC 0.1-2 loaded
```

```r
flow <- iris
shapiro.test(flow$Sepal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  flow$Sepal.Length
## W = 0.97609, p-value = 0.01018
```

```r
shapiro.test(flow$Sepal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  flow$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

```r
shapiro.test(flow$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  flow$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

```r
shapiro.test(flow$Petal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  flow$Petal.Width
## W = 0.90183, p-value = 1.68e-08
```

Based on an alpha level of 0.05 and Shapiro-Wilk's tests, we fail to reject the null hypothesis that Sepal Width is normally distributed. However, we do reject the null hypothesis that Sepal Length, Petal Length, and Petal Width are normally distributed. So we can do an ANOVA for Sepal Width but must use the Kruskal-Wallis test to analyze the rest of the variables.

```r
summary(aov(Sepal.Width~Species, data = flow))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## Species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on an alpha level of 0.05 and the ANOVA, we reject the null hypothesis that the mean Sepal Width among groups are the same. Now we will find where the difference lie using Tukey's HSD.

```r
TukeyHSD(aov(Sepal.Width~Species, data = flow))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = Sepal.Width ~ Species, data = flow)
## 
## $Species
##                        diff         lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

Based on Tukey's HSD test, we reject the null hypothesis that there are not pairwise differences in means of sepal width. We can conclude the mean sepal width for each group are different.

Now we will look at the rest of the variables using the Kruskal-Wallis test.

```r
kruskal.test(flow$Sepal.Length, flow$Species)
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  flow$Sepal.Length and flow$Species
## Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

Based on the Kruskal-Wallis test and an alpha level of 0.05, we reject the null hypothesis that the median sepal length in each group are the same. To find where the differences are, we will use the Wilcoxon-Rank Sum test.

```r
wilcox.test(flow[flow$Species == "versicolor",]$Sepal.Length,
            flow[flow$Species == "setosa",]$Sepal.Length)
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  flow[flow$Species == "versicolor", ]$Sepal.Length and flow[flow$Species == "setosa", ]$Sepal.L
## W = 2331.5, p-value = 8.346e-14
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(flow[flow$Species == "virginica",]$Sepal.Length,
            flow[flow$Species == "setosa",]$Sepal.Length)
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
```

```
## data:  flow[flow$Species == "virginica", ]$Sepal.Length and flow[flow$Species == "setosa", ]$Sepal.Le
## W = 2461.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(flow[flow$Species == "virginica",]$Sepal.Length,
            flow[flow$Species == "versicolor",]$Sepal.Length)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "virginica", ]$Sepal.Length and flow[flow$Species == "versicolor", ]$Sepa
## W = 1974, p-value = 5.869e-07
## alternative hypothesis: true location shift is not equal to 0
```

Based on the Wilcoxon Rank Sum tests we reject the null hypothesis that there are no pairwise differences in median of sepal length. We can conclude that the median sepal length across groups are difference.

```r
kruskal.test(flow$Petal.Length, flow$Species)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  flow$Petal.Length and flow$Species
## Kruskal-Wallis chi-squared = 130.41, df = 2, p-value < 2.2e-16
```

Based on the Kruskal-Wallis test and an alpha level of 0.05, we reject the null hypothesis that the median petal length in each group are the same. To find where the differences are, we will use the Wilcoxon-Rank Sum test.

```r
wilcox.test(flow[flow$Species == "versicolor",]$Petal.Length,
            flow[flow$Species == "setosa",]$Petal.Length)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "versicolor", ]$Petal.Length and flow[flow$Species == "setosa", ]$Petal.L
## W = 2500, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(flow[flow$Species == "virginica",]$Petal.Length,
            flow[flow$Species == "setosa",]$Petal.Length)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "virginica", ]$Petal.Length and flow[flow$Species == "setosa", ]$Petal.Le
## W = 2500, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(flow[flow$Species == "virginica",]$Petal.Length,
            flow[flow$Species == "versicolor",]$Petal.Length)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "virginica", ]$Petal.Length and flow[flow$Species == "versicolor", ]$Peta
## W = 2455.5, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Based on the Wilcoxon Rank Sum tests we reject the null hypothesis that there are no pairwise differences in median of petal length. We can conclude that the median petal length across groups are difference.

```
kruskal.test(flow$Petal.Width, flow$Species)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  flow$Petal.Width and flow$Species
## Kruskal-Wallis chi-squared = 131.19, df = 2, p-value < 2.2e-16
```

Based on the Kruskal-Wallis test and an alpha level of 0.05, we reject the null hypothesis that the median petal width in each group are the same. To find where the differences are, we will use the Wilcoxon-Rank Sum test.

```
wilcox.test(flow[flow$Species == "versicolor",]$Petal.Width,
            flow[flow$Species == "setosa",]$Petal.Width)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "versicolor", ]$Petal.Width and flow[flow$Species == "setosa", ]$Petal.Wi
## W = 2500, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(flow[flow$Species == "virginica",]$Petal.Width,
            flow[flow$Species == "setosa",]$Petal.Width)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "virginica", ]$Petal.Width and flow[flow$Species == "setosa", ]$Petal.Wid
## W = 2500, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(flow[flow$Species == "virginica",]$Petal.Width,
            flow[flow$Species == "versicolor",]$Petal.Width)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  flow[flow$Species == "virginica", ]$Petal.Width and flow[flow$Species == "versicolor", ]$Petal
## W = 2451, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Based on the Wilcoxon Rank Sum tests we reject the null hypothesis that there are no pairwise differences in median of petal width We can conclude that the median petal length across groups are difference.

Summary

Based on this analysis we can say that the mean sepal width is different across groups and the median sepal length, petal length, and petal width are difference across groups.

Problem 2. (25 points) We are interested in finding the important predictors of accumulated wealth at the time of retirement, assess their adjusted effect sizes (in direction and magnitude) and use the best linear regression model for interpretation and prediction. We want to analyze the Pension.txt dataset (available on Blackboard) that contains 194 observations on 17 variables: pyears - years of employment, prftshr - indicator for profit sharing company, choice - indicator for company giving a choice to contribute, female, married, age, educ - years of education, finc25, finc35, finc50, finc75, finc100, finc101- indicators for 25, 35, 50, 75, 100 and 101 levels of retirement contribution, wealth89 - wealth in thousands of dollars, race, stckin89 - percent of the portfolio in stock, irain89 - percent of the portfolio in IRA. Perform all necessary data analysis steps and write a section summarizing the findings.

```r
 pension <-read.table("~ChrisWatkins/Desktop/Biostats(Grad)/Data Sets/Pension.txt"
                       , header = T)
head(pension)
```

```
##   pyears prftshr choice female married age educ finc25 finc35 finc50
## 1      1       0      1      0       1  64   12      0      0      1
## 2      6       1      1      1       1  56   13      0      0      0
## 3     25       1      1      0       1  56   12      0      0      0
## 4     20       1      0      1       1  63   12      1      0      0
## 5     35       0      1      0       1  67   12      0      1      0
## 6     13       1      0      0       1  64   11      0      0      0
##   finc75 finc100 finc101 wealth89 race stckin89 irain89
## 1      0       0       0   77.900    0        1       1
## 2      1       0       0  154.900    0        1       1
## 3      1       0       0  154.900    0        1       1
## 4      0       0       0  232.500    0        1       1
## 5      0       0       0  179.000    0        0       1
## 6      1       0       0  120.025    0        1       0
```

```r
summary(pension)
```

```
##      pyears         prftshr           choice           female      
##  Min.   : 0.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.: 4.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median : 9.0   Median :0.0000   Median :1.0000   Median :1.0000  
##  Mean   :11.3   Mean   :0.2062   Mean   :0.6134   Mean   :0.6031  
##  3rd Qu.:16.0   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :45.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##  NA's   :3                                                        
##     married            age             educ           finc25      
##  Min.   :0.0000   Min.   :54.00   Min.   : 8.00   Min.   :0.0000  
##  1st Qu.:1.0000   1st Qu.:57.00   1st Qu.:12.00   1st Qu.:0.0000  
```

```
##  Median :1.0000   Median :60.00   Median :12.00   Median :0.0000
##  Mean   :0.7577   Mean   :60.48   Mean   :13.57   Mean   :0.2062
##  3rd Qu.:1.0000   3rd Qu.:64.00   3rd Qu.:16.00   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :73.00   Max.   :18.00   Max.   :1.0000
##
##      finc35           finc50           finc75          finc100
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.000
##  Mean   :0.1753   Mean   :0.2371   Mean   :0.134   Mean   :0.134
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.000
##
##     finc101          wealth89           race           stckin89
##  Min.   :0.00000   Min.   :-580.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:  65.45   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median : 140.00   Median :0.0000   Median :0.0000
##  Mean   :0.06186   Mean   : 207.37   Mean   :0.1134   Mean   :0.3402
##  3rd Qu.:0.00000   3rd Qu.: 251.00   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1485.00   Max.   :1.0000   Max.   :1.0000
##
##     irain89
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5155
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Looking at the data it looks like it is in a form ready for fitting a linear model.

```r
library(MASS)
pension <- na.omit(pension)
base.pension <- lm(wealth89~., data = pension)
best.pension1 <- stepAIC(base.pension, direction = "both",
                         trace = FALSE, data = pension)
```

After using step AIC for a main effects model, we find that age, finc75, finc100, finc101, stckin89, and irain89 are significant. However, the R squared value is only 0.29. I will now look at 2-way interations and the square of age.

```r
best.pension2 <- stepAIC(base.pension, ~.^2 + I(age^2),
                          direction = "both", trace = FALSE, date = pension)
summary(best.pension2)
```

```
##
## Call:
## lm(formula = wealth89 ~ pyears + prftshr + choice + female +
##     married + age + educ + finc50 + finc75 + finc100 + finc101 +
##     race + stckin89 + irain89 + married:finc75 + age:finc101 +
##     female:finc101 + pyears:finc50 + finc101:irain89 + female:irain89 +
##     prftshr:finc100 + female:finc100 + female:age + pyears:age +
##     choice:finc50 + prftshr:finc75 + prftshr:finc101 + educ:finc101 +
##     pyears:finc101 + choice:finc101 + finc101:stckin89 + married:stckin89 +
##     age:finc100 + educ:finc50 + race:irain89, data = pension)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -277.51  -83.03   -4.97   51.31  967.72
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1357.2558   490.3856  -2.768  0.00633 **
## pyears              50.1361    22.0609   2.273  0.02442 *
## prftshr             -5.0036    36.0069  -0.139  0.88966
## choice              11.6297    32.0365   0.363  0.71709
## female            1098.6495   427.7513   2.568  0.01116 *
## married             23.0794    41.2837   0.559  0.57694
## age                 20.2261     7.8804   2.567  0.01122 *
## educ                13.9549     6.1570   2.266  0.02480 *
## finc50             310.1337   182.1977   1.702  0.09073 .
## finc75             903.9643   169.5816   5.331 3.40e-07 ***
## finc100          -1055.6358   625.1455  -1.689  0.09330 .
## finc101          -4853.5572  2038.8081  -2.381  0.01850 *
## race               -37.8020    51.4382  -0.735  0.46351
## stckin89           -10.9083    60.4771  -0.180  0.85710
## irain89            119.0651    43.6835   2.726  0.00716 **
## married:finc75    -825.3649   172.1496  -4.794 3.79e-06 ***
## age:finc101        149.1101    26.5349   5.619 8.67e-08 ***
## female:finc101    1363.4674   268.4506   5.079 1.08e-06 ***
## pyears:finc50        6.4800     3.0576   2.119  0.03566 *
## finc101:irain89  -1105.2580   208.7101  -5.296 4.00e-07 ***
## female:irain89     -67.7683    52.7681  -1.284  0.20096
## prftshr:finc100    169.1147    98.1464   1.723  0.08687 .
## female:finc100     217.4788    74.5411   2.918  0.00405 **
## female:age         -17.6923     6.9409  -2.549  0.01178 *
## pyears:age          -0.8285     0.3561  -2.327  0.02128 *
## choice:finc50     -147.1176    59.2673  -2.482  0.01412 *
## prftshr:finc75    -174.3577    88.9518  -1.960  0.05177 .
## prftshr:finc101  -1714.9650   329.5719  -5.204 6.11e-07 ***
## educ:finc101      -374.6280    80.4911  -4.654 6.93e-06 ***
## pyears:finc101      70.5327    21.3444   3.304  0.00118 **
## choice:finc101     996.0279   248.1238   4.014 9.26e-05 ***
## finc101:stckin89   717.0916   245.7146   2.918  0.00404 **
## married:stckin89   176.3666    68.5414   2.573  0.01102 *
## age:finc100         16.5860    10.1157   1.640  0.10311
## educ:finc50        -21.2994    13.0153  -1.636  0.10377
## race:irain89      -109.9657    83.5664  -1.316  0.19015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 160.4 on 155 degrees of freedom
## Multiple R-squared:  0.6547, Adjusted R-squared:  0.5768
## F-statistic: 8.397 on 35 and 155 DF,  p-value: < 2.2e-16
```

After including 2-way interactions we find a model that has an R squared of 0.65, which is much better than the main effects model.

Summary

After fitting linear models for both main effects and including 2-way interactions, the best is with the 2-way interactions. Including 2-way interactions increases the R squared from 0.29 to 0.65, which is a large change. An interesting observation is that finc100 and finc101 were both significant and had positive effects on wealth in the main effects model but had large negative effects in the model that included 2-way interactions (Note: finc100 close to significant at 0.09). In fact, fin101 had the largest negative effect on wealth with a coefficient of -4853.56. This means that at the 101 level of retirement contribution, while keeping everything else constant, will decrease your wealth by 4853.56 at the time of retirement. The largest positive effect was the interaction variable between female and finc101 which had a coefficient of 1363.47 and a very significant p-value of 1.08e-6. It seems that being female and interaction terms with female increase wealth at retirement. There are a lot of interactions terms that would need to be studied more to assess what they would actually mean. For subsequent analyses I could engineer new variables. The variable pyears is very sparse, so I would group 0-10, 11-20, and >20. Age is pretty sparse above 65 so I would group <65 and >=65. I could also group those that went to college and did not go to college.

Problem 3. (25 points) We are interested in finding the important predictors of online customers booking a room at a hotel, assess their adjusted effect sizes (in direction and magnitude) and use the best logistic regression model for interpretation and prediction. We want to analyze the Travel.txt dataset (available on Blackboard) that contains 20,000 observations on 26 variables (description of all variables is presented in the Data_Dictionary_Travel file available on Blackboard). Perform all necessary data analysis steps and write a section summarizing the findings.

```r
travel <-read.table("~ChrisWatkins/Desktop/Biostats(Grad)/Data Sets/Travel.txt"
                    , header = T)
summary(travel)
```

```
##                date_time     user_location_region   user_location_city
##  2015-03-13 23:15:00:    3   CA     : 2924        NEW YORK    :  366
##  2015-05-27 19:37:00:    3   NY     : 1236        LOS ANGELES:  266
##  2015-06-09 13:11:00:    3   TX     : 1163        TORONTO     :  232
##  2015-06-10 17:31:00:    3   FL     : 1063        HOUSTON     :  229
##  2015-07-07 20:11:00:    3   ON     :  938        CHICAGO     :  215
##  2015-07-25 11:51:00:    3   (Other):12669        CALGARY     :  179
##  (Other)            :19982   NA's   :    7        (Other)     :18513
##  user_location_latitude       user_location_longitude
##  NULL     : 4242        NULL                 : 4242
##  40.75512 :  366        -73.98300900000001:  366
##  34.059768:  266        -118.312427       :  266
##  43.667179:  232        -79.390203        :  232
##  29.769607:  229        -95.42647         :  229
##  41.89042 :  215        -87.62904         :  215
##  (Other)  :14450        (Other)           :14450
##  orig_destination_distance    user_id             is_mobile
##  NULL    : 4242            Min.   :-2.147e+09   Min.   :0.0000
```

```
##  227.5021:   34            1st Qu.:-1.030e+09   1st Qu.:0.0000
##  0.6328  :   17            Median : 2.398e+07   Median :0.0000
##  0.1175  :   16            Mean   : 1.761e+07   Mean   :0.2233
##  342.2687:   16            3rd Qu.: 1.086e+09   3rd Qu.:0.0000
##  196.1892:   15            Max.   : 2.147e+09   Max.   :1.0000
##  (Other) :15660
##    is_package       channel          srch_ci            srch_co
##  Min.   :0.0000   Min.   :231.0   2015-09-04:  134   2015-09-07:  152
##  1st Qu.:0.0000   1st Qu.:293.0   2015-07-03:  124   2015-07-05:  140
##  Median :0.0000   Median :510.0   2015-09-05:  121   2015-07-26:  124
##  Mean   :0.1948   Mean   :418.6   2015-08-14:  117   2015-08-09:  123
##  3rd Qu.:0.0000   3rd Qu.:541.0   2015-07-31:  114   2015-08-30:  122
##  Max.   :1.0000   Max.   :541.0   2015-08-07:  114   2016-01-02:  120
##                                   (Other)   :19276   (Other)   :19219
##  srch_adults_cnt srch_children_cnt  srch_rm_cnt    srch_destination_id
##  Min.   :0.000   Min.   :0.0000    Min.   :0.000   Min.   :     8152
##  1st Qu.:2.000   1st Qu.:0.0000    1st Qu.:1.000   1st Qu.:  5527175
##  Median :2.000   Median :0.0000    Median :1.000   Median :  5626298
##  Mean   :2.056   Mean   :0.3108    Mean   :1.077   Mean   : 67753049
##  3rd Qu.:2.000   3rd Qu.:0.0000    3rd Qu.:1.000   3rd Qu.:187465121
##  Max.   :9.000   Max.   :8.0000    Max.   :8.000   Max.   :196871823
##
##                   hotel_country     is_booking         hotel_id
##  UNITED STATES OF AMERICA:12009   Min.   :0.00000   Min.   :      402
##  CANADA                  : 1141   1st Qu.:0.00000   1st Qu.:   725600
##  MEXICO                  : 1072   Median :0.00000   Median : 21533932
##  ITALY                   :  541   Mean   :0.08765   Mean   : 60301548
##  UNITED KINGDOM          :  426   3rd Qu.:0.00000   3rd Qu.: 77027722
##  FRANCE                  :  377   Max.   :1.00000   Max.   :410748015
##  (Other)                 : 4434
##  prop_is_branded  prop_starrating distance_band hist_price_band
##  Min.   :0.0000   Min.   :0.000   C :5130       H :4065
##  1st Qu.:0.0000   1st Qu.:3.000   F :2732       L :3873
##  Median :1.0000   Median :4.000   M :7631       M :8078
##  Mean   :0.6165   Mean   :3.528   VC:3155       VH:2108
##  3rd Qu.:1.0000   3rd Qu.:4.000   VF:1352       VL:1876
##  Max.   :1.0000   Max.   :5.000
##
##  popularity_band      cnt
##  H :5974         Min.   : 1.000
##  L : 721         1st Qu.: 1.000
##  M :5213         Median : 1.000
##  VH:7970         Mean   : 1.421
##  VL: 122         3rd Qu.: 1.000
##                  Max.   :38.000
##
```

```
names(travel)
```

```
##  [1] "date_time"                 "user_location_region"
##  [3] "user_location_city"        "user_location_latitude"
##  [5] "user_location_longitude"   "orig_destination_distance"
##  [7] "user_id"                   "is_mobile"
##  [9] "is_package"                "channel"
## [11] "srch_ci"                   "srch_co"
```

```
## [13] "srch_adults_cnt"           "srch_children_cnt"
## [15] "srch_rm_cnt"               "srch_destination_id"
## [17] "hotel_country"             "is_booking"
## [19] "hotel_id"                  "prop_is_branded"
## [21] "prop_starrating"           "distance_band"
## [23] "hist_price_band"           "popularity_band"
## [25] "cnt"
```

```
total_cnt <- travel$srch_adults_cnt+travel$srch_children_cnt
travel <- cbind(travel, total_cnt)
travel <- travel[,-c(1,2,3,4,5,6,7,11,12,13,14,15,16,17,19)]
travel$prop_starrating <- as.factor(travel$prop_starrating)
travel$channel <- as.factor(travel$channel)
travel <- travel[travel$total_cnt > 0,]
travel <- na.omit(travel)
```

After looking at the data there are several variables that I have decided to remove. These include date_time, all user location variables, origin destination distance, user id, check in and out date, number of adults, number of children, number of hotel rooms specified, destination id, hotel country, and hotel id. A lot of these I felt were irrelevant to the question of predicting the booking of a hotel. All ID information is not helpful because it does not tell us anything. The location of the user or destination does not matter because people book hotels from all over the world no matter where they are traveling. Also the check-in and check-out dates are not very helpful because people travel year round depending on their schedule. The number of hotel rooms is a continuous variable that is already redefined in the binary variable of booking that is our variable of interest for the logistic regression. I engineered a new variable total_cnt , which is the total number of people specified for a room rather than have two variables for adults and kids. I also made the star rating and channel a categorical variable rather than a continuous variable. I found that the total_cnt variable that I created had counts of 0 people in the room which makes no sense. I decided to remove these 13 people from the data set.

```
w <- sample(1:length(travel$is_booking), round(0.8*length(travel$is_booking)))
train <- travel[w,]
val <- travel[-w,]
library(pROC)
```

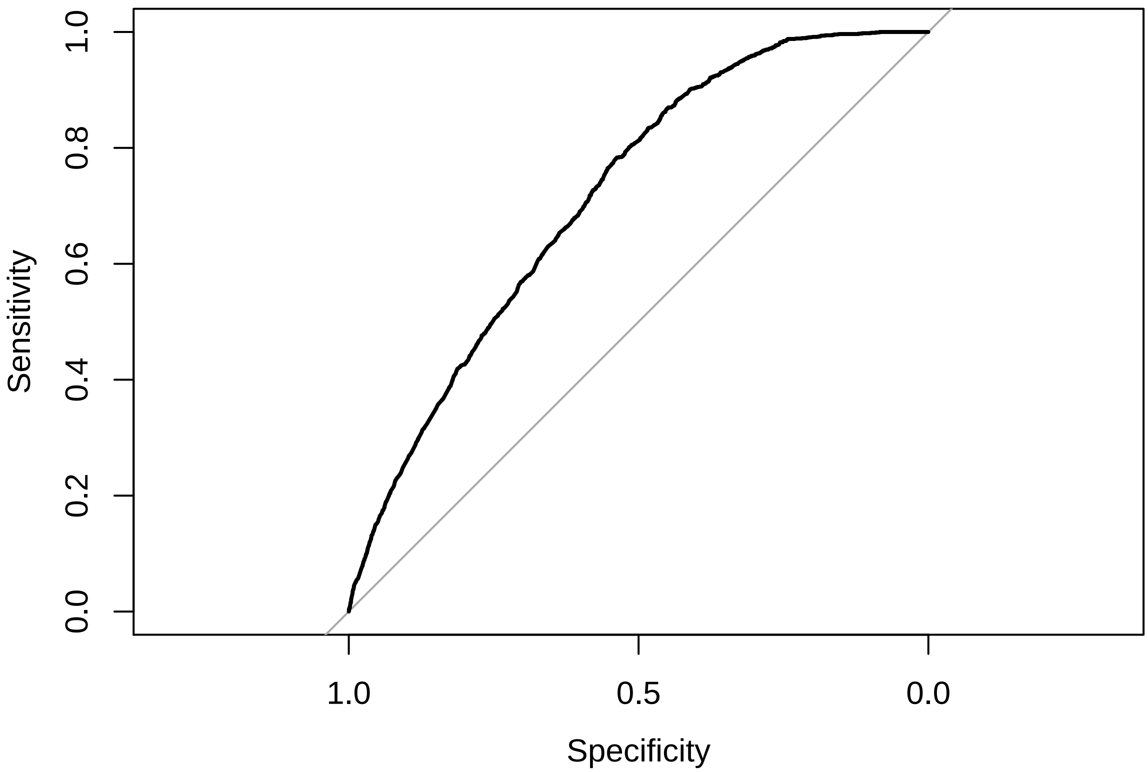
```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(MKmisc)
base.travel <- glm(is_booking~., family = binomial, data = train)
best.travel1 <- stepAIC(base.travel, direction = "both", trace = FALSE, data = train)
summary(best.travel1)
```

```
##
## Call:
## glm(formula = is_booking ~ is_package + channel + prop_is_branded +
##     prop_starrating + popularity_band + cnt + total_cnt, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.8878  -0.5173  -0.3951  -0.1334   3.7254
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.20404    0.44723   0.456 0.648226
## is_package          -0.90370    0.09678  -9.338  < 2e-16 ***
## channel262          -0.13806    0.12325  -1.120 0.262634
## channel293          -0.41539    0.11591  -3.584 0.000339 ***
## channel324           0.10463    0.12909   0.811 0.417646
## channel355           0.40646    0.22371   1.817 0.069236 .
## channel386           0.28811    0.20569   1.401 0.161315
## channel417          -9.75878  139.01412  -0.070 0.944035
## channel448          -1.01466    0.42777  -2.372 0.017694 *
## channel479           0.43029    0.55629   0.773 0.439228
## channel510          -0.06602    0.10529  -0.627 0.530648
## channel541          -0.05581    0.08821  -0.633 0.526933
## prop_is_branded      0.30569    0.06156   4.966 6.83e-07 ***
## prop_starrating1     1.01584    0.71497   1.421 0.155370
## prop_starrating2     0.95326    0.37353   2.552 0.010710 *
## prop_starrating3     0.92610    0.36744   2.520 0.011722 *
## prop_starrating4     0.52058    0.36909   1.410 0.158407
## prop_starrating5     0.28366    0.37764   0.751 0.452574
## popularity_bandL    -0.76228    0.20357  -3.745 0.000181 ***
## popularity_bandM    -0.10270    0.07853  -1.308 0.190902
## popularity_bandVH    0.32915    0.07028   4.684 2.82e-06 ***
## popularity_bandVL   -0.36144    0.43065  -0.839 0.401309
## cnt                 -2.78755    0.23379 -11.924  < 2e-16 ***
## total_cnt           -0.11156    0.02633  -4.236 2.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9553.9  on 15989  degrees of freedom
## Residual deviance: 8661.7  on 15966  degrees of freedom
## AIC: 8709.7
##
## Number of Fisher Scoring iterations: 10
```

```r
po1 <- predict(best.travel1, type = "response")
roc(is_booking~po1,plot=T,data=train)
```



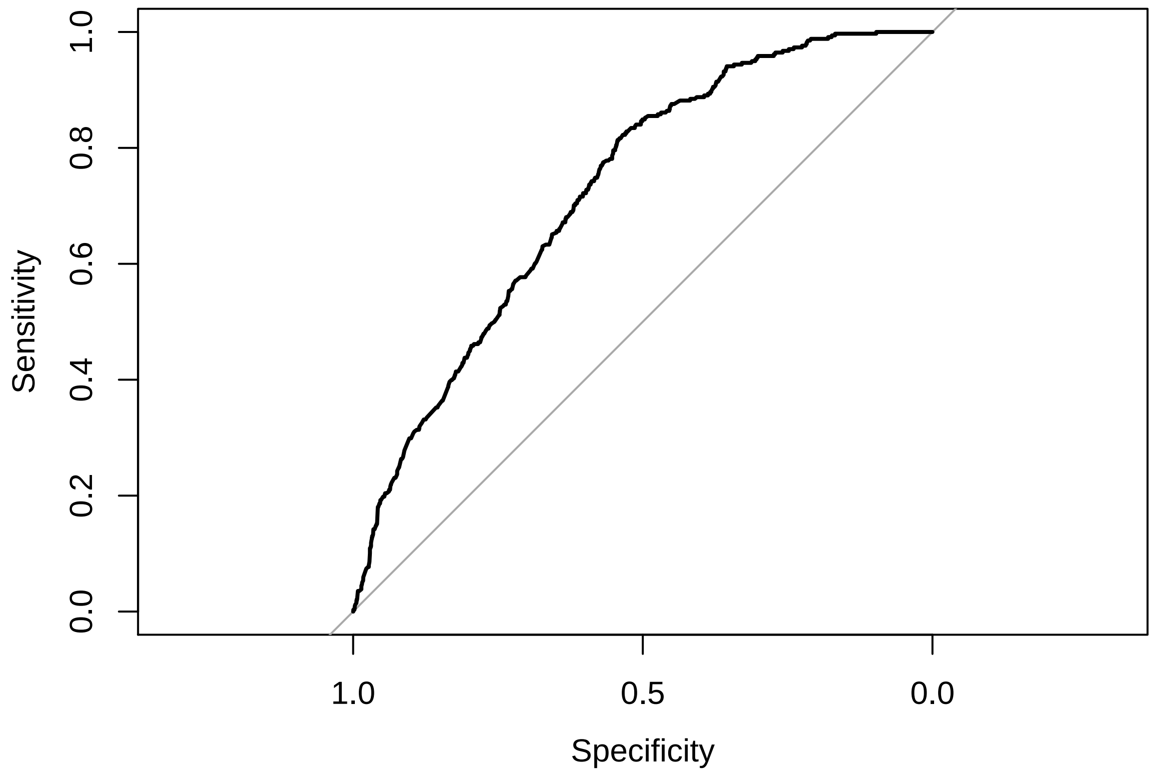

```
##
## Call:
## roc.formula(formula = is_booking ~ po1, data = train, plot = T)
##
## Data: po1 in 14577 controls (is_booking 0) < 1413 cases (is_booking 1).
## Area under the curve: 0.7161
```

```
par(cex=1.2)
HLgof.test(po1,train$is_booking)
```

```
## $C
##
##  Hosmer-Lemeshow C statistic
##
## data:  po1 and train$is_booking
## X-squared = 13.709, df = 8, p-value = 0.08967
##
##
## $H
##
##  Hosmer-Lemeshow H statistic
##
## data:  po1 and train$is_booking
## X-squared = 13.171, df = 8, p-value = 0.1061
```

The best main effects model has an area under ther curve of over 0.7 for the training data, which is very good. Also, the Hosmer-Lemshow test fails to reject the null hypothesis that this is a good model. However we want to see how it handles the validation data.

```
po11 <- predict(best.travel1, type = "response", newdata = val)
roc(is_booking~po11, plot = T, data = val)
```




```
##
## Call:
## roc.formula(formula = is_booking ~ po11, data = val, plot = T)
##
## Data: po11 in 3659 controls (is_booking 0) < 338 cases (is_booking 1).
## Area under the curve: 0.7261
```

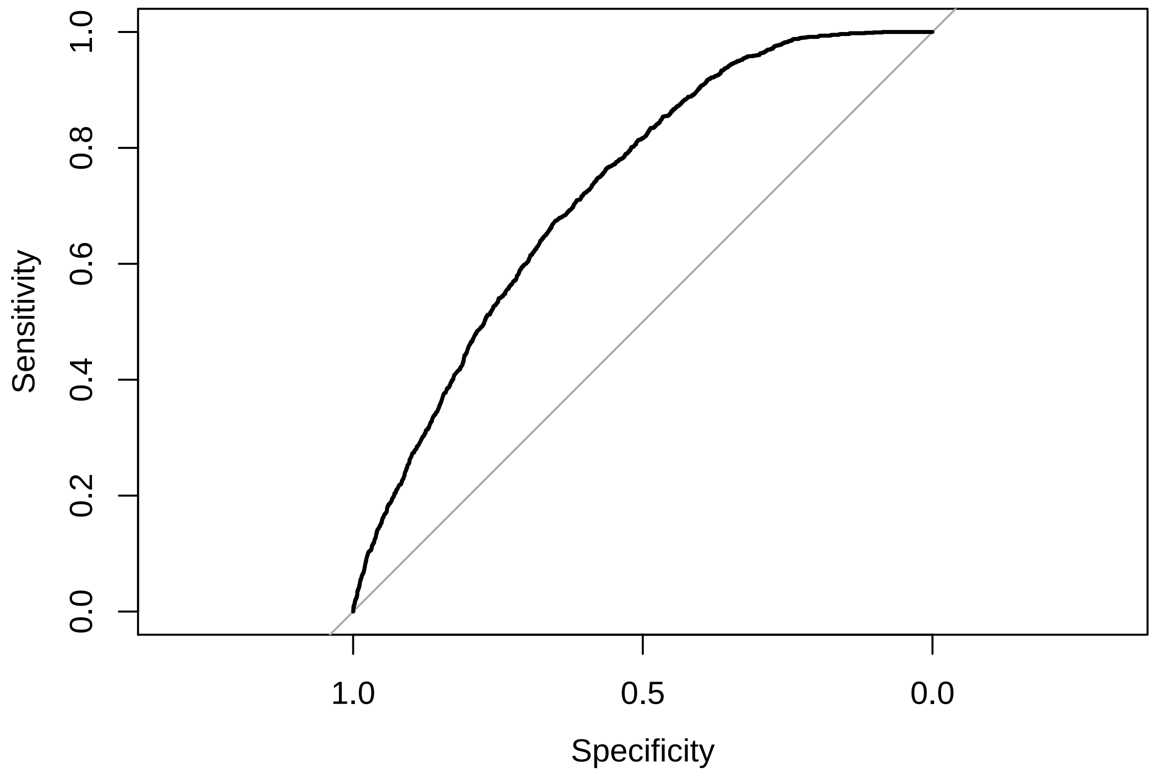
Looking at the area under the curve we see that the best main effects model performs well for the validation data. This means that this model is indeed a good model. We will now look at a model with two way interations.

```
best.travel2 <- stepAIC(base.travel, ~.^2, direction = "both",
                        trace = FALSE, data = train)
summary(best.travel2)
```

```
##
## Call:
## glm(formula = is_booking ~ is_package + channel + prop_is_branded +
##     prop_starrating + distance_band + popularity_band + cnt +
##     total_cnt + prop_is_branded:prop_starrating + distance_band:total_cnt +
##     prop_is_branded:total_cnt + cnt:total_cnt, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9846  -0.5050  -0.3888  -0.1279   3.7324
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -0.18810    0.73213  -0.257 0.797243
```

```
## is_package                      -0.88798    0.09699  -9.156  < 2e-16 ***
## channel262                      -0.13744    0.12373  -1.111 0.266634
## channel293                      -0.41215    0.11624  -3.546 0.000392 ***
## channel324                       0.11940    0.12952   0.922 0.356571
## channel355                       0.38194    0.22481   1.699 0.089330 .
## channel386                       0.27622    0.20640   1.338 0.180810
## channel417                     -12.64813  622.48111  -0.020 0.983789
## channel448                      -1.03196    0.42853  -2.408 0.016034 *
## channel479                       0.47991    0.55762   0.861 0.389435
## channel510                      -0.05067    0.10573  -0.479 0.631772
## channel541                      -0.04947    0.08857  -0.559 0.576478
## prop_is_branded                -11.26719  131.16184  -0.086 0.931544
## prop_starrating1                 0.48551    0.82817   0.586 0.557707
## prop_starrating2                 0.27975    0.39549   0.707 0.479349
## prop_starrating3                 0.50176    0.37511   1.338 0.181014
## prop_starrating4                 0.41497    0.37527   1.106 0.268824
## prop_starrating5                 0.34193    0.40626   0.842 0.399988
## distance_bandF                  -0.58193    0.22785  -2.554 0.010649 *
## distance_bandM                  -0.42624    0.17170  -2.482 0.013048 *
## distance_bandVC                 -0.44314    0.22043  -2.010 0.044391 *
## distance_bandVF                  0.29328    0.29052   1.009 0.312735
## popularity_bandL                -0.75447    0.20395  -3.699 0.000216 ***
## popularity_bandM                -0.10038    0.07893  -1.272 0.203479
## popularity_bandVH                0.34470    0.07066   4.878 1.07e-06 ***
## popularity_bandVL               -0.34002    0.43085  -0.789 0.430011
## cnt                             -1.97772    0.60044  -3.294 0.000988 ***
## total_cnt                        0.22596    0.29643   0.762 0.445888
## prop_is_branded:prop_starrating1  12.90235  131.16919   0.098 0.921643
## prop_is_branded:prop_starrating2  12.40478  131.16191   0.095 0.924652
## prop_is_branded:prop_starrating3  11.99445  131.16182   0.091 0.927137
## prop_is_branded:prop_starrating4  11.50967  131.16182   0.088 0.930074
## prop_is_branded:prop_starrating5  11.32569  131.16194   0.086 0.931189
## distance_bandF:total_cnt         0.17170    0.09062   1.895 0.058138 .
## distance_bandM:total_cnt         0.19138    0.07130   2.684 0.007275 **
## distance_bandVC:total_cnt        0.16166    0.09138   1.769 0.076872 .
## distance_bandVF:total_cnt       -0.09894    0.12509  -0.791 0.428976
## prop_is_branded:total_cnt       -0.10025    0.05542  -1.809 0.070468 .
## cnt:total_cnt                   -0.38825    0.28523  -1.361 0.173458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9553.9  on 15989  degrees of freedom
## Residual deviance: 8601.9  on 15951  degrees of freedom
## AIC: 8679.9
##
## Number of Fisher Scoring iterations: 13
```

```r
po2 <- predict(best.travel2, type = "response")
roc(is_booking~po2,plot=T,data=train)
```



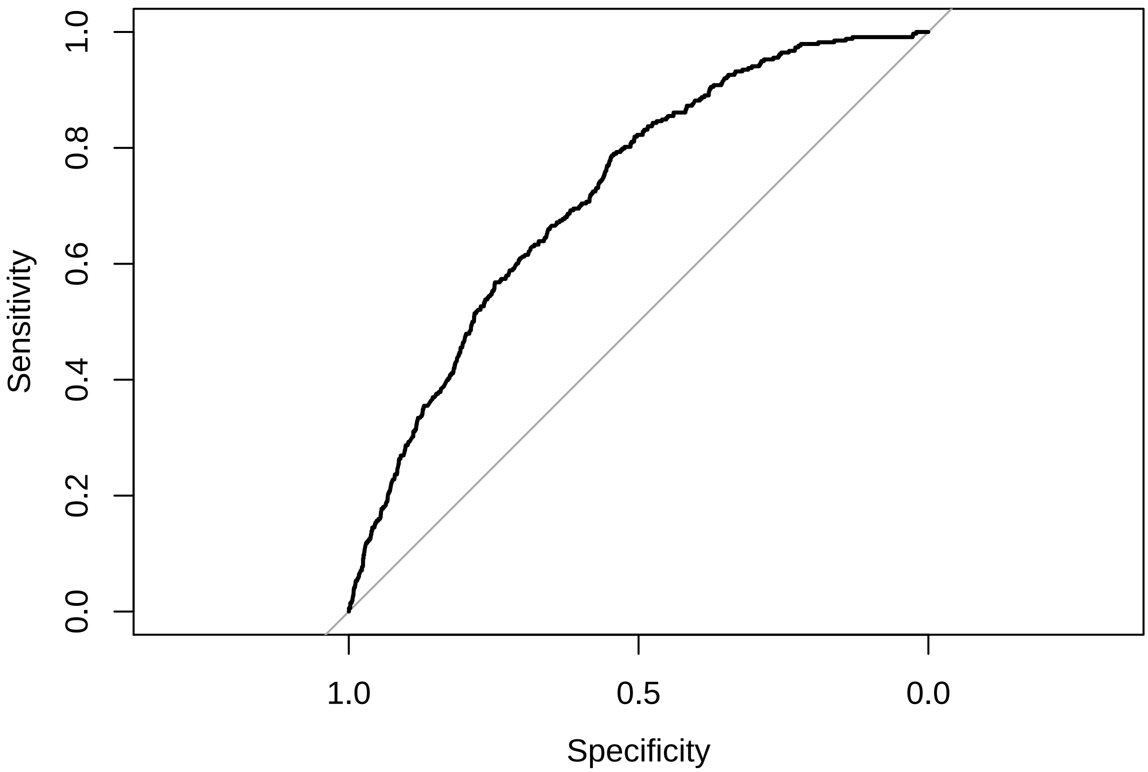

```
##
## Call:
## roc.formula(formula = is_booking ~ po2, data = train, plot = T)
##
## Data: po2 in 14577 controls (is_booking 0) < 1413 cases (is_booking 1).
## Area under the curve: 0.7258
```

```
HLgof.test(po2,train$is_booking)
```

```
## $C
##
##  Hosmer-Lemeshow C statistic
##
## data:  po2 and train$is_booking
## X-squared = 5.1619, df = 8, p-value = 0.7401
##
##
## $H
##
##  Hosmer-Lemeshow H statistic
##
## data:  po2 and train$is_booking
## X-squared = 4.7177, df = 8, p-value = 0.7873
```

The area under the curve is slightly better but not significantly difference from the main effects model. Also, the Hosmer-Lemeshow test fails to reject the null hypothesis that this is a good model.

```
po12 <- predict(best.travel2, type = "response", newdata = val)
roc(is_booking~po12, plot = T, data = val)
```




```
##
## Call:
## roc.formula(formula = is_booking ~ po12, data = val, plot = T)
##
## Data: po12 in 3659 controls (is_booking 0) < 338 cases (is_booking 1).
## Area under the curve: 0.7212
```

With the validation data the model has a slightly lower area under the curve.

Summary

Both the main effects model and the model with interaction were good models with an area under the curve was over 0.7 for both the training data and validation data. Also, both performed well with the Hosmer-Lemeshow test. Although the model with interaction had a higher area under the curve, it was only a 0.01 increase on the training data than the main effects model and on the validation data the area under the curve was roughly the same as the main effects model. Also, the model with interaction had a lot of variables. Therefore, I would choose the main effects model because it is simpler, has less variables, and we don't lose anything. Significant coefficients with positive effects on booking a hotel room included if the hotel are a brand name, a 2 or 3 star rating, and how often it was booked is very high. Significant coefficients with negative effects on booking a hotel room are if it is a package, channel 1293 and 1448, how often it is booked is low, number of clicks/bookings in the same session, and the total amount of people in the room. The big takeaways are that more clicks mean the log odds are less to book, being a big brand with a high number of bookings and a higher star rating leads to a higher log odds of booking, and having low popularity leads to a lower log odds of booking.

Problem 4. (25 points) Poisson Regression using my own data set.

For this problem I will be using the package Lahman, which contains baseball data from 1871-2016. I am going to subset the data set Teams for 2010-2016, which contains 7 years of yearly statistics for teams. With 30 teams and 7 years of data it will be 210 observations of 48 variables. My goal will be to use Poisson Regression for predicting wins, which is a count variable.

```
library(Lahman)
MLB2010.2016 <- Teams[Teams$yearID >2009,]
MLB2010.2016 <- MLB2010.2016[,-c(1,2,3,4,5,6,7,8,10,11,12,13,14,16,
                            18,19,24,25,28,29,30,31,32,33,39,40,41,42,43,44,45,46,47,48)]
names(MLB2010.2016)
```

```
##  [1] "W"   "R"   "H"   "HR"  "BB"  "SO"  "SB"  "SF"  "RA"  "HA"  "HRA"
## [12] "BBA" "SOA" "E"
```

After careful considerations of variables, I decided to remove 36 variables and leave 14 for analysis. I removed these for many reason including: Several different ID terms, same variables in different forms, and some variables not being important. I specifially removed runs and runs allowed because these variables would dominate the model. I want to look and the underlying statistics that contribute to wins. The remaining variables include wins, hits, homeruns, walks, strikeouts, stolen bases, sacrifice flies, runs, runs allowed, hits allowed, homeruns allowed, walks allowed, strikeouts by pitchers, and errors.

```
library(readr)
MLBTeamOffenseStats_2010_2017_<-read_csv("~ChrisWatkins/Desktop/MLB Data/MLBTeamOffenseStats(2010-2017)
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Tm = col_character(),
##   BatAge = col_double(),
##   `R/G` = col_double(),
##   BA = col_double(),
##   OBP = col_double(),
##   SLG = col_double(),
##   OPS = col_double()
## )

## See spec(...) for full column specifications.
```

```
MLBOff10.16 <- MLBTeamOffenseStats_2010_2017_[MLBTeamOffenseStats_2010_2017_$Year <2017,]
OPSp <- MLBOff10.16$`OPS+`
LOB <- MLBOff10.16$LOB
MLB2010.2016 <- cbind(MLB2010.2016, OPSp, LOB)
```

I have a second data set with more offensive metrics that I want to add that were not in the data set, which are OPS+ (On base plus slugging adjusted) and left on base (LOB).

```
base.MLB <- glm(W~., family = poisson, data = MLB2010.2016)
best.MLB1 <- stepAIC(base.MLB, direction = "both", trace = FALSE,
                     data = MLB2010.2016)
summary(best.MLB1)
```

```
##
## Call:
## glm(formula = W ~ R + RA, family = poisson, data = MLB2010.2016)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -1.09356  -0.28258   0.00546   0.27832   1.63147
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.5342368  0.1026237   44.18   <2e-16 ***
## R            0.0011438  0.0001101   10.38   <2e-16 ***
## RA          -0.0013574  0.0001036  -13.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 313.104  on 209  degrees of freedom
## Residual deviance:  42.777  on 207  degrees of freedom
## AIC: 1356
##
## Number of Fisher Scoring iterations: 3
```

```
best.MLB2 <- stepAIC(base.MLB, ~.^2, direction = "both", trace = FALSE,
                     data = MLB2010.2016)
summary(best.MLB2)
```

```
##
## Call:
## glm(formula = W ~ R + RA, family = poisson, data = MLB2010.2016)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.09356  -0.28258   0.00546   0.27832   1.63147
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.5342368  0.1026237   44.18   <2e-16 ***
## R            0.0011438  0.0001101   10.38   <2e-16 ***
## RA          -0.0013574  0.0001036  -13.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 313.104  on 209  degrees of freedom
## Residual deviance:  42.777  on 207  degrees of freedom
## AIC: 1356
##
## Number of Fisher Scoring iterations: 3
```

```
pchisq(deviance(best.MLB1), df.residual(best.MLB1), lower.tail = FALSE)
```

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

In both the main effects model and the model with 2-way interaction the two significant coefficients were runs and runs allowed where runs have a positive effect and runs allowed had a negative effect. The deviance goodness of fit test confirms it is a good model. This seems likely and not ground breaking since more runs means more wins and less runs allowed mean more wins. I will now look deeper and build a Poisson Regression Model for runs.

```r
MLB2010.2016 <- MLB2010.2016[,-c(1,9,19,11,12,13,14)]
```

First, I need to remove variables that have no effect on runs. These variables include: wins, runs allowed, hits allowed, home runs against, walks against, strikeouts by pitcher and errors.

```r
base.MLB2 <- glm(R~., family = poisson, data = MLB2010.2016)
best.rMLB1 <- stepAIC(base.MLB2, direction = "both", trace = FALSE, data = MLB2010.2016)
summary(best.rMLB1)
```

```
##
## Call:
## glm(formula = R ~ H + HR + BB + SB + SF, family = poisson, data = MLB2010.2016)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6304  -0.4926  -0.0012   0.5353   3.4268
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 4.967e+00  5.527e-02  89.871  < 2e-16 ***
## H           7.390e-04  3.964e-05  18.640  < 2e-16 ***
## HR          1.392e-03  8.529e-05  16.325  < 2e-16 ***
## BB          4.906e-04  4.820e-05  10.178  < 2e-16 ***
## SB          2.774e-04  8.954e-05   3.098  0.00195 **
## SF          1.009e-03  3.792e-04   2.662  0.00777 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1432.72  on 209  degrees of freedom
## Residual deviance:  159.57  on 204  degrees of freedom
## AIC: 1930.2
##
## Number of Fisher Scoring iterations: 3
```

```r
pchisq(deviance(best.rMLB1), df.residual(best.rMLB1), lower.tail = FALSE)
```

```
## [1] 0.9905969
```

```r
best.rMLB2 <- stepAIC(base.MLB2, ~.^2, direction = "both", trace = FALSE, data = MLB2010.2016)
summary(best.rMLB2)
```

```
##
## Call:
## glm(formula = R ~ H + HR + BB + SB + SF + LOB + BB:SF + BB:LOB,
##     family = poisson, data = MLB2010.2016)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5621  -0.4575  -0.0581   0.5268   3.3852
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.397e+00  4.849e-01   9.068  < 2e-16 ***
## H            7.542e-04  4.009e-05  18.813  < 2e-16 ***
```

```
## HR          1.388e-03  8.577e-05  16.186  < 2e-16 ***
## BB          1.702e-03  9.520e-04   1.788  0.07373 .
## SB          2.579e-04  9.049e-05   2.851  0.00436 **
## SF         -4.378e-03  2.995e-03  -1.462  0.14382
## LOB         6.997e-04  4.231e-04   1.654  0.09822 .
## BB:SF       1.076e-05  5.920e-06   1.817  0.06922 .
## BB:LOB     -1.498e-06  8.391e-07  -1.786  0.07416 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1432.72  on 209  degrees of freedom
## Residual deviance:  152.34  on 201  degrees of freedom
## AIC: 1928.9
##
## Number of Fisher Scoring iterations: 3
```

```r
pchisq(deviance(best.rMLB2), df.residual(best.rMLB2), lower.tail = FALSE)
```

```
## [1] 0.995634
```

Both the main effects model and model with interaction are good based on the deviance goodness of fit statistic.

Summary

I first started the analysis by using Poisson Regression to model wins. I found that two factors, runs and runs against, were the only variables in the model. This was not a very interesting result so I looked deeper at what are the biggest factors of runs. Both the main effects model and model with interation were good models based on the deviance goodness of fit statistic. The model with interaction has a slightly better AIC, so I will analyze the main effects model as the best model. All variables, hits, homeruns, walks, stolen bases, and sacrafice flies were significant.

```r
exp(1.392e-3)
```

```
## [1] 1.001393
```

In looking at the exponential of the homerun variable we see that on average for 1 more homerun the expected runs go up 0.1%. This means that for 10 more homeruns the expected runs go up 1% on average. While 1% may seem small, it can make all the difference in making the playoffs and not making the playoffs. The most interesting part of this result is that homeruns have a bigger effect than walks, OPS+ was not in the model, and strikeouts were not in the model. In the baseball industry OPS+ is regarded at the big statistic for increasing run production. Also, players that hit a lot of homeruns but do not get on base (i.e. more walks) are not sought after. Walks are important, but the model seems to show that it does not matter how much you strike out if the homerun numbers are high. There are new metrics being developed in baseball all the time for better prediction of players increasing wins. I did not use data sets from 2017, which should to be noted because in 2017 the number of homeruns hit and number of strikeouts were the highest in the history of baseball. This shows a shift in philosophy, that backs up this model.