Final Quiz

Ryan Richardson

21/12/2020

Question 1 - Retirement

Preprocessing

Retirement data is read in. The first two columns are removed as they are both variations of an ID column.

NAs in the various pct categories were filled with 0s to ensure the model could be properly built in a stepwise fashion.

```r
retirementData <- read.csv("pension.csv", header=TRUE)
retirementData<-retirementData[,-c(1,2)] #remove id and row count
nafill(retirementData[,c(2,3,4,5,8,9,10,11,12,13,15,16,17)], fill=0) # fill missing factors with 0s, on
```

```
## [[1]]
##   [1] 0 1 1 1 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
##  [38] 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0
##  [75] 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
## [112] 0 1 0 0 0 1 0 1 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
## [149] 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0
## [186] 0 0 1 1 0 1 0 1 1
##
## [[2]]
##   [1] 1 1 1 0 1 0 1 0 0 1 1 0 1 0 1 1 1 1 1 0 0 1 1 0 1 0 0 0 0 0 1 1 1 0 0 1 1
##  [38] 0 0 1 0 1 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 1 0 1 1 1 1 1 1 0
## [112] 0 1 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 0 0 1 1 1 0 1 0 1 1 1 1 1 0
## [149] 0 1 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 1 0 1 0 0 1 0 1
## [186] 1 0 0 1 1 0 1 0 0
##
## [[3]]
##   [1] 0 1 0 1 0 0 1 1 1 0 1 1 0 1 0 1 1 0 0 1 1 0 1 1 1 1 1 1 0 1 1 0 0 0 1 1 0
##  [38] 1 1 0 1 1 1 0 1 1 1 0 0 1 0 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 0
##  [75] 0 1 0 0 1 0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 0 0
## [112] 1 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 0 0 1 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 0 1
## [149] 1 1 1 1 0 1 1 0 1 1 1 0 1 1 1 0 0 1 0 0 1 1 1 1 0 1 0 1 0 1 1 1 1 0 0 0 1
## [186] 0 0 0 0 1 1 0 1 0
##
## [[4]]
##   [1] 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 1 1 1 0 0 1
##  [38] 1 1 1 0 0 0 1 0 1 1 1 1 0 1 1 0 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 1 0 1 0 1 1 1 1
## [112] 1 0 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1
## [149] 0 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 0 1 0 1 0 1 0 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1
```

```
##
## [[5]]
##   [1] 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0
##  [38] 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 1 1 0 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0
## [149] 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 1 1 0 0 0
## [186] 0 1 0 0 0 1 0 1 1
##
## [[6]]
##   [1] 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 1
##  [38] 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0
##  [75] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1
## [112] 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0
## [186] 0 0 0 0 0 0 0 0 0
##
## [[7]]
##   [1] 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 1 0 0 1 0
##  [38] 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0
##  [75] 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 1 0 0 0
## [112] 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 1 1 1 0 0 1 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0 0
## [149] 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0
## [186] 1 0 1 0 0 0 0 0 0
##
## [[8]]
##   [1] 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
##  [75] 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 1 1 1 0
## [149] 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [186] 0 0 0 1 0 0 0 0 0
##
## [[9]]
##   [1] 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1
##  [75] 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [186] 0 0 0 0 1 0 1 0 0
##
## [[10]]
##   [1] 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
##  [38] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
## [149] 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
## [186] 0 0 0 0 0 0 0 0 0
##
## [[11]]
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0
##  [38] 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 1 0 0 0
## [112] 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [186] 0 1 0 0 0 0 0 0 0
##
## [[12]]
##   [1] 1 1 1 1 0 1 0 0 0 1 1 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 1 0 1 0 0
##  [38] 1 0 0 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1
##  [75] 1 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0
## [112] 1 0 1 1 0 1 0 0 0 1 1 1 0 0 1 0 1 0 1 0 0 1 0 1 0 0 1 0 0 1 0 0 0 0 1 0 0
## [149] 0 1 1 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0 0
## [186] 0 0 0 0 0 0 1 0 0
##
## [[13]]
##   [1] 1 1 1 1 1 0 1 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 0 0 1 0 0 1 1 0 0 1 1 1 0 1 0
##  [38] 1 0 0 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0
##  [75] 1 1 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0
## [112] 1 1 0 1 1 0 0 1 0 0 1 1 0 1 1 1 1 0 1 0 0 1 0 0 1 1 1 1 0 1 1 1 0 0 1 0 1
## [149] 1 1 1 1 1 1 1 1 0 1 1 0 1 0 0 0 1 1 1 1 0 0 1 0 1 0 1 0 1 1 1 0 0 0 1 1 0
## [186] 0 0 1 0 0 0 1 0 0
```

```r
retirementData[,c(2,3,4,5,8,9,10,11,12,13,15,16,17)] <- lapply(retirementData[,c(2,3,4,5,8,9,10,11,12,1
retirementData<- na.omit(retirementData)
summary(retirementData)
```

```
##      pyears     prftshr choice  female  married      age             educ
##  Min.   : 0.0   0:151   0: 74   0: 75   0: 47   Min.   :54.00   Min.   : 8.00
##  1st Qu.: 4.0   1: 40   1:117   1:116   1:144   1st Qu.:57.00   1st Qu.:12.00
##  Median : 9.0                                   Median :60.00   Median :12.00
##  Mean   :11.3                                   Mean   :60.52   Mean   :13.53
##  3rd Qu.:16.0                                   3rd Qu.:64.00   3rd Qu.:16.00
##  Max.   :45.0                                   Max.   :73.00   Max.   :18.00
##  finc25  finc35  finc50  finc75  finc100 finc101    wealth89      black
##  0:151   0:157   0:146   0:165   0:165   0:181   Min.   :  -6.3   0:169
##  1: 40   1: 34   1: 45   1: 26   1: 26   1: 10   1st Qu.:  65.8   1: 22
##                                                  Median : 140.0
##                                                  Mean   : 212.0
##                                                  3rd Qu.: 253.4
##                                                  Max.   :1485.0
##  stckin89 irain89    pctstck
##  0:126    0:93    Min.   :  0.00
##  1: 65    1:98    1st Qu.:  0.00
##                   Median : 50.00
##                   Mean   : 48.43
##                   3rd Qu.:100.00
##                   Max.   :100.00
```

```r
retirementModel <- lm(wealth89 ~ ., data = retirementData)
retirementInteractionModel <- lm(wealth89~.^2, data=retirementData)
retirementStep <- stepAIC(retirementModel, trace=0)
retirementIneractionStep <- stepAIC(retirementInteractionModel, trace =0)


summary(retirementStep)
```

```
##
## Call:
```

```
## lm(formula = wealth89 ~ age + finc50 + finc75 + finc100 + finc101 +
##     stckin89 + irain89, data = retirementData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -413.65 -113.98  -46.41   69.79 1147.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -593.247    227.590  -2.607 0.009897 **
## age           10.677      3.736   2.858 0.004758 **
## finc501       58.452     39.542   1.478 0.141065
## finc751      168.494     48.878   3.447 0.000703 ***
## finc1001     151.098     47.842   3.158 0.001857 **
## finc1011     350.426     70.951   4.939 1.76e-06 ***
## stckin891    109.376     34.821   3.141 0.001963 **
## irain891      90.154     33.367   2.702 0.007542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 211.1 on 183 degrees of freedom
## Multiple R-squared:  0.2938, Adjusted R-squared:  0.2668
## F-statistic: 10.88 on 7 and 183 DF,  p-value: 1.888e-11
```

```
summary(retirementIneractionStep)
```

```
##
## Call:
## lm(formula = wealth89 ~ pyears + prftshr + choice + female +
##     married + age + educ + finc25 + finc35 + finc50 + finc75 +
##     finc100 + finc101 + black + stckin89 + irain89 + pctstck +
##     pyears:age + pyears:finc25 + pyears:finc35 + pyears:finc50 +
##     pyears:finc75 + pyears:finc100 + pyears:finc101 + pyears:stckin89 +
##     prftshr:married + prftshr:finc101 + prftshr:black + prftshr:stckin89 +
##     choice:married + choice:educ + choice:finc25 + choice:finc35 +
##     choice:finc100 + choice:finc101 + choice:stckin89 + choice:irain89 +
##     female:age + female:educ + female:finc25 + female:finc35 +
##     female:finc50 + female:finc75 + female:finc100 + female:stckin89 +
##     female:irain89 + female:pctstck + married:age + married:finc25 +
##     married:finc35 + married:finc50 + married:finc75 + married:finc100 +
##     married:stckin89 + married:pctstck + age:finc75 + age:finc100 +
##     age:finc101 + educ:finc25 + educ:finc35 + educ:finc50 + educ:finc75 +
##     educ:finc100 + educ:finc101 + educ:irain89 + finc25:black +
##     finc25:stckin89 + finc25:irain89 + finc25:pctstck + finc35:black +
##     finc35:stckin89 + finc35:irain89 + finc35:pctstck + finc50:black +
##     finc50:stckin89 + finc50:irain89 + finc75:stckin89 + finc75:irain89 +
##     finc75:pctstck + finc100:irain89 + finc100:pctstck + black:stckin89 +
##     black:pctstck, data = retirementData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -267.14  -63.13   -3.03   45.96  716.32
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -8098.5609  1187.7594  -6.818 5.67e-10 ***
## pyears                62.4838    27.6563   2.259 0.025892 *
## prftshr1             -23.9464    96.1356  -0.249 0.803769
## choice1             -490.8660   195.5814  -2.510 0.013578 *
## female1             3091.5883   638.2468   4.844 4.32e-06 ***
## married1             454.4384   632.5852   0.718 0.474087
## age                   33.7457    13.3483   2.528 0.012929 *
## educ                 414.9635    58.7316   7.065 1.70e-10 ***
## finc251             5694.1030   811.5925   7.016 2.16e-10 ***
## finc351             5313.2173   851.8555   6.237 9.07e-09 ***
## finc501             6134.0235   823.8556   7.446 2.59e-11 ***
## finc751             5823.0858  1117.6492   5.210 9.23e-07 ***
## finc1001            2523.7210  1089.4405   2.317 0.022435 *
## finc1011            3227.6682  1799.3189   1.794 0.075665 .
## black1               415.5876   154.2692   2.694 0.008201 **
## stckin891            337.4968   154.6438   2.182 0.031266 *
## irain891            -289.6711   233.3623  -1.241 0.217212
## pctstck               -0.8516     1.2920  -0.659 0.511222
## pyears:age            -1.4008     0.4110  -3.409 0.000922 ***
## pyears:finc251        23.7974     9.4945   2.506 0.013700 *
## pyears:finc351        26.2542     9.1084   2.882 0.004771 **
## pyears:finc501        26.9819     9.0007   2.998 0.003382 **
## pyears:finc751        17.0941     9.5147   1.797 0.075221 .
## pyears:finc1001       16.3112     8.8386   1.845 0.067738 .
## pyears:finc1011       91.1912    22.6012   4.035 0.000103 ***
## pyears:stckin891      13.9932     4.1735   3.353 0.001107 **
## prftshr1:married1    136.6050   100.1090   1.365 0.175254
## prftshr1:finc1011  -2069.7916   369.5861  -5.600 1.67e-07 ***
## prftshr1:black1    -2045.8026   334.2212  -6.121 1.56e-08 ***
## prftshr1:stckin891  -362.0428    84.5862  -4.280 4.08e-05 ***
## choice1:married1      98.6887    74.9988   1.316 0.191030
## choice1:educ          19.2869    13.0550   1.477 0.142516
## choice1:finc251      318.6459    90.6814   3.514 0.000648 ***
## choice1:finc351      166.1198    84.4197   1.968 0.051680 .
## choice1:finc1001    -161.1465   101.0216  -1.595 0.113624
## choice1:finc1011    1404.0247   217.0701   6.468 3.05e-09 ***
## choice1:stckin891   -143.1790    79.0033  -1.812 0.072742 .
## choice1:irain891      87.8452    66.0656   1.330 0.186456
## female1:age          -24.6097     9.4642  -2.600 0.010631 *
## female1:educ         -28.3712    12.1152  -2.342 0.021043 *
## female1:finc251    -1312.1696   205.2312  -6.394 4.34e-09 ***
## female1:finc351    -1152.6486   198.4103  -5.809 6.53e-08 ***
## female1:finc501    -1260.4743   190.5621  -6.615 1.52e-09 ***
## female1:finc751    -1161.2663   210.9718  -5.504 2.56e-07 ***
## female1:finc1001    -954.3958   187.7976  -5.082 1.60e-06 ***
## female1:stckin891    221.7983    83.5362   2.655 0.009139 **
## female1:irain891    -104.5060    66.2370  -1.578 0.117573
## female1:pctstck        1.2572     0.9768   1.287 0.200848
## married1:age         -11.7878    10.0042  -1.178 0.241298
## married1:finc251     255.2812   159.4110   1.601 0.112236
## married1:finc351     352.1483   181.6968   1.938 0.055245 .
## married1:finc501     215.9407   163.8411   1.318 0.190322
## married1:finc751    -888.5169   255.3484  -3.480 0.000728 ***
## married1:finc1001    933.6802   249.9010   3.736 0.000302 ***
## married1:stckin891   128.1676    94.2450   1.360 0.176707    
## married1:pctstck      -1.7991     1.0912  -1.649 0.102148    
## age:finc751           22.2846    12.1517   1.834 0.069453 .  
## age:finc1001          53.9224    12.5828   4.285 4.00e-05 ***
## age:finc1011         126.5629    19.9049   6.358 5.13e-09 ***
## educ:finc251        -382.0073    59.0345  -6.471 3.01e-09 ***
## educ:finc351        -367.6033    60.1644  -6.110 1.64e-08 ***
## educ:finc501        -404.0051    59.3892  -6.803 6.12e-10 ***
## educ:finc751        -396.4842    61.1957  -6.479 2.90e-09 ***
## educ:finc1001       -422.7306    60.5111  -6.986 2.51e-10 ***
## educ:finc1011       -815.9719   107.0343  -7.623 1.06e-11 ***
## educ:irain891        -27.3758    11.7666  -2.327 0.021872 *  
## finc251:black1      -306.8674   157.9092  -1.943 0.054605 .  
## finc251:stckin891   -487.0534   117.1365  -4.158 6.49e-05 ***
## finc251:irain891     661.8288   160.7863   4.116 7.60e-05 ***
## finc251:pctstck        1.5625     0.9770   1.599 0.112709    
## finc351:black1      -282.4027   171.5972  -1.646 0.102754    
## finc351:stckin891   -501.4788   132.4692  -3.786 0.000253 ***
## finc351:irain891     779.7560   154.1886   5.057 1.77e-06 ***
## finc351:pctstck        2.6565     1.0525   2.524 0.013076 *  
## finc501:black1      -190.1708   152.5641  -1.246 0.215304    
## finc501:stckin891   -334.0412   103.2410  -3.236 0.001616 ** 
## finc501:irain891     791.2109   153.9876   5.138 1.26e-06 ***
## finc751:stckin891   -397.0207   112.9333  -3.516 0.000645 ***
## finc751:irain891     700.9223   156.3582   4.483 1.86e-05 ***
## finc751:pctstck        4.5420     1.3514   3.361 0.001078 ** 
## finc1001:irain891    508.6392   151.6738   3.354 0.001104 ** 
## finc1001:pctstck       2.8752     1.3068   2.200 0.029946 *  
## black1:stckin891    -235.0598   155.7495  -1.509 0.134191    
## black1:pctstck        -3.8256     1.4495  -2.639 0.009551 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 152 on 107 degrees of freedom
## Multiple R-squared:  0.786,  Adjusted R-squared:  0.6199 
## F-statistic: 4.734 on 83 and 107 DF,  p-value: 7.349e-14
```

Interpretation

The models were built in a stepwise fashion.

The first model used only first order interactions, with an Adj-R Squared value of 0.27. The significant components of the initial model were a person's age, whehter how much they put away for their retirement, and whether or not they had money in stock or in an ira.

The base case was someone who was 0 years old, had a retirement contribution level of finc25, had no money in stock and no money in an IRA. This type of person doesn't exist. What most stuck out was the intercept showed that most people started at a negative wealth value, and by the time they were ~55, they would be at a value of 0.

Unsurprisingly, the greater the finance level was and an indicator that they had money in both stock and IRA would increase their wealth at retirement significantly.

The second model used third order interactions and the Adj-R Squared increased singificantly to 0.62. In this model, every variable was considered signifiant, either on its own or beause of an interaction it had with other variables.

The base case was someone employed for 0 years, without a profit sharing option, and without a choice of participation in their company's retirement contribution. They were male, unmarried, at 0 years old, with 0 years of education. They were white, with no money in stocks or an IRA, and with no financing level. Though this person doesn't exist, it stood out just how far the intercept was below 0, essentially meaning that a newly born white male started with an 8.1 million dollar 'negative wealth' value, which in practice doesn't really make sense.

To sum up the most significant findings: For each year employed, wealth at retirement increased by 62.48, being female increased the starting wealth at retirement by 3091, and for every year olde ra person increased their wealthy by 33.75. Every year of education increased wealth by 414.96, and each of the investment levels increased the ending wealth significantly. The best finance level was 50 for retirement contribution. Black indivdiuals had a higher started wealth of 415.58, and having socks in the retirement investment was more significant than having an IRA. In fact the IRA on its own was a negative.

Some of the more surprising take aways were that contirbution rates on their own becames less effective above 50% saw a decrease in total wealth at retirement. Additionally, females saw negative wealth accumulating across all financing levels, as did years of education.

Looking more thoroughly at the effects of the second order interaction terms, it's hard to tell if this model is the correct choice. A lot of the effects may be significant, and may be considered optimal from a stepwise model building algorithm, but the size and direction of their effects have me concerned that what we are seeing are many non-linear trends in the data that are being coerced into a linear model throwing off the effect sizes and directions.

Question 2 - Travel

Preprocessing

Many variables are factor variables and are converted to such (region, city, mobile, package, channel, desId, hotel country, booking, hotel id, branded, starrating, distance, hist price, and popularity).

Location latitude, logintude, and origin distance are converted to numeric.

Due to issues with the ISO 8601 timestamp in date_time, that column needs to be removed as its current state does not give an accurate representation of the true time that the user accessed the site wthout the timezone information attached. Latitude and longitude are both based on the user_location_city, as it is already encoded it is dropped. The user and Hotel IDs are dropped as they are insignificant or overly specific.

User_location_city was dropped due to the limited number of observations across each city throughout the entire dataset.

Orig_destination_distance is dropped due to excessive NA's.

The bottom 15% of user location region were grouped into a level called "other" as these destinations were exceedingly rare compared to the most common destinations. This significantly sped up the processing time of the model. As an example, the location_region was reduced to 45 total levels, with the lowest having 85 instances and the highest having almost 3000.

The same was done for srch_destination_id, but the bottom 60% was grouped. The smallest group now had 47 instances and the largest had close to 900. Lower thresholds were tried but often resulted in hundreds of levels that had single digit numbers, and did not make sense to include. Overall, 3793 levels were reduced to 64.

```
travel = read.table("Travel.txt", sep="", header=TRUE)
travel[,c(2,3,8,9,10,16,17,18,19,20,21,22,23,24)] <- lapply(travel[,c(2,3,8,9,10,16,17,18,19,20,21,22,23
travel[,c(4,5,6)] <- lapply(travel[,c(4,5,6)], as.numeric)
travelData = travel[,-c(1,3,4,5,6,7,11,12,19)]
travelData=group_category(travelData, "user_location_region", 0.15, update=TRUE)
travelData=group_category(travelData, "srch_destination_id", 0.60, update=TRUE)
```

```r
travelData[,c(1,8)] <- lapply(travelData[,c(1,8)], factor)

summary(travelData)
```

```
##  user_location_region is_mobile is_package    channel     srch_adults_cnt
##  OTHER  :3001         0:15533   0:16103    541    :7805   Min.   :0.000  
##  CA     :2924         1: 4467   1: 3897    510    :3079   1st Qu.:2.000  
##  NY     :1236                              231    :2615   Median :2.000  
##  TX     :1163                              293    :2579   Mean   :2.056  
##  FL     :1063                              262    :1797   3rd Qu.:2.000  
##  ON     : 938                              324    :1277   Max.   :9.000  
##  (Other):9675                              (Other): 848                  
##  srch_children_cnt  srch_rm_cnt    srch_destination_id
##  Min.   :0.0000    Min.   :0.000   OTHER  :12026      
##  1st Qu.:0.0000    1st Qu.:1.000   5526679:  881      
##  Median :0.0000    Median :1.000   5527206:  521      
##  Mean   :0.3108    Mean   :1.077   5579968:  373      
##  3rd Qu.:0.0000    3rd Qu.:1.000   5527237:  369      
##  Max.   :8.0000    Max.   :8.000   5527578:  268      
##                                    (Other): 5562      
##                    hotel_country   is_booking prop_is_branded prop_starrating
##  UNITED STATES OF AMERICA:12009   0:18247    0: 7671         0: 283         
##  CANADA                  : 1141   1: 1753    1:12329         1:  38         
##  MEXICO                  : 1072                              2:1993         
##  ITALY                   :  541                              3:6836         
##  UNITED KINGDOM          :  426                              4:8227         
##  FRANCE                  :  377                              5:2623         
##  (Other)                 : 4434                                             
##  distance_band hist_price_band popularity_band      cnt        
##  C :5130       H :4065         H :5974         Min.   : 1.000  
##  F :2732       L :3873         L : 721         1st Qu.: 1.000  
##  M :7631       M :8078         M :5213         Median : 1.000  
##  VC:3155       VH:2108         VH:7970         Mean   : 1.421  
##  VF:1352       VL:1876         VL: 122         3rd Qu.: 1.000  
##                                                Max.   :38.000  
## 
```

```r
#travelModel <- glm(is_booking~., data=travelData, family=binomial)
#travelStep<- stepAIC(travelModel, trace=0)
travelBest <- glm(formula = is_booking ~ is_package + channel + srch_adults_cnt +
    srch_children_cnt + srch_rm_cnt + srch_destination_id + prop_is_branded +
    prop_starrating + distance_band + hist_price_band + popularity_band +
    cnt, family = binomial, data = travelData)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(travelBest)
```

```
## 
## Call:
## glm(formula = is_booking ~ is_package + channel + srch_adults_cnt + 
##     srch_children_cnt + srch_rm_cnt + srch_destination_id + prop_is_branded + 
##     prop_starrating + distance_band + hist_price_band + popularity_band + 
##     cnt, family = binomial, data = travelData)
## 
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9889  -0.5174  -0.3692  -0.1416   3.6564
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -8.551e-01  6.955e-01  -1.229  0.21889
## is_package1                  -8.183e-01  9.275e-02  -8.823  < 2e-16 ***
## channel262                   -1.115e-01  1.117e-01  -0.998  0.31831
## channel293                   -4.071e-01  1.058e-01  -3.847  0.00012 ***
## channel324                    1.195e-01  1.179e-01   1.014  0.31052
## channel355                    4.850e-01  2.034e-01   2.384  0.01711 *
## channel386                    3.285e-01  1.851e-01   1.775  0.07596 .
## channel417                   -1.497e+01  1.199e+03  -0.012  0.99003
## channel448                   -9.101e-01  3.730e-01  -2.440  0.01470 *
## channel479                    6.017e-01  5.138e-01   1.171  0.24158
## channel510                   -5.856e-02  9.597e-02  -0.610  0.54172
## channel541                    8.512e-03  8.019e-02   0.106  0.91547
## srch_adults_cnt              -1.017e-01  3.569e-02  -2.849  0.00438 **
## srch_children_cnt            -1.073e-01  3.839e-02  -2.795  0.00518 **
## srch_rm_cnt                   1.239e-01  7.334e-02   1.689  0.09115 .
## srch_destination_id18654373   8.068e-01  6.895e-01   1.170  0.24197
## srch_destination_id18659209   2.576e-01  7.516e-01   0.343  0.73184
## srch_destination_id18667672   5.775e-01  7.915e-01   0.730  0.46563
## srch_destination_id186729088  9.037e-01  7.635e-01   1.184  0.23652
## srch_destination_id18690054  -3.669e-01  9.314e-01  -0.394  0.69366
## srch_destination_id187796077  1.450e+00  7.265e-01   1.996  0.04590 *
## srch_destination_id187952038 -1.308e+01  3.104e+02  -0.042  0.96638
## srch_destination_id190301218 -1.284e+01  3.290e+02  -0.039  0.96887
## srch_destination_id24800154  -7.697e-02  7.862e-01  -0.098  0.92201
## srch_destination_id24801518   1.032e+00  6.500e-01   1.588  0.11233
## srch_destination_id24802324   1.148e-02  7.832e-01   0.015  0.98830
## srch_destination_id24802510   1.239e+00  7.075e-01   1.752  0.07977 .
## srch_destination_id24803657  -1.269e+01  2.942e+02  -0.043  0.96561
## srch_destination_id290562     7.561e-01  8.420e-01   0.898  0.36920
## srch_destination_id33572      1.937e+00  6.640e-01   2.917  0.00353 **
## srch_destination_id4012763    8.870e-01  7.020e-01   1.263  0.20642
## srch_destination_id46251     -6.877e-01  1.173e+00  -0.586  0.55765
## srch_destination_id5525222    1.008e+00  7.181e-01   1.404  0.16030
## srch_destination_id5525315    1.178e+00  6.964e-01   1.691  0.09077 .
## srch_destination_id5525377    2.207e-01  8.452e-01   0.261  0.79405
## srch_destination_id5525532    9.435e-01  6.746e-01   1.399  0.16191
## srch_destination_id5525811    6.943e-01  6.622e-01   1.049  0.29440
## srch_destination_id5525966    2.848e-01  7.898e-01   0.361  0.71845
## srch_destination_id5525997    2.948e-01  8.447e-01   0.349  0.72714
## srch_destination_id5526276    2.715e-01  9.411e-01   0.288  0.77298
## srch_destination_id5526338    5.052e-01  7.942e-01   0.636  0.52470
## srch_destination_id5526400    1.018e+00  7.945e-01   1.282  0.19995
## srch_destination_id5526679    8.753e-01  6.112e-01   1.432  0.15211
## srch_destination_id5526772    8.974e-01  6.582e-01   1.364  0.17272
## srch_destination_id5526803    3.309e-01  6.974e-01   0.474  0.63517
## srch_destination_id5526989    7.174e-01  6.608e-01   1.086  0.27764
## srch_destination_id5527051    4.628e-01  8.458e-01   0.547  0.58426
## srch_destination_id5527144    1.044e+00  7.424e-01   1.406  0.15971
## srch_destination_id5527175    5.028e-01  7.122e-01   0.706  0.48018    
## srch_destination_id5527206    9.277e-01  6.181e-01   1.501  0.13341    
## srch_destination_id5527237    3.758e-01  6.419e-01   0.585  0.55828    
## srch_destination_id5527361    1.387e+00  7.217e-01   1.922  0.05467 .  
## srch_destination_id5527392    5.739e-01  7.920e-01   0.725  0.46870    
## srch_destination_id5527516   -2.629e-01  9.347e-01  -0.281  0.77852    
## srch_destination_id5527547    3.830e-01  6.787e-01   0.564  0.57253    
## srch_destination_id5527578    9.052e-01  6.410e-01   1.412  0.15789    
## srch_destination_id5527640    1.058e+00  6.852e-01   1.544  0.12247    
## srch_destination_id5527733    1.101e-01  9.396e-01   0.117  0.90672    
## srch_destination_id5527857    8.324e-01  7.350e-01   1.133  0.25737    
## srch_destination_id5527888    1.061e+00  7.051e-01   1.505  0.13240    
## srch_destination_id5527981    1.131e+00  6.565e-01   1.723  0.08498 .  
## srch_destination_id5576775    2.324e-01  9.414e-01   0.247  0.80499    
## srch_destination_id5576806    1.064e+00  7.636e-01   1.394  0.16332    
## srch_destination_id5576961    2.313e-01  7.099e-01   0.326  0.74454    
## srch_destination_id5576992    8.456e-01  6.906e-01   1.224  0.22078    
## srch_destination_id5577023    7.227e-01  7.570e-01   0.955  0.33974    
## srch_destination_id5579534    1.383e+00  7.217e-01   1.916  0.05533 .  
## srch_destination_id5579875    4.115e-01  7.298e-01   0.564  0.57285    
## srch_destination_id5579968    2.773e-01  6.780e-01   0.409  0.68252    
## srch_destination_id5580619   -8.528e-01  1.171e+00  -0.728  0.46635    
## srch_destination_id5580774    4.170e-01  9.400e-01   0.444  0.65731    
## srch_destination_id5581115    3.986e-01  8.400e-01   0.475  0.63513    
## srch_destination_id5581518    1.169e+00  7.413e-01   1.578  0.11465    
## srch_destination_id5582386    3.584e-01  8.425e-01   0.425  0.67053    
## srch_destination_id5582510    1.382e+00  7.210e-01   1.916  0.05536 .  
## srch_destination_id5583409    1.985e-02  9.359e-01   0.021  0.98308    
## srch_destination_id5601079    9.894e-01  7.177e-01   1.379  0.16799    
## srch_destination_idOTHER      1.157e+00  5.939e-01   1.948  0.05137 .  
## prop_is_branded1              2.916e-01  5.701e-02   5.115 3.14e-07 ***
## prop_starrating1              6.506e-01  6.716e-01   0.969  0.33265    
## prop_starrating2              7.585e-01  2.910e-01   2.606  0.00915 ** 
## prop_starrating3              6.618e-01  2.823e-01   2.344  0.01908 *  
## prop_starrating4              2.450e-01  2.845e-01   0.861  0.38916    
## prop_starrating5             -3.060e-02  2.953e-01  -0.104  0.91749    
## distance_bandF               -2.252e-01  9.119e-02  -2.469  0.01354 *  
## distance_bandM               -1.388e-02  6.591e-02  -0.211  0.83324    
## distance_bandVC              -9.638e-02  8.307e-02  -1.160  0.24594    
## distance_bandVF               4.017e-02  1.095e-01   0.367  0.71379    
## hist_price_bandL             -3.278e-02  8.967e-02  -0.366  0.71469    
## hist_price_bandM             -3.035e-02  7.413e-02  -0.409  0.68222    
## hist_price_bandVH             1.435e-01  1.024e-01   1.401  0.16109    
## hist_price_bandVL            -2.458e-01  1.189e-01  -2.066  0.03881 *  
## popularity_bandL             -8.247e-01  1.830e-01  -4.506 6.61e-06 ***
## popularity_bandM             -1.177e-01  7.148e-02  -1.646  0.09974 .  
## popularity_bandVH             3.438e-01  6.389e-02   5.381 7.42e-08 ***
## popularity_bandVL            -6.220e-01  4.277e-01  -1.454  0.14582    
## cnt                          -2.596e+00  1.873e-01 -13.864  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
```

```
##     Null deviance: 11883  on 19999  degrees of freedom
## Residual deviance: 10591  on 19903  degrees of freedom
## AIC: 10785
##
## Number of Fisher Scoring iterations: 15
```

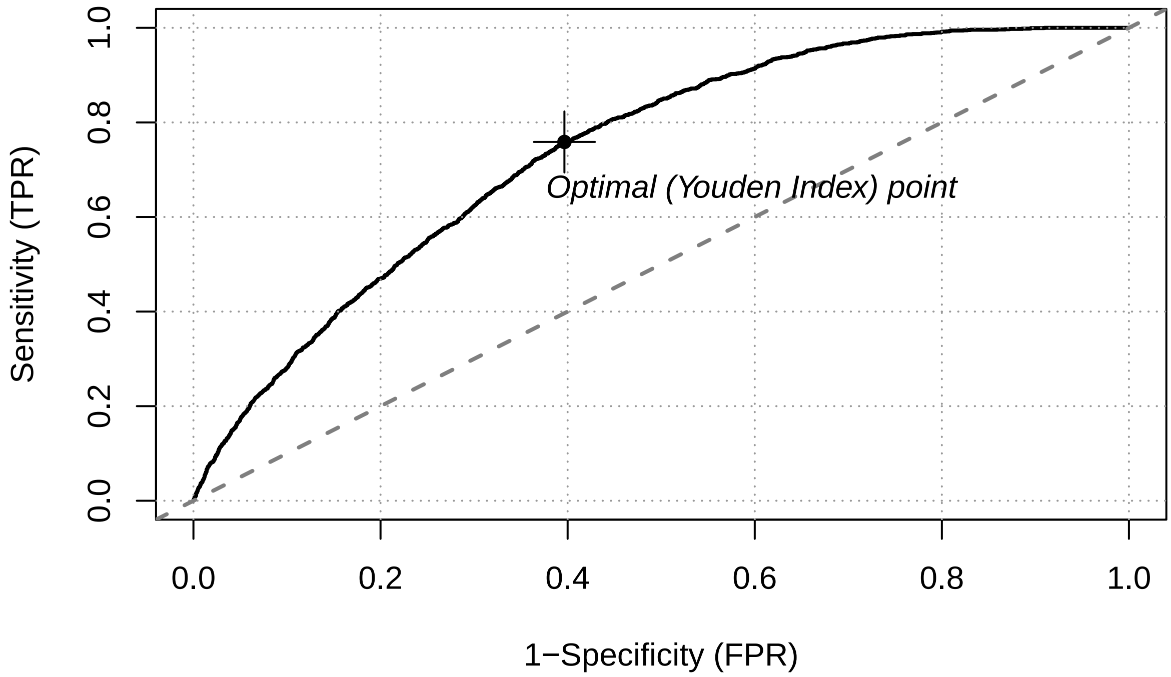
```
class <- travelBest$y
score <- qlogis(travelBest$fitted.values)
travelEmp <- rocit(score=score, class=class, method="emp")
travelBin <- rocit(score = score,class = class,method = "bin")
travelNon <- rocit(score = score,class = class,method = "non")
```

Interpretation

```
summary(travelEmp)
```

```
##
##  Method used: empirical
##  Number of positive(s): 1753
##  Number of negative(s): 18247
##  Area under curve: 0.74
```

```
plot(travelEmp, col = c(1,"gray50"),legend = FALSE, YIndex = TRUE)
```



```
cbind( exp(coef(travelBest)))
```

```
##                                         [,1]
## (Intercept)                     4.252310e-01
## is_package1                  4.411923e-01
## channel262                   8.945338e-01
## channel293                   6.655789e-01
## channel324                   1.126954e+00
## channel355                   1.624210e+00
## channel386                   1.388941e+00
## channel417                   3.140328e-07
## channel448                   4.024954e-01
## channel479                   1.825168e+00
## channel510                   9.431217e-01
## channel541                   1.008548e+00
## srch_adults_cnt              9.033139e-01
## srch_children_cnt            8.982551e-01
## srch_rm_cnt                  1.131891e+00
## srch_destination_id18654373  2.240677e+00
## srch_destination_id18659209  1.293763e+00
## srch_destination_id18667672  1.781494e+00
## srch_destination_id186729088 2.468841e+00
## srch_destination_id18690054  6.929070e-01
## srch_destination_id187796077 4.264181e+00
## srch_destination_id187952038 2.078383e-06
## srch_destination_id190301218 2.651582e-06
## srch_destination_id24800154  9.259180e-01
## srch_destination_id24801518  2.806752e+00
## srch_destination_id24802324  1.011546e+00
## srch_destination_id24802510  3.453792e+00
## srch_destination_id24803657  3.096720e-06
## srch_destination_id290562    2.129928e+00
## srch_destination_id33572     6.939017e+00
## srch_destination_id4012763   2.427782e+00
## srch_destination_id46251     5.027157e-01
## srch_destination_id5525222   2.740968e+00
## srch_destination_id5525315   3.247355e+00
## srch_destination_id5525377   1.246887e+00
## srch_destination_id5525532   2.568951e+00
## srch_destination_id5525811   2.002338e+00
## srch_destination_id5525966   1.329450e+00
## srch_destination_id5525997   1.342806e+00
## srch_destination_id5526276   1.311918e+00
## srch_destination_id5526338   1.657354e+00
## srch_destination_id5526400   2.768649e+00
## srch_destination_id5526679   2.399532e+00
## srch_destination_id5526772   2.453332e+00
## srch_destination_id5526803   1.392188e+00
## srch_destination_id5526989   2.049017e+00
## srch_destination_id5527051   1.588524e+00
## srch_destination_id5527144   2.840080e+00
## srch_destination_id5527175   1.653364e+00
## srch_destination_id5527206   2.528595e+00
## srch_destination_id5527237   1.456122e+00
## srch_destination_id5527361   4.001875e+00
## srch_destination_id5527392   1.775195e+00
## srch_destination_id5527516   7.688285e-01
## srch_destination_id5527547   1.466689e+00
## srch_destination_id5527578   2.472515e+00
## srch_destination_id5527640   2.881344e+00
## srch_destination_id5527733   1.116392e+00
## srch_destination_id5527857   2.298935e+00
## srch_destination_id5527888   2.889056e+00
## srch_destination_id5527981   3.098275e+00
## srch_destination_id5576775   1.261647e+00
## srch_destination_id5576806   2.899130e+00
## srch_destination_id5576961   1.260262e+00
## srch_destination_id5576992   2.329433e+00
## srch_destination_id5577023   2.059948e+00
## srch_destination_id5579534   3.986702e+00
## srch_destination_id5579875   1.509119e+00
## srch_destination_id5579968   1.319585e+00
## srch_destination_id5580619   4.262161e-01
## srch_destination_id5580774   1.517443e+00
## srch_destination_id5581115   1.489747e+00
## srch_destination_id5581518   3.220344e+00
## srch_destination_id5582386   1.431033e+00
## srch_destination_id5582510   3.980958e+00
## srch_destination_id5583409   1.020044e+00
## srch_destination_id5601079   2.689670e+00
## srch_destination_idOTHER     3.180959e+00
## prop_is_branded1             1.338542e+00
## prop_starrating1             1.916787e+00
## prop_starrating2             2.135142e+00
## prop_starrating3             1.938333e+00
## prop_starrating4             1.277634e+00
## prop_starrating5             9.698671e-01
## distance_bandF               7.983773e-01
## distance_bandM               9.862190e-01
## distance_bandVC              9.081178e-01
## distance_bandVF              1.040986e+00
## hist_price_bandL             9.677517e-01
## hist_price_bandM             9.701043e-01
## hist_price_bandVH            1.154313e+00
## hist_price_bandVL            7.821105e-01
## popularity_bandL             4.383746e-01
## popularity_bandM             8.889892e-01
## popularity_bandVH            1.410300e+00
## popularity_bandVL            5.368579e-01
## cnt                          7.453454e-02
```

using a step wise model building process we can create a model that can accurately predict if someone is willing to book a hotel with an AUC of ~0.74 at a threshold of close to 0.7. The base model is assuming an unpackaged hotel for 0 rooms, 0 adults, 0 children, on channel 231, search destination 18652358, unbranded, with a 0 star randing, at C distance, H price, H popularity, and a count of 0 interactions on the page. While this user likely does not exist, it indicates that a user is only 42.5% likely to book at hotel at the start of their experience.

Packaged bookings are 44.1% likely to book

And teh channel that a user books with have a high impact on the probability of booking. Channel 324,355,386, and 479 all show an increase in the chance of making a booking over the base odds. The worst was Channel 417 which should a near 0% likelihood of booking a hotel.

As the number of adutls increased, the likely hood of booking a hotel decreased by about 10% per adult. As the number of children increased, the likelihood of booking a hotel decreased by about 11% per child.

Interestingly, as the number of rooms increased, the likelihood of booking increased by 13% per room. This is likely due to the relative rarity of booking multiple hotel rooms for a standard trip. In the case that multiple rooms need to be booked, there are likely external pressures involved that require the individual to make a booking wherever there is space (such as a sports trip or something siimlar).

Search destination was also a key factor in predicting if a booking took place or not. Compared to the intial destination, most places had a significantly higher chance of attracking bookings, with the highest being destination 5527361 which was 4x more likely to lead to a booking than any other place. Booking 5576534 was also close to 4x more likely to lead to a booking, indicating that destination played a substantial role in bookings. As it's not possible to tell what these destinations are from the id, my best guess is that they are relatively common vacation or conference locations. We also see that the 'other' category we created is 3x more likely to attract bookings over the base location. This surprises me as so many of those locations were rarely searched, but it's possible these could be smaller locations where people go to visit family or something to that effect. It's hard to tell without more information about the reasons someone would be travelling.

Distance bands that were very far from the center were more 4% more likely to get a booking, which is a little surprising. Though most of the bands were between 0% and 10% as likely to book as the base rate, with the only exception being the 'Far' band which was 21% less likely to get a booking. This may be due to the types of hotels that are very from other other hotels being seen as more exclusive or higher class. Or they are really cheap and the price factor wins out. It's hard to tell without more information.

Branded hotels were 33% more likely to get a booking compared to non branded hotels, signifying people tend to prioritise a 'known' chain over an unknown.

A hotel with any number of stars was more likely to get a booking than one without stars, except for 5-star hotels. They were 3% less likely to get a booking, likely due to their price. However, there is likely something else at play, as for the price levels, only the Very Highly priced hotels were more likely to get a booking compared to the Highly Priced hotels. This really surprised me as I would assume that cheaper hotels would get more bookings but that doesn't seem to be the case. There may be other effects at play, but it's hard to tell without knowing what was considered a high price.

hotel popularity only helped if the hotel was in the extrmely popular category compared to other hotels nearby. This seems to be a bit of a self-fulfilling prophecy. A hotel is very popular and will get more bookings, but it's also 41% more likely to be boked because of how popular it is. We can't attribute causation either way, by default we should expect that more popular hotels are more likely to be booked, and we see that hold.

Lastly, the number of events/clicks in one user session reduced the likelihood of booking by a 25% per event. To me, this suggests that users would come back across multiple sessions, finding the best prices maybe in a long session without booking, only to come back later and to book exactly what they wanted right away. It seems that it would be best to encourage peopel to book as quickly as possible, as the longer they are on the site and the more they look, the less likely they are to a booking.

Question 3 - BoneDesnity

Preprocessing

Data is converted to factors, NA values are dropped due to their relative infrequency. Allc is dropped due to being an id value. Frx and Nosp are droppped due to being indicator of numNosp.

```
boneData<-  read.delim("FITglm2.txt",sep="\t")
boneData[,c(3,4,6,11,13,14,15,16,17,18)] <- lapply(boneData[,c(3,4,6,11,13,14,15,16,17,18)], factor)
boneData = boneData[complete.cases(boneData),]
boneData = boneData[,-c(1, 3,4)]
summary(boneData)
```

```
##      ra_age         numnosp       trt01       p3_weigh         htotbmd
##  Min.   :54.00   Min.   :0.0000   0:3177   Min.   : 36.30   Min.   :0.3700
##  1st Qu.:64.00   1st Qu.:0.0000   1:3189   1st Qu.: 56.90   1st Qu.:0.6350
##  Median :68.00   Median :0.0000            Median : 63.10   Median :0.6980
##  Mean   :68.12   Mean   :0.1528            Mean   : 64.54   Mean   :0.6925
##  3rd Qu.:73.00   3rd Qu.:0.0000            3rd Qu.: 70.80   3rd Qu.:0.7540
##  Max.   :81.00   Max.   :4.0000            Max.   :124.60   Max.   :0.9860
##       nbmd           trialyrs                riskcat4        tneck
##  Min.   :0.3370   Min.   :0.005476               :   0   Min.   :-4.342
##  1st Qu.:0.5420   1st Qu.:3.014374   0: LOW RISK :5239   1st Qu.:-2.633
##  Median :0.5900   Median :4.027379   1: HIGH RISK:1127   Median :-2.233
##  Mean   :0.5842   Mean   :3.795054                       Mean   :-2.282
##  3rd Qu.:0.6350   3rd Qu.:4.484600                       3rd Qu.:-1.858
##  Max.   :0.7830   Max.   :4.821355                       Max.   :-0.625
##  bmd25    hplac    htrt     lplac    ltrt
##  0:4312   0:5800   0:5805   0:3755   0:3738
##  1:2054   1: 566   1: 561   1:2611   1:2628
##
##
##
##
##                                  rtgroup
##                                      :   0
##  1:HIGH FALL RISK, PLACEBO GROUP  : 566
##  2:HIGH FALL RISK, TREATMENT GROUP: 561
##  3:LOW FALL RISK, PLACEBO GROUP   :2611
##  4:LOW FALL RISK, TREATMENT GROUP :2628
##
```

```
boneModel <- glm(numnosp~., data=boneData, family="poisson")
#bestBone <- stepAIC(boneModel, trace=0)
bestBone<- glm(formula = numnosp ~ ra_age + trt01 + p3_weigh + htotbmd +
    trialyrs + riskcat4 + bmd25, family = "poisson", data = boneData)

pchisq(bestBone$deviance, df=bestBone$df.residual, lower.tail = FALSE, log.p=TRUE)
```

```
## [1] -7.998056e-130
```

```
summary(bestBone)
```

```
##
## Call:
## glm(formula = numnosp ~ ra_age + trt01 + p3_weigh + htotbmd +
##     trialyrs + riskcat4 + bmd25, family = "poisson", data = boneData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9892  -0.5848  -0.5126  -0.4462   4.7979
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.456423   0.609678  -4.029 5.60e-05 ***
## ra_age                0.010998   0.005593   1.966 0.049262 *
## trt011               -0.146430   0.064322  -2.277 0.022814 *
## p3_weigh              0.015164   0.003083   4.918 8.75e-07 ***
## htotbmd              -2.746337   0.506964  -5.417 6.05e-08 ***
## trialyrs              0.175405   0.046173   3.799 0.000145 ***
## riskcat41: HIGH RISK  0.184264   0.079613   2.315 0.020640 *
## bmd251                0.217483   0.086210   2.523 0.011646 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4115.6  on 6365  degrees of freedom
## Residual deviance: 4003.1  on 6358  degrees of freedom
## AIC: 5741.1
##
## Number of Fisher Scoring iterations: 6
```

```
cbind(exp(coef(bestBone))-1)
```

```
##                             [,1]
## (Intercept)          -0.91425889
## ra_age                0.01105909
## trt011               -0.13621418
## p3_weigh              0.01527927
## htotbmd              -0.93583757
## trialyrs              0.19172842
## riskcat41: HIGH RISK  0.20233372
## bmd251                0.24294404
```

Interpretation

Stepwise model is again used to create a model. The model is significant according to a chisq test of the model deviances. A model with ineractions could not be completed using the stepwise method.

The base case indicates an age of 0 years, no treatment, 0bmd, low risk category, and without osteoperosis. For very year old, a woman increases here number of non-spinal fractures by 0.01. If she is treated, she decreases her count by 0.13. For every pound over 100 she increases her count by 0.015 fractures, but the closer her bone mass density is to 1 the less likely she is to experience a fracture, which is in line with expectation. Women that were being followed up on generally experienced an additional 0.2 fractures per period they were folowed. High risk women experienced 0.2 more fractures than low risk women, and women with osteoperosis increased the number of fractures by 0.25.

The model findings are in line with expectations. Lighter weight, younger women, with higher bone density are very unlikely to have any non spinal fractures. Those being treated for osteoperosis or low bone density also experience fewer fractures. Most women start with a low chance of fracture occuring, which is in line with the mean number of fractures being 0.15 across all participants included.

Question 4 - Wine Data

Preprocessing

The wine data came preprocessed.We end up dropping everything other than Alcohol, Ash, Alcalinity, Phenols, Flavanoids, and Proanthocyanins as the rest are not significant predictors and adversely affect the predictive ability of the model.

```
wineData <- wine[,c(1,2,4,5,7,8,10)]
summary(wineData)
```

```
##  Type      Alcohol          Ash           Alcalinity       Phenols
##  1:59   Min.   :11.03   Min.   :1.360   Min.   :10.60   Min.   :0.980
##  2:71   1st Qu.:12.36   1st Qu.:2.210   1st Qu.:17.20   1st Qu.:1.742
##  3:48   Median :13.05   Median :2.360   Median :19.50   Median :2.355
##         Mean   :13.00   Mean   :2.367   Mean   :19.49   Mean   :2.295
##         3rd Qu.:13.68   3rd Qu.:2.558   3rd Qu.:21.50   3rd Qu.:2.800
##         Max.   :14.83   Max.   :3.230   Max.   :30.00   Max.   :3.880
##    Flavanoids    Proanthocyanins
##  Min.   :0.340   Min.   :0.410
##  1st Qu.:1.205   1st Qu.:1.250
##  Median :2.135   Median :1.555
##  Mean   :2.029   Mean   :1.591
##  3rd Qu.:2.875   3rd Qu.:1.950
##  Max.   :5.080   Max.   :3.580
```

```
wineModel <- multinom(Type ~. ,data = wineData, maxit = 1000)
```

```
## # weights:  24 (14 variable)
## initial  value 195.552987
## iter  10 value 44.221495
## iter  20 value 9.251744
## iter  30 value 6.433651
## iter  40 value 4.282992
## iter  50 value 3.560243
## iter  60 value 3.426442
## iter  70 value 3.415113
## iter  80 value 3.384415
## iter  90 value 3.304622
## iter 100 value 3.232115
## iter 110 value 3.181833
## iter 120 value 3.168420
## iter 130 value 3.150584
## iter 140 value 3.112894
## iter 150 value 3.051188
## iter 160 value 2.963782
## iter 170 value 2.938810
## iter 180 value 2.932499
## iter 190 value 2.917349
## iter 200 value 2.875854
## iter 210 value 2.761784
## iter 220 value 2.650254
## iter 230 value 2.602950
## iter 240 value 2.599994
## iter 250 value 2.581361
## iter 260 value 2.520786
## iter 270 value 2.465096
## iter 280 value 2.439120
## iter 290 value 2.420140
## iter 300 value 2.416344
## iter 310 value 2.396167
## iter 320 value 2.316940
## iter 330 value 2.268689
## iter 340 value 2.216879
## iter 350 value 2.198561
## iter 360 value 2.196020
## iter 370 value 2.182690
## iter 380 value 2.158483
## iter 390 value 2.129171
## iter 400 value 2.113066
## iter 410 value 2.110033
## iter 420 value 2.106255
## iter 430 value 2.093317
## iter 440 value 2.076424
## iter 450 value 2.065032
## iter 460 value 2.052842
## iter 470 value 2.050928
## iter 480 value 2.043868
## iter 490 value 2.031700
## iter 500 value 2.013810
## iter 510 value 1.999114
## iter 520 value 1.995490
## iter 530 value 1.992716
## iter 540 value 1.985942
## iter 550 value 1.962111
## iter 560 value 1.936499
## iter 570 value 1.920844
## iter 580 value 1.918417
## iter 590 value 1.915311
## iter 600 value 1.906795
## iter 610 value 1.896329
## iter 620 value 1.885069
## iter 630 value 1.874986
## iter 640 value 1.864931
## iter 650 value 1.857984
## iter 660 value 1.854489
## iter 670 value 1.845276
## iter 680 value 1.836391
## iter 690 value 1.832895
## iter 700 value 1.825035
## iter 710 value 1.816415
## iter 720 value 1.812281
## iter 730 value 1.802631
## iter 740 value 1.797247
## iter 750 value 1.791405
## iter 760 value 1.785987
## iter 770 value 1.781405
## iter 780 value 1.774296
## iter 790 value 1.771983
## iter 800 value 1.764443
## iter 810 value 1.761038
## iter 820 value 1.755158
## iter 830 value 1.749394
## iter 840 value 1.744569
## iter 850 value 1.741453
## iter 860 value 1.735346
## iter 870 value 1.725778
## iter 880 value 1.715955
## iter 890 value 1.708961
## iter 900 value 1.701639
## iter 910 value 1.697797
## iter 920 value 1.692367
## iter 930 value 1.680937
## iter 940 value 1.675308
## iter 950 value 1.670356
## iter 960 value 1.666745
## iter 970 value 1.663378
## iter 980 value 1.657840
## iter 990 value 1.650692
## iter1000 value 1.645459
## final  value 1.645459
## stopped after 1000 iterations
```

```
modelSummary <- summary(wineModel)

z <- modelSummary$coefficients/modelSummary$standard.errors
p <- (1-pnorm(abs(z),0,1))*2 # I am using two-tailed z test
typeResults <- rbind(modelSummary$coefficients[2, ],modelSummary$standard.errors[2, ],z[2, ],p[2, ])
typeResults
```

```
##      (Intercept)    Alcohol          Ash   Alcalinity     Phenols    Flavanoids
## [1,]  -187.02748 12.7520058 -12.4547570 18.471784577 -21.1815934 -151.86641248
## [2,]    11.32875 10.2635747  24.4040776  5.733084711  13.0494900   66.92951629
## [3,]   -16.50911  1.2424527  -0.5103556  3.221962610  -1.6231740   -2.26904990
## [4,]     0.00000  0.2140696   0.6098024  0.001273157   0.1045522    0.02326529
##      Proanthocyanins
## [1,]   -33.891405981
## [2,]    11.484645722
## [3,]    -2.951018847
## [4,]     0.003167276
```

```
modelSummary
```

```
## Call:
## multinom(formula = Type ~ ., data = wineData, maxit = 1000)
##
## Coefficients:
##   (Intercept)   Alcohol       Ash Alcalinity   Phenols Flavanoids
## 2    579.4334 -42.09109 -83.00739   10.59658  19.00547  -39.76319
## 3   -187.0275  12.75201 -12.45476   18.47178 -21.18159 -151.86641
##   Proanthocyanins
## 2        11.79003
## 3       -33.89141
##
## Std. Errors:
##   (Intercept)   Alcohol      Ash Alcalinity  Phenols Flavanoids Proanthocyanins
## 2    12.86216  2.750265 11.04432   1.342911 11.84245   9.421742        8.372075
## 3    11.32875 10.263575 24.40408   5.733085 13.04949  66.929516       11.484646
##
## Residual Deviance: 3.290919
## AIC: 31.29092
```

```
cbind(exp(modelSummary$coefficients))
```

```
##     (Intercept)      Alcohol          Ash   Alcalinity      Phenols
## 2 4.413022e+251 5.248926e-19 8.919638e-37     39997.69 1.794608e+08
## 3  5.956578e-82 3.452437e+05 3.899130e-06 105243216.52 6.323403e-10
##     Flavanoids Proanthocyanins
## 2 5.383516e-18    1.319299e+05
## 3 1.109826e-66    1.910510e-15
```

Interpretation

Many of the variables were not significant in the modelling of the type of wine. Even so, we are able to find a few significant variables by manually creating the model and evaluating a normalised p-value of the coefficients. However, in doing so, we see the effect sizes and directions shift drastically from adding or removing single variables. The model selected is the closest I could get, despite some extreme predictive values for the types.

By Default, the type 2 wine is extremely more likely to be chosen than any other wine, and type 3 is drastically less likely to be the chosen wine.

As the alcohol content increases type two becomes extremely less likely to be the proper wine than type 1, and type 3 wine becomes more likely to be the wine over type one.

For ash, both Type 2 and Type 3 wine are very less likely to be the type of wine predicted as the ash content increases. For alcalinity, it is the opposite. As the alcalinity increases, type 2 wine is much more likely than type 1, and type 3 is even more likely than type 1.

As phenols increase, Type 2 is much more likely than type 1, and type 3 is much less likely than type 1. Flavanoids share a similar trend to Ash in that both Type 2 and 3 are much less likely as the amount increases. Lastly, Proanthocyanins mimic Phenols in that Type 2 is more likely as they increase, and type 3 is much less likely as they increase.

Overall, this model has too many weird effects and the values are too extreme to be reasonable. I do not think it is a good predictor of wine type and I'm sure it can be improved.

Question 5 - Lung Cancer

Preprocecssing

institution, status, age, sex, ecog are converted to factors. There are too few observations to safely drop NA's. meal.cal may be dropped due to tis abnormally high numbers of NAs. Time may also be dropped as it encodes much of the data contained in status. Instituion is later dropped due to not being a significant predictor. Time was also dropped due to encoding too much of the surivaval status

```r
lungData <- lung
summary(lungData)
```

```
##       inst            time            status           age       
##  Min.   : 1.00   Min.   :   5.0   Min.   :1.000   Min.   :39.00  
##  1st Qu.: 3.00   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00  
##  Median :11.00   Median : 255.5   Median :2.000   Median :63.00  
##  Mean   :11.09   Mean   : 305.2   Mean   :1.724   Mean   :62.45  
##  3rd Qu.:16.00   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00  
##  Max.   :33.00   Max.   :1022.0   Max.   :2.000   Max.   :82.00  
##  NA's   :1                                                       
##       sex           ph.ecog          ph.karno        pat.karno     
##  Min.   :1.000   Min.   :0.0000   Min.   : 50.00   Min.   : 30.00  
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00  
##  Median :1.000   Median :1.0000   Median : 80.00   Median : 80.00  
##  Mean   :1.395   Mean   :0.9515   Mean   : 81.94   Mean   : 79.96  
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00  
##  Max.   :2.000   Max.   :3.0000   Max.   :100.00   Max.   :100.00  
##                  NA's   :1        NA's   :1        NA's   :3       
##     meal.cal         wt.loss      
##  Min.   :  96.0   Min.   :-24.000  
##  1st Qu.: 635.0   1st Qu.:  0.000  
##  Median : 975.0   Median :  7.000  
##  Mean   : 928.8   Mean   :  9.832  
##  3rd Qu.:1150.0   3rd Qu.: 15.750  
##  Max.   :2600.0   Max.   : 68.000  
##  NA's   :47       NA's   :14
```

```
lungData[,c(1,3,5,6)] = lapply(lungData[,c(1,3,5,6)], factor)
lungData = lungData[,-c(1,2)]
lungData<- na.omit(lungData)
lungModel <- glm(status~.^2, data=lungData, family=binomial, control = list(maxit = 100))
lungStep<- stepAIC(lungModel, trace=0)
lungBest<- glm(formula = status ~ sex + ph.ecog + ph.karno + pat.karno +
    meal.cal + wt.loss + sex:ph.ecog + sex:ph.karno + sex:pat.karno +
    sex:meal.cal + sex:wt.loss + ph.ecog:pat.karno + ph.ecog:meal.cal +
    ph.ecog:wt.loss + ph.karno:wt.loss + pat.karno:meal.cal +
    meal.cal:wt.loss, family = binomial, data = lungData, control = list(maxit = 100))
summary(lungBest)
```

```
## 
## Call:
## glm(formula = status ~ sex + ph.ecog + ph.karno + pat.karno + 
##     meal.cal + wt.loss + sex:ph.ecog + sex:ph.karno + sex:pat.karno + 
##     sex:meal.cal + sex:wt.loss + ph.ecog:pat.karno + ph.ecog:meal.cal + 
##     ph.ecog:wt.loss + ph.karno:wt.loss + pat.karno:meal.cal + 
##     meal.cal:wt.loss, family = binomial, data = lungData, control = list(maxit = 100))
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1730  -0.2341   0.1810   0.6343   1.8861  
## 
## Coefficients: (4 not defined because of singularities)
##                        Estimate Std. Error z value Pr(>|z|)   
## (Intercept)          -1.279e+01  9.847e+00  -1.299 0.193882   
## sex2                 -9.509e+00  8.556e+00  -1.111 0.266426   
## ph.ecog1             -1.952e+00  6.834e+00  -0.286 0.775159   
## ph.ecog2              5.505e+00  8.109e+00   0.679 0.497238   
## ph.ecog3              6.344e+00  1.455e+03   0.004 0.996522   
## ph.karno              6.865e-02  4.502e-02   1.525 0.127317   
## pat.karno             1.488e-01  1.015e-01   1.466 0.142573   
## meal.cal              1.127e-02  5.648e-03   1.994 0.046101 * 
## wt.loss               8.500e-01  3.564e-01   2.385 0.017077 * 
## sex2:ph.ecog1         6.992e+00  2.142e+00   3.263 0.001100 **
## sex2:ph.ecog2        -4.712e+00  5.206e+00  -0.905 0.365387   
## sex2:ph.ecog3                NA         NA      NA       NA   
## sex2:ph.karno         1.479e-01  8.484e-02   1.743 0.081340 . 
## sex2:pat.karno       -1.616e-01  6.886e-02  -2.347 0.018900 * 
## sex2:meal.cal         5.183e-03  1.990e-03   2.605 0.009200 **
## sex2:wt.loss          1.037e-01  5.728e-02   1.810 0.070310 . 
## ph.ecog1:pat.karno -1.047e-02  7.136e-02  -0.147 0.883333   
## ph.ecog2:pat.karno  2.510e-01  1.033e-01   2.429 0.015135 * 
## ph.ecog3:pat.karno          NA         NA      NA       NA   
## ph.ecog1:meal.cal   2.952e-03  2.073e-03   1.424 0.154345   
## ph.ecog2:meal.cal  -8.764e-03  3.442e-03  -2.546 0.010885 * 
## ph.ecog3:meal.cal           NA         NA      NA       NA
## ph.ecog1:wt.loss   -2.378e-01  8.041e-02  -2.957 0.003104 **
## ph.ecog2:wt.loss   -7.278e-01  2.138e-01  -3.403 0.000666 ***
## ph.ecog3:wt.loss           NA         NA      NA       NA
## ph.karno:wt.loss   -9.368e-03  3.852e-03  -2.432 0.015002 *
## pat.karno:meal.cal -1.879e-04  6.675e-05  -2.815 0.004881 **
## meal.cal:wt.loss    1.745e-04  7.599e-05   2.297 0.021641 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 199.16  on 167  degrees of freedom
## Residual deviance: 118.01  on 144  degrees of freedom
## AIC: 166.01
##
## Number of Fisher Scoring iterations: 14
```

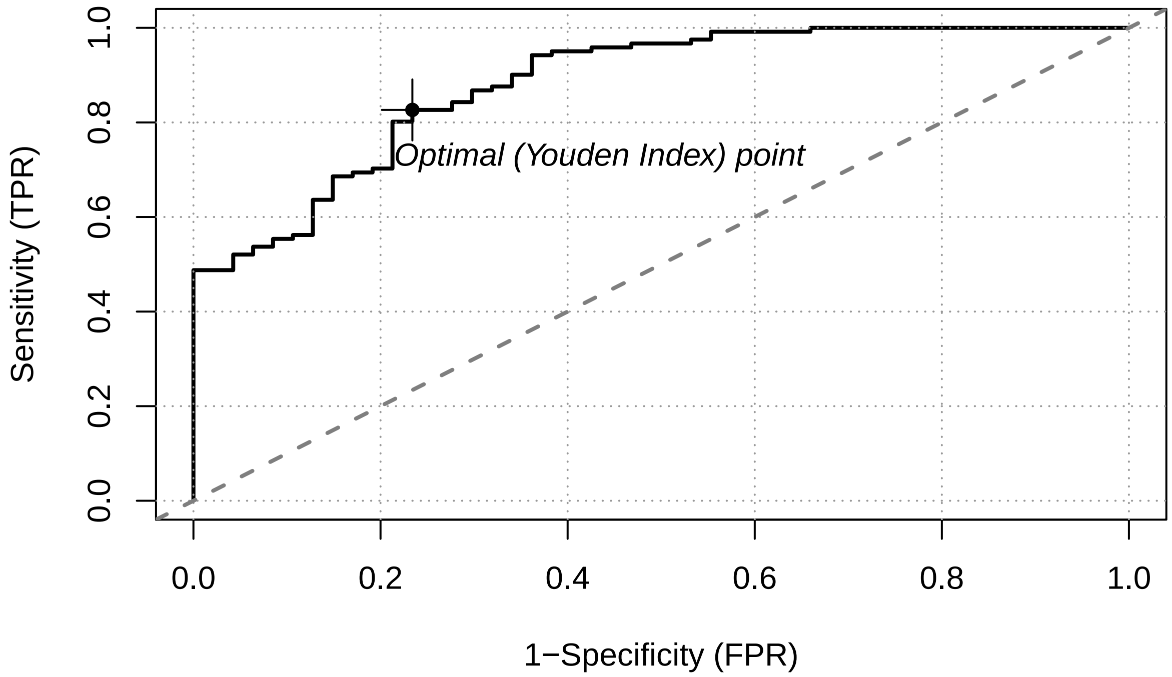
```
lungclass <- lungBest$y
lungscore <- qlogis(lungBest$fitted.values)
lungEmp <- rocit(score = lungscore, class=lungclass, method="emp")
```

Interpretation

```
summary(lungEmp)
```

```
##
##  Method used: empirical
##  Number of positive(s): 121
##  Number of negative(s): 47
##  Area under curve: 0.881
```

```
plot(lungEmp, col = c(1,"gray50"),legend = FALSE, YIndex = TRUE)
```



```r
cbind(exp(coef(lungBest)))
```

```
##                            [,1]
## (Intercept)        2.779271e-06
## sex2               7.419291e-05
## ph.ecog1           1.419946e-01
## ph.ecog2           2.458761e+02
## ph.ecog3           5.692581e+02
## ph.karno           1.071057e+00
## pat.karno          1.160402e+00
## meal.cal           1.011329e+00
## wt.loss            2.339600e+00
## sex2:ph.ecog1      1.087647e+03
## sex2:ph.ecog2      8.985484e-03
## sex2:ph.ecog3                NA
## sex2:ph.karno      1.159360e+00
## sex2:pat.karno     8.507416e-01
## sex2:meal.cal      1.005196e+00
## sex2:wt.loss       1.109229e+00
## ph.ecog1:pat.karno 9.895825e-01
## ph.ecog2:pat.karno 1.285357e+00
## ph.ecog3:pat.karno           NA
## ph.ecog1:meal.cal  1.002956e+00
## ph.ecog2:meal.cal  9.912746e-01
## ph.ecog3:meal.cal            NA
## ph.ecog1:wt.loss   7.883641e-01
## ph.ecog2:wt.loss   4.829813e-01
## ph.ecog3:wt.loss             NA
## ph.karno:wt.loss   9.906755e-01
## pat.karno:meal.cal 9.998121e-01
## meal.cal:wt.loss   1.000175e+00
```

The model is relatively small and much easier to look at the stepwise effects as a result. However, the number of variables in the step wise model lead to an extremely high AUC, which suggests overfitting took place. The AUC was 0.881 at a threshold of close to 0.8. The base incidence rate is looking at asymptomatic males with low KARNO scores, eating no calroies and with no weight loss.

The primary risk factors are being male, with risk increasing as the ECOG performance gets worse (gets higher). The Karno Score indicates an increase risk on it sown which is not inline with expectations. Even though the increas is small, you would expect that the better a physician rates you the lower your risk factors are for chance of death.

The amount of calories you eat does not have a large effect on your risk factors, but the amount of weightloss does have a large increase in the risk of death from lung cancer. Surprisingly, age is not a major factor in risk of death due to lung cancer.

There are some intesreting effects between the variable as well. What stands out is that asymptomatic females are slightly more at risk than asymptomatic males. However, their patient KARNO scores often indicate a lesser risk, which could mean that they understate how healthty they are in their own mind, while their physicians may over state how healthy they are compared to males. Females also have an additional risk factor due to weight loss than males.

The interaction of severity of symptoms and KARNO scores doesn't seem to increase or decrease risk factors in a meaningful way, same with the calories. However, the ECOG score and the weight loss did have a negative effect on the risk factoris for asymptomatic and nearly asymptomatic individuals suggesting that their weight loss may not be a result of the disease or an indicator of increased risk at that time. Calorie intake had very limited interactions.

Overall the risk factors are pretty much in line with expectations outside of the KARNO scores seemingly being reversed. Weight loos and increased severity of symptoms increase the risk of death. Females are much less at risk to begin with but have some more complex interactions in their risk factors that may not be accounted for in this model.