Target Hospital Analysis

Alice Zormelo

Introduction

An orthopedic equipment company is very interested in finding target hospitals or potential ways to increase equipment sales to hospitals in the United States.

Objective

Analyze orthopedic data from 4,703 hospitals to provide a model that selects hospitals where a high number of orthopedic equipment sales is expected.

Note: A subset of hospitals from specific states can be used instead for the analysis, but it has to be between 500 and 800 hospitals.

What we will do in this analysis

- Perform exploratory data analysis to assess what model to use.
- Go through the standard model building procedures.
- Evaluate the final model's classification capabilities.

Variables

ZIP : US postal code

HID : Hospital ID

CITY : City Name

STATE : State Name

BEDS : # of hospital beds

RBEDS : # of rehabilitation beds

OUT-V : # of outpatient visits

ADM : Administrative Cost (in $1000's per year)

SIR : Revenue from Inpatient

SALESY : Sales of rehabilitation equipment since January 1st

SALES12 : Sales of rehabilitation equipment for the last 12 months

HIP95 : # of hip operations in 1995

KNEE95 : # of knee operations in 1995

Teach : Teaching hospital? 0 or 1

TRAUMA : Do they have a trauma unit? 0 or 1

REHAB : Do they have a rehabilitation unit? 0 or 1

HIP96 : # of hip operations in 1996

KNEE96 : # of knee operations in 1996

FEMUR96 : # of femur operations in 1996

Loading Data

First, let's read in the data. In this analysis, we will only focus on some of the west coast states which are California, Oregon, and Washington state.

```r
ortho_data <- read.table("ortho.txt", header = T)

states <- c("CA", "WA", "OR")

west_coast <- ortho_data %>% filter(STATE %in% states)
```

Based on the summary statistics, a lot of the variables seem skewed and unbalanced. In addition, there are 3 categorical variables that will be used in the analysis; Teach, TRAUMA, and REHAB. Furthermore, no missing values are present.

```r
head(west_coast, 5)
```

```
##     ZIP    HID       CITY STATE BEDS RBEDS   OUT   ADM   Rev SALESY SALES12
## 1 90007 166093 LosAngeles    CA  158     0     0  2026  3226     31      56
## 2 90017 154093 LosAngeles    CA  357    23  9125 12776  6094      9       9
## 3 90024 175593 LosAngeles    CA  610     0 17155 21753 12310     64      64
## 4 90027 156093 LosAngeles    CA  504     0     0 24654 13876     57      57
## 5 90027 168093 LosAngeles    CA  306    15 36500 17608  7211      1       5
##   HIP95 KNEE95 Teach TRAUMA REHAB HIP96 KNEE96 FEMUR96
## 1   176     70     1      0     0   158     61      49
## 2   131     64     1      0     1   134     51      86
## 3   160     84     1      1     0   136     97     125
## 4   151     68     1      0     0   138     92     122
## 5    44      5     1      0     1    48      7      95
```

```r
summary(west_coast)
```

```
##       ZIP            HID                CITY              STATE          
##  Min.   :90004   Length:589         Length:589         Length:589        
##  1st Qu.:92104   Class :character   Class :character   Class :character  
##  Median :94063   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :94188                                                           
##  3rd Qu.:95932                                                           
##  Max.   :99403                                                           
##       BEDS            RBEDS              OUT              ADM       
##  Min.   :   5.0   Min.   :  0.000   Min.   :     0   Min.   :    0  
##  1st Qu.:  66.0   1st Qu.:  0.000   1st Qu.:     0   1st Qu.: 2101  
##  Median : 129.0   Median :  0.000   Median : 15076   Median : 4604  
##  Mean   : 173.8   Mean   :  7.031   Mean   : 35867   Mean   : 6752  
##  3rd Qu.: 225.0   3rd Qu.:  0.000   3rd Qu.: 34675   3rd Qu.: 9696  
##  Max.   :1476.0   Max.   :180.000   Max.   :942251   Max.   :66439  
##       Rev            SALESY          SALES12           HIP95       
##  Min.   :    0   Min.   :  0.0   Min.   :  0.00   Min.   :  0.00  
```

```
##  1st Qu.: 1599   1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.: 13.00
##  Median : 3292   Median :  3.0   Median :  5.00   Median : 32.00
##  Mean   : 4433   Mean   : 22.6   Mean   : 36.33   Mean   : 57.73
##  3rd Qu.: 6094   3rd Qu.: 23.0   3rd Qu.: 35.00   3rd Qu.: 78.00
##  Max.   :45157   Max.   :438.0   Max.   :735.00   Max.   :606.00
##      KNEE95           Teach             TRAUMA            REHAB
##  Min.   :  0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  5.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 21.00   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   : 39.59   Mean   :0.2377   Mean   :0.1087   Mean   :0.2105
##  3rd Qu.: 53.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :375.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      HIP96           KNEE96           FEMUR96
##  Min.   :  0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 12.0   1st Qu.:  3.00   1st Qu.: 14.00
##  Median : 32.0   Median : 20.00   Median : 37.00
##  Mean   : 58.6   Mean   : 40.07   Mean   : 49.66
##  3rd Qu.: 78.0   3rd Qu.: 52.00   3rd Qu.: 72.00
##  Max.   :633.0   Max.   :388.00   Max.   :350.00
```

```
sum(is.na(west_coast))
```

```
## [1] 0
```

```
west_coast$Teach <- as.factor(west_coast$Teach)
west_coast$TRAUMA <- as.factor(west_coast$TRAUMA)
west_coast$REHAB <- as.factor(west_coast$REHAB)
```

EDA

The response variable, SALES12, is heavily skewed to the right. The best approach is to use a log transformation on SALES12 when comparing with the other predictor variables.

```
west_coast %>% ggplot(aes(SALES12)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

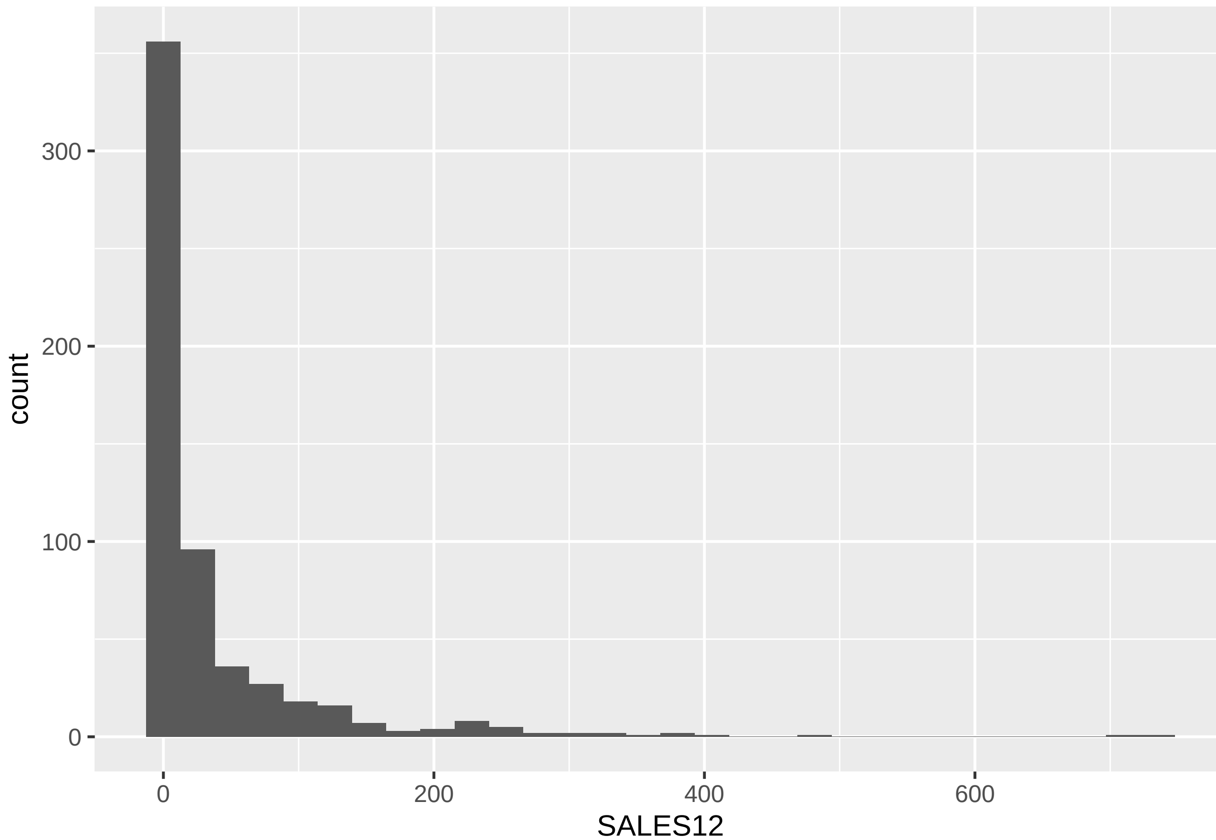


Figure 1: Histogram of equipment sales for the past 12 months

Even with a log transformation on SALES12, the normality assumption still seems to be violated in most if not all of the visuals.

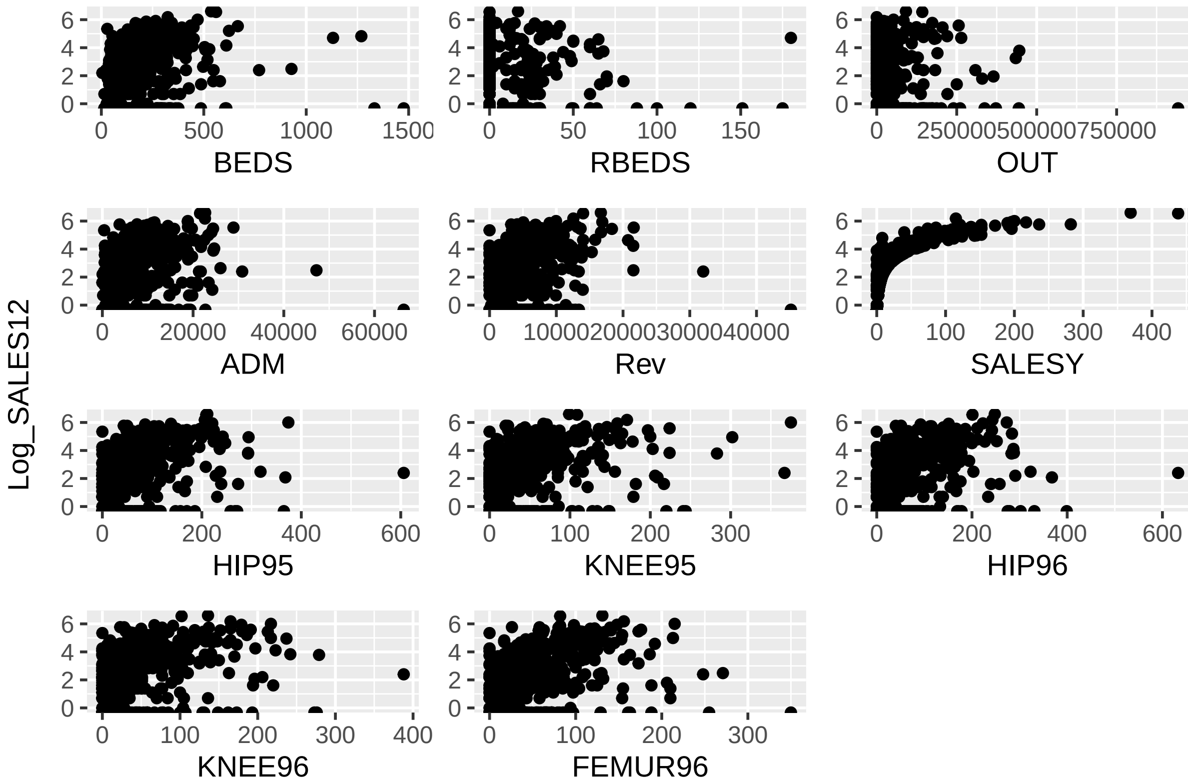


Figure 2: Scatter plots for continuous data

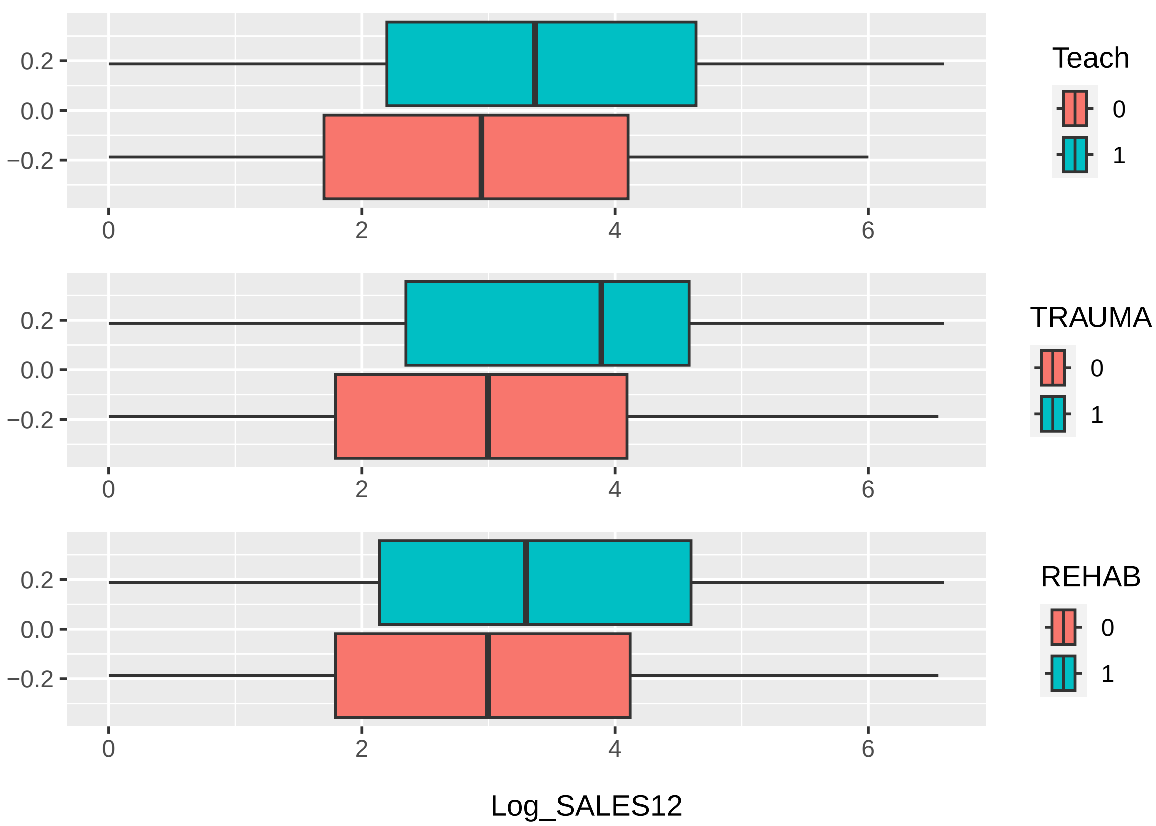


Figure 3: Box plot for categorical variables

Based on the correlation plot, we are also dealing with highly correlated variables. Whichever model we choose, some type of variable reduction method is needed.

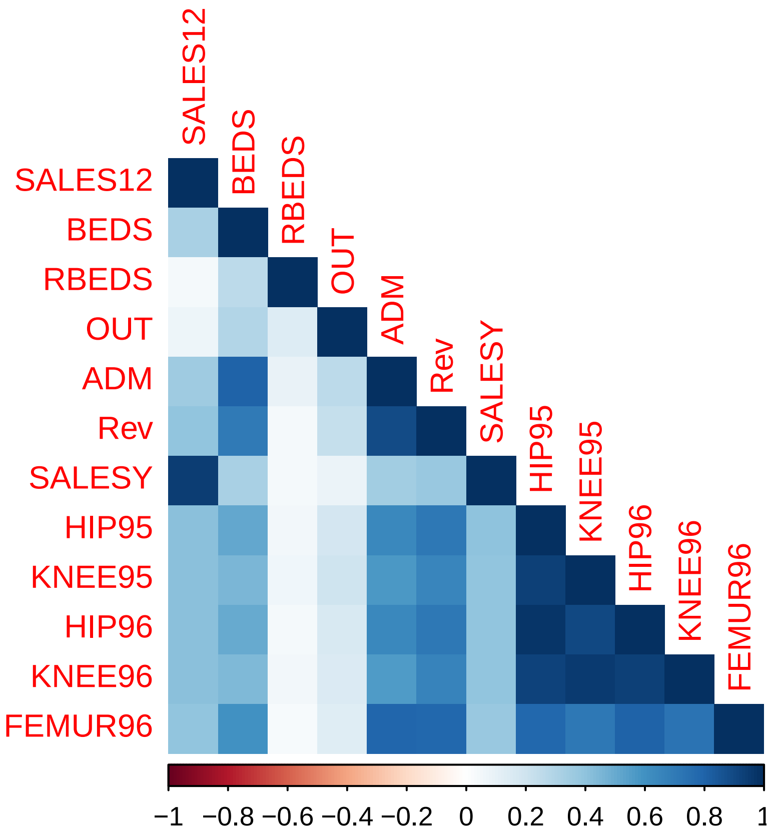


Figure 4: Correlation plot

Model Building

Based on the EDA visuals, it seems that fitting a normal regression model, even with transformations, is not the right approach. Since we want to identify hospitals that have high consumption and low sales, let's use SALES12 and ADM to create a new variable that classifies which hospitals meet this criterion and which ones do not. We can instead perform logistic regression based on the new outcome variable and check the model's assumptions to see if it's an appropriate method to use.

Creating new response variable

We will use the mean of ADM and SALES12 as the cutoff value to create the new variable. In other words, classify the hospitals with administrative costs greater than 6,752 dollars and sales of rehabilitation equipment less than 36 dollars.

```
summary(west_coast[, c(8, 11)])
```

```
##       ADM            SALES12      
##  Min.   :    0   Min.   :  0.00  
##  1st Qu.: 2101   1st Qu.:  0.00  
##  Median : 4604   Median :  5.00  
##  Mean   : 6752   Mean   : 36.33  
##  3rd Qu.: 9696   3rd Qu.: 35.00  
##  Max.   :66439   Max.   :735.00
```

```
w_c <- west_coast %>% mutate(y = ifelse(ADM > 6752 & SALES12 < 36,
                                        1, 0))



# Slight class imbalance
length(w_c$y[w_c$y == 0])/length(w_c$y)
```

```
## [1] 0.7928693
```

```
length(w_c$y[w_c$y == 1])/length(w_c$y)
```

```
## [1] 0.2071307
```

```
fin_data <- w_c[, -c(1:4, 8, 11 )]



fin_data$y <- as.factor(fin_data$y)
```

Fit Initial Model

We first use glm() to fit a logistic regression model on all the relevant variables. Not only are several variables insignificant, at least 3 variables have a vif (variance inflation factor) greater than 5 meaning high multicollinearity is present. This confirms what we observed in the correlation plot.

```r
first_mod = glm(y ~ ., family = binomial(),
                data = fin_data)

summary(first_mod)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(), data = fin_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3156  -0.3614  -0.1838  -0.0008   2.3528
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.403e+00  3.845e-01 -11.453  < 2e-16 ***
## BEDS         5.020e-03  1.759e-03   2.854  0.00432 **
## RBEDS       -1.499e-02  1.515e-02  -0.989  0.32255
## OUT          2.235e-06  2.449e-06   0.913  0.36136
## Rev          5.201e-04  1.048e-04   4.963 6.93e-07 ***
## SALESY      -1.359e-01  1.924e-02  -7.063 1.63e-12 ***
## HIP95        9.743e-03  1.311e-02   0.743  0.45732
## KNEE95      -2.413e-02  1.436e-02  -1.680  0.09294 .
## Teach1       1.021e+00  4.117e-01   2.480  0.01313 *
## TRAUMA1     -8.446e-02  6.230e-01  -0.136  0.89216
## REHAB1       6.718e-01  6.133e-01   1.096  0.27329
## HIP96        8.620e-03  1.149e-02   0.750  0.45301
## KNEE96      -1.417e-02  1.295e-02  -1.095  0.27357
## FEMUR96      1.918e-02  9.189e-03   2.087  0.03688 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 600.93  on 588  degrees of freedom
## Residual deviance: 268.54  on 575  degrees of freedom
## AIC: 296.54
##
## Number of Fisher Scoring iterations: 8
```

```r
vif(first_mod) # high multicollinearity
```

```
##       BEDS      RBEDS         OUT        Rev     SALESY      HIP95     KNEE95      Teach
```

```
##  1.999622  2.642330  1.266034  3.665446  1.941998 22.825950 18.594181  1.327866
##    TRAUMA     REHAB     HIP96    KNEE96   FEMUR96
##  1.237794  2.345265 20.428376 16.916064  4.820145
```

Reducing the number of variables

Based off of the initial model, let's use step() as a dimension reduction procedure to remove unnecessary variables. The AIC (Akaike information criterion) will assess the quality of all possible models and the model with the lowest AIC value will be considered the best model in this context.

Note: the results from running this code were too long to print so they were omitted.

```
step(first_mod, direction = "both", k = 2)
```

Reduced Model

The smallest AIC given was 288.2. The chosen variables were BEDS, Rev, SALESY, TEACH, KNEE96, and FEMUR96. We now see that all variables have become significant and they are no longer variables with high vif values. Unfortunately, there was one observation (507) that appeared to be an influential outlier so it was removed.

```
second_mod <- glm(y ~ BEDS + Rev + SALESY + Teach + KNEE96 +
                    FEMUR96, family = binomial(), data = fin_data)

summary(second_mod)
```

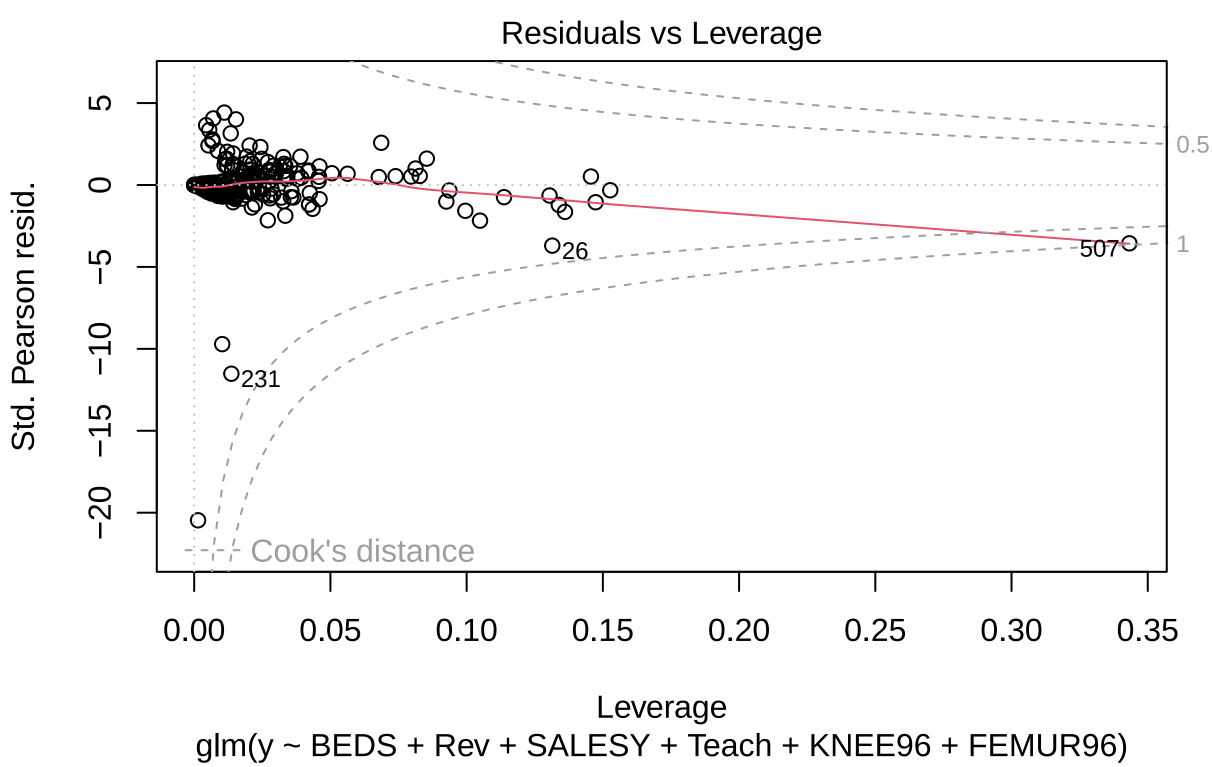
```
##
## Call:
## glm(formula = y ~ BEDS + Rev + SALESY + Teach + KNEE96 + FEMUR96,
##     family = binomial(), data = fin_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4751  -0.3602  -0.1892  -0.0011   2.4541
##
## Coefficients:
```

```
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -4.323e+00  3.727e-01 -11.601  < 2e-16 ***
## BEDS         4.833e-03  1.494e-03   3.235 0.001215 ** 
## Rev          5.305e-04  9.947e-05   5.333 9.67e-08 ***
## SALESY      -1.338e-01  1.821e-02  -7.347 2.02e-13 ***
## Teach1       1.007e+00  3.929e-01   2.563 0.010387 *  
## KNEE96      -1.883e-02  5.307e-03  -3.548 0.000388 ***
## FEMUR96      2.563e-02  7.388e-03   3.469 0.000523 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 600.93  on 588  degrees of freedom
## Residual deviance: 274.17  on 582  degrees of freedom
## AIC: 288.17
##
## Number of Fisher Scoring iterations: 8
```

```
vif(second_mod)
```

```
##     BEDS      Rev   SALESY    Teach   KNEE96  FEMUR96 
## 1.673554 3.421183 1.866998 1.246253 2.987671 3.278494
```

```
plot(second_mod,5)
```



```r
fin_data2 <- fin_data[-507,]
```

Final Model with removed observation

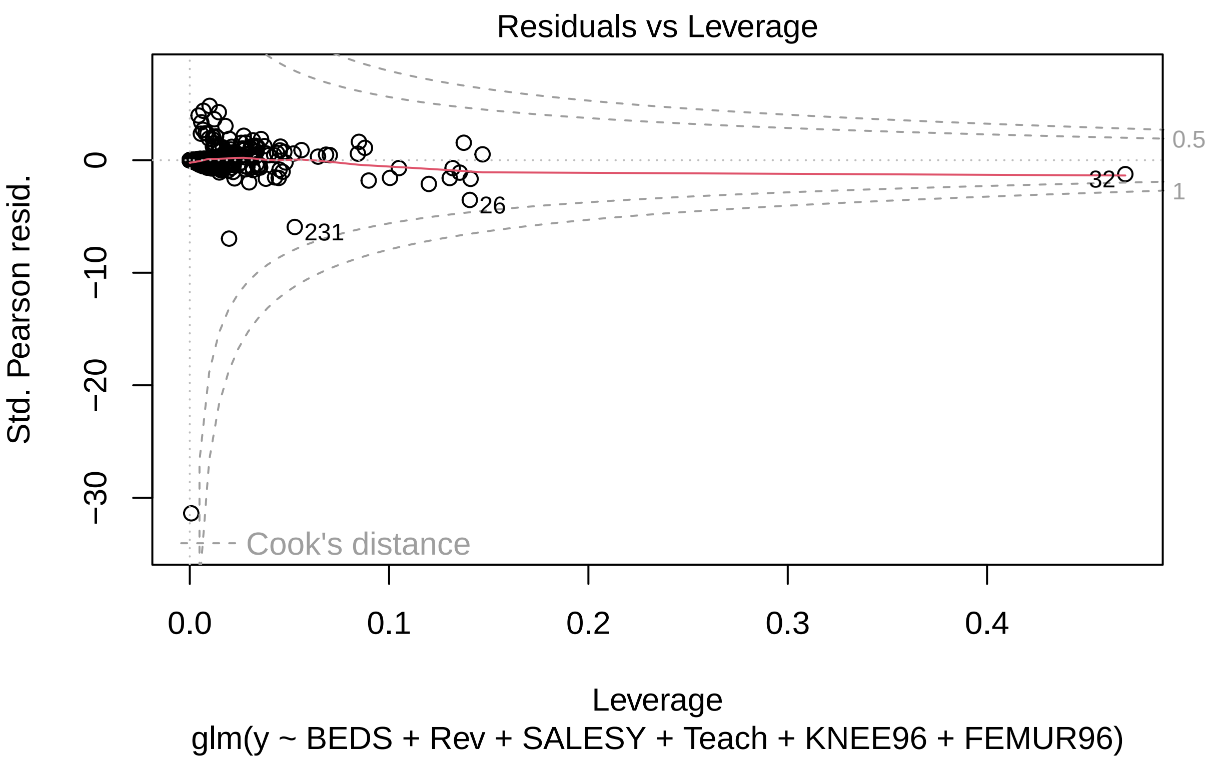
```r
fin_mod <-  glm(y ~ BEDS + Rev + SALESY + Teach + KNEE96 + FEMUR96,
                family = binomial(), data = fin_data2)

summary(fin_mod)
```

```
##
## Call:
## glm(formula = y ~ BEDS + Rev + SALESY + Teach + KNEE96 + FEMUR96,
##     family = binomial(), data = fin_data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7127  -0.3588  -0.1782  -0.0005   2.5217
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.520e+00  3.916e-01 -11.544  < 2e-16 ***
## BEDS         8.355e-03  1.780e-03   4.695 2.67e-06 ***
## Rev          4.778e-04  9.836e-05   4.857 1.19e-06 ***
## SALESY      -1.450e-01  1.947e-02  -7.446 9.62e-14 ***
## Teach1       7.011e-01  4.079e-01   1.719 0.085650 .
## KNEE96      -1.982e-02  5.385e-03  -3.681 0.000232 ***
## FEMUR96      2.561e-02  7.415e-03   3.454 0.000553 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 600.47  on 587  degrees of freedom
## Residual deviance: 265.62  on 581  degrees of freedom
## AIC: 279.62
##
## Number of Fisher Scoring iterations: 8
```

```r
vif(fin_mod)
```

```
##     BEDS      Rev   SALESY    Teach   KNEE96  FEMUR96
## 2.286207 3.420973 2.212432 1.291580 3.159329 3.244743
```

```r
plot(fin_mod,5)
```



Classification capabilites

To visually assess the classification capabilities of our chosen model, we will utilize the ROC Curve. Given that the specificity is 0.86, the sensitivity is 0.93, and the area under the curve is 0.95, the model's overall accuracy is pretty robust.

The threshold is set at 0.149. **In other words, if the fitted probability from the model is at least 0.149 we will classify the hospital as being a target hospital to sell equipment to.**

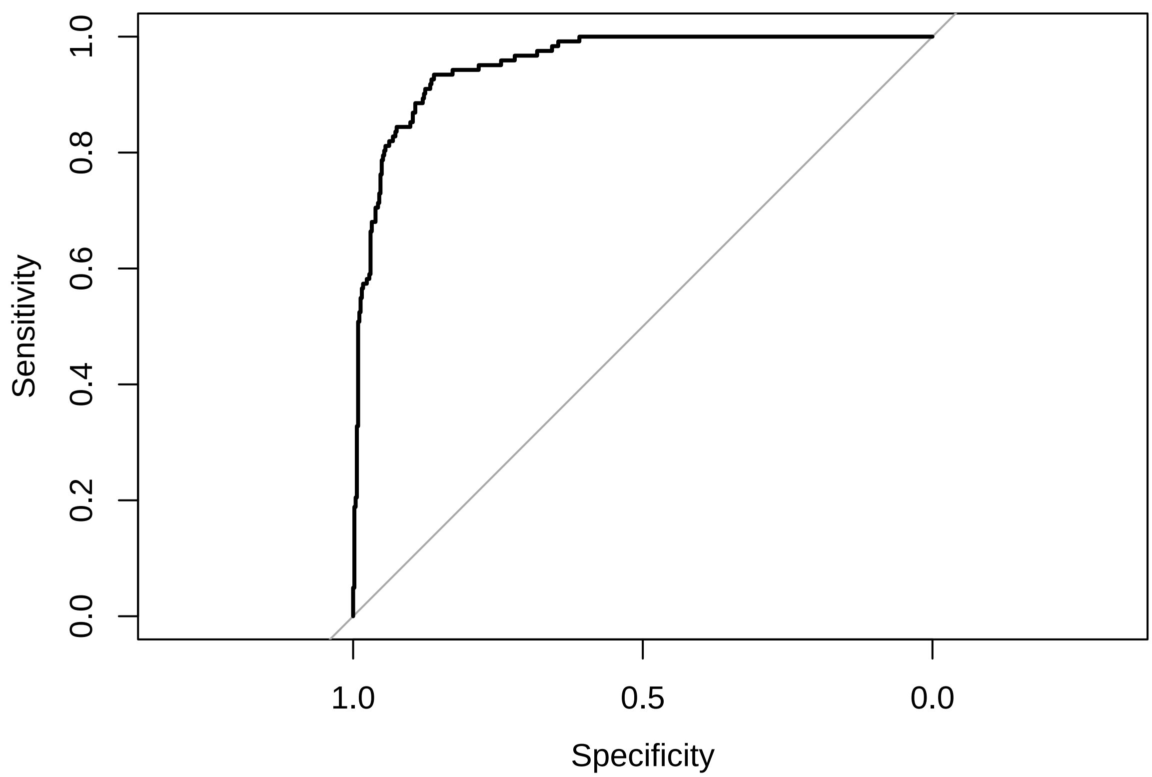
```r
final_pred = predict(fin_mod, type = "response")


roc_curve = roc(fin_data2$y, final_pred, auc = TRUE)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
print(roc_curve)
```

```
##
## Call:
## roc.default(response = fin_data2$y, predictor = final_pred, auc = TRUE)
##
## Data: final_pred in 466 controls (fin_data2$y 0) < 122 cases (fin_data2$y 1).
## Area under the curve: 0.9551
```

```
plot(roc_curve)
```




```
coords(roc_curve, "b", best.method = "youden")
```

```
##   threshold specificity sensitivity
## 1 0.1490317    0.860515   0.9344262
```