STAT 685 - Directed Studies

Ken Marciel

2/7/2022

Risk Factors for Postoperative Nausea and Vomiting (PONV)

The incidence of post-operative nausea and vomiting (PONV) is generally in the range of 20-40% (Apfel **et al**). There are several well-documented models for predicting PONV, to help guide prudent administration of prophylaxis. These models are typically developed using logistic regression and stepwise backward elimination for variable selection. The most common measure of accuracy for these models is the area under the receiver operating characteristic curve (ROC). These AUC values tend to fall in the range of 0.6 to 0.7 (Thomas **et al**, van den Bosch **et al**).

Packages

```
library(dplyr) # rename variables
library(alr4) # marginal model plots
library(leaps) # regression subset plots
library(car) # regression subset plots
library(rms) # logistic regression
library(pROC) # ROC curve
library(caret) # data splitting, resampling
```

Data set

```
# Obtain raw data set
path <- 'C:/Users/keoka/OneDrive - Texas A&M University/Courses/STAT_685/Data'
setwd(path) # set working directory to location of data file
ponv <- read.csv('PONV.csv') # read data from file

# Review data set
#class(ponv) # data frame
#dim(ponv) # 916 rows, 26 columns
#str(ponv) # display structure of data set
#names(ponv) # display names of variables in data set
# Rename selected variables
ponv <- rename(ponv,
               AgeOver50 = Ageover50,
               AnesthesiaDuration = Anaesthesiaduration,
               AnesthesiaOverHour = Anaesthesiaoverhour,
               Nonsmoker = Smoking,
               KinetosisHistory = Kinethosishist,
               PONVhistory = PONVhist,
               SerotoninBlocker = Serotoninblock)
#summary(ponv) # all columns are encoded as numeric values
```

```
#head(ponv) # display first 6 rows
#tail(ponv) # display last 6 rows

# Clean data set
ponv <- ponv[complete.cases(ponv),] # remove rows with missing values
#dim(ponv) # 916 rows, 26 columns (none of the records have missing values)
ponv <- unique(ponv) # remove duplicate records
#dim(ponv) # 823 rows, 26 columns (93 duplicate rows removed)
```

The raw data set has 916 rows and 26 columns. None of the rows have missing values. After removing the 93 duplicates, the cleaned data set has 823 rows and 26 columns.

Sample size

```
attach(ponv)
# Identify number of records of patients who took prophylaxis = 362
#nrow(ponv[Glukocorticoid==1 | Metoklopramid==1 | Serotoninblock==1,])
# Remove 362 records of patients who took prophylaxis
ponv <- ponv[Glukocorticoid==0 & Metoklopramid==0 & SerotoninBlocker==0,]
detach(ponv)

# Review data set.
#dim(ponv) # 461 rows, 26 columns (362 rows removed)
n <- nrow(ponv) # sample size = 461
```

After removing the 362 records of patients who took prophylaxis, the data set now has 461 rows and 26 columns.

Variables

```
names(ponv) # column names for the 26 variables in the original data set
```

```
##  [1] "Patient"             "AgeOver50"           "Age"
##  [4] "Gender"              "Diagnosis"           "Surgery"
##  [7] "AnesthesiaDuration" "AnesthesiaOverHour" "Weight"
## [10] "BMI"                 "Nonsmoker"           "KinetosisHistory"
## [13] "PONVhistory"         "Glukocorticoid"      "Metoklopramid"
## [16] "SerotoninBlocker"    "Cormackliheene"      "PONV0to2"
## [19] "PONV2to24"           "PONV0to24"           "Headache0to2"
## [22] "Headache2to24"       "Headache0to24"       "SinclairScore"
## [25] "ApfelScore"          "LelaScore"
```

Next, we will make sure that the original 26 variables are properly encoded for data analysis.

```
# Identification variable = 1
Patient <- as.character(ponv$Patient) # integer changed to character

# Response variables = 9
PONV0to2 <- as.factor(ponv$PONV0to2) # integer changed to factor
PONV2to24 <- as.factor(ponv$PONV2to24) # integer changed to factor
PONV0to24 <- as.factor(ponv$PONV0to24) # integer changed to factor
Headache0to2 <- as.factor(ponv$Headache0to2) # integer changed to factor
Headache2to24 <- as.factor(ponv$Headache2to24) # integer changed to factor
Headache0to24 <- as.factor(ponv$Headache0to24) # integer changed to factor
```

```
SinclairScore <- ponv$SinclairScore # numeric
ApfelScore <- ponv$ApfelScore # numeric
LelaScore <- ponv$LelaScore # integer

# Predictor variables = 16
AgeOver50 <- as.factor(ponv$AgeOver50) # integer changed to factor
Age <- ponv$Age # integer
Gender <- as.factor(ponv$Gender) # integer changed to factor
Diagnosis <- as.factor(ponv$Diagnosis) # integer changed to factor
Surgery <- as.factor(ponv$Surgery) # numeric changed to factor
AnaesthesiaDuration <- ponv$AnaesthesiaDuration # integer
AnaesthesiaOverHour <- as.factor(ponv$AnaesthesiaOverHour) # integer changed to factor
Weight <- ponv$Weight # numeric
BMI <- ponv$BMI # numeric
Nonsmoker <- as.factor(ponv$Nonsmoker) # integer changed to factor
KinetosisHistory <- as.factor(ponv$KinetosisHistory) # integer changed to factor
PONVhistory <- as.factor(ponv$PONVhistory) # integer changed to factor
Glukocorticoid <- as.factor(ponv$Glukocorticoid) # integer changed to factor
Metoklopramid <- as.factor(ponv$Metoklopramid) # integer changed to factor
SerotoninBlocker <- as.factor(ponv$SerotoninBlocker) # integer changed to factor
Cormackliheene <- as.factor(ponv$Cormackliheene) # numeric changed to factor
```

The 26 variables from the original data set consist of the ID variable, 9 response variables, and 16 predictor variables.

Variables considered for analysis of PONV incidence within 24 hours

Response variable selected:

Y = PONV0to24 (binary) = incidence of PONV within 24 hours of operation

For the purpose of developing a predictive model for PONV, we exclude the eight predictor variables corresponding to anesthetic and postoperative patient risk factors. In the full model, we consider the remaining eight predictor variables corresponding to the preoperative patient risk factors:

x_1 = Age (integer)
x_2 = Gender (binary)
$x_3 \dots x_{27}$ = Diagnosis (categorical with 26 levels)
$x_{28} \dots x_{34}$ = Surgery (categorical with 8 levels)
x_{35} = BMI (real)
x_{36} = Nonsmoker (binary)
x_{37} = Kinetosis history (binary)
x_{38} = PONV history (binary)

The model includes 25 dummy variables for the 26 levels of *Diagnosis*, and 7 dummy variables for the 8 levels of *Surgery*. This is a total of 38 variables, when the factors with more than two levels are taken into full account.

```
# Data set for regression analysis of PONV incidence within 24 hours
ponv <- data.frame(Patient, # ID variable
                   PONV0to24, # response variable
                   # 8 predictor variables
                   Age, Gender, Diagnosis, Surgery,
                   BMI, Nonsmoker, KinetosisHistory, PONVhistory)
#class(ponv) # data frame
#dim(ponv) # 461 rows, 10 columns
```

```
n <- nrow(ponv) # sample size = 461
p <- ncol(ponv) - 2 # predictor variables for full model = 8
# Observed incidence of PONV = 37.3%
PONVincidence = sum(as.numeric(ponv$PONV0to24)-1) /
  length(as.numeric(ponv$PONV0to24)-1)
```

The 10 variables in the data set for the full model consist of the ID variable, the response variable, and 8 predictor variables. The observed incidence of PONV for this data set is 0.37.

Full model for logistic regression

We begin by considering the following generalized linear model for the binary response variable:

$Y = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_{27} x_{27} + \beta_{28} x_{28} + ... + \beta_{34} x_{34} + \beta_{35} x_{35} + \beta_{36} x_{36} + \beta_{37} x_{37} + \beta_{38} x_{38} + e)$

where $e \sim$ iid $N(0, 1)$.

To model the binary response variable through a generalized linear model, we use the log odds ratio (logit) as the link function as follows:

$g^{-1}(Y) = \log(\frac{\theta(Y)}{1-\theta(Y)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_{27} x_{27} + \beta_{28} x_{28} + ... + \beta_{34} x_{34} + \beta_{35} x_{35} + \beta_{36} x_{36} + \beta_{37} x_{37} + \beta_{38} x_{38} + e$

θ is the parameter of the binomial distribution, which is related to a transformation of the logit as follows:

$\theta(Y) = \frac{\exp(Y)}{1+\exp(Y)} = \frac{1}{1+\exp(-Y)}$

For this analysis, our θ of interest is the proportion of patients diagnosed with PONV within the 24 hours following surgery.

The logistic regression model is fitted using the generalized linear method of least squares:

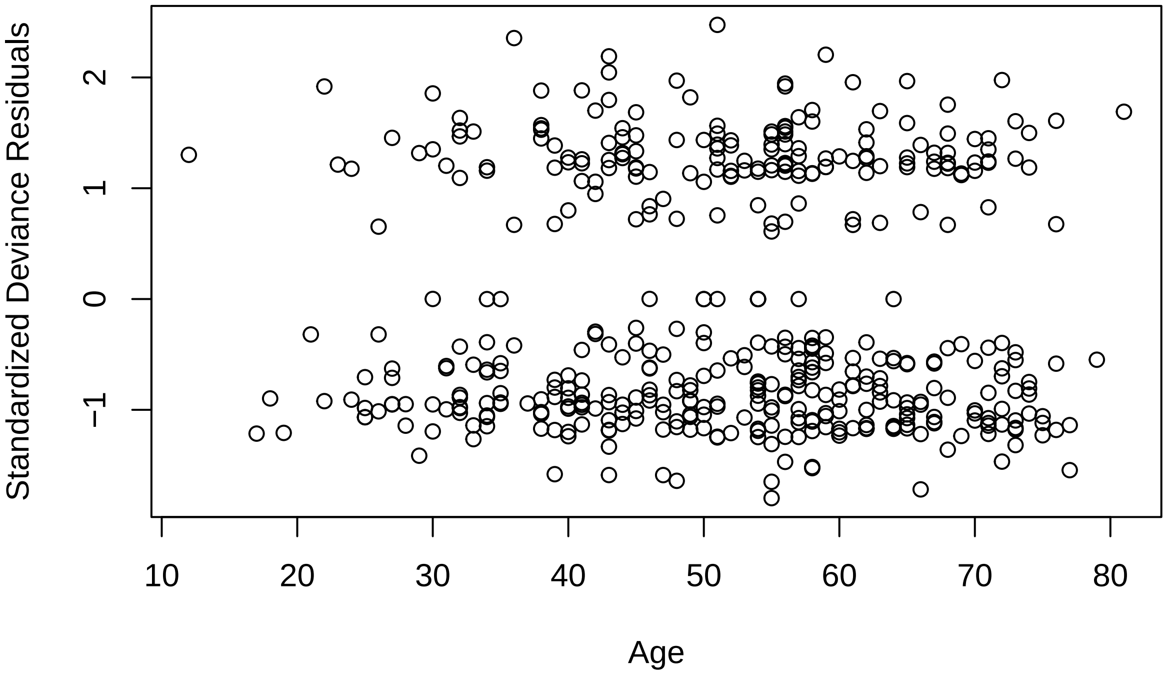
```
# Use glm (generalized linear model) function to fit logistic regression
glmFit1 <- glm(PONV0to24 ~ Age + Gender + Diagnosis + Surgery +
                 BMI + Nonsmoker + KinetosisHistory + PONVhistory,
               family = binomial, data = ponv)
summary(glmFit1)
```

```
##
## Call:
## glm(formula = PONV0to24 ~ Age + Gender + Diagnosis + Surgery +
##     BMI + Nonsmoker + KinetosisHistory + PONVhistory, family = binomial,
##     data = ponv)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7753  -0.9663  -0.5579   1.1567   2.3369
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.817e+01  1.171e+03  -0.016   0.9876
## Age                 -3.662e-03  8.834e-03  -0.415   0.6785
## Gender1              8.547e-01  3.217e-01   2.656   0.0079 **
## Diagnosis1           9.179e-01  6.314e-01   1.454   0.1460
## Diagnosis2           1.079e+00  5.780e-01   1.868   0.0618 .
## Diagnosis3           1.081e+00  7.968e-01   1.357   0.1748
## Diagnosis4           1.108e+00  6.124e-01   1.809   0.0705 .
## Diagnosis5          -1.851e-01  8.955e-01  -0.207   0.8362
## Diagnosis6         1.136e+00  9.282e-01   1.224   0.2211
## Diagnosis7         1.493e+00  1.674e+00   0.892   0.3726
## Diagnosis8        -7.698e-01  1.217e+00  -0.632   0.5272
## Diagnosis9        -6.555e-01  1.450e+00  -0.452   0.6513
## Diagnosis10        1.634e+00  2.022e+00   0.808   0.4191
## Diagnosis11       -1.439e+01  2.400e+03  -0.006   0.9952
## Diagnosis12       -1.392e+01  1.696e+03  -0.008   0.9935
## Diagnosis13       -3.879e-02  8.915e-01  -0.044   0.9653
## Diagnosis14        1.801e+01  1.223e+03   0.015   0.9883
## Diagnosis15        6.021e-01  8.127e-01   0.741   0.4587
## Diagnosis16        7.782e-01  9.200e-01   0.846   0.3976
## Diagnosis17       -1.397e+01  1.340e+03  -0.010   0.9917
## Diagnosis18        1.664e+01  2.400e+03   0.007   0.9945
## Diagnosis19        5.787e-01  9.616e-01   0.602   0.5473
## Diagnosis20        8.271e-01  1.384e+00   0.598   0.5500
## Diagnosis23        1.838e+01  2.400e+03   0.008   0.9939
## Diagnosis24       -1.458e+01  2.400e+03  -0.006   0.9952
## Diagnosis25        1.896e+01  2.400e+03   0.008   0.9937
## Surgery1           1.530e+01  1.171e+03   0.013   0.9896
## Surgery2           1.522e+01  1.171e+03   0.013   0.9896
## Surgery3           3.240e+01  2.670e+03   0.012   0.9903
## Surgery4           1.548e+01  1.171e+03   0.013   0.9894
## Surgery5          -2.295e+00  1.693e+03  -0.001   0.9989
## Surgery7           1.426e+01  1.171e+03   0.012   0.9903
## BMI                2.241e-02  2.415e-02   0.928   0.3536
## Nonsmoker1         5.353e-01  2.454e-01   2.181   0.0292 *
## KinetosisHistory1  8.598e-02  4.972e-01   0.173   0.8627
## PONVhistory1       1.432e+00  3.414e-01   4.194 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 609.06  on 460  degrees of freedom
## Residual deviance: 522.95  on 425  degrees of freedom
## AIC: 594.95
##
## Number of Fisher Scoring iterations: 15
```

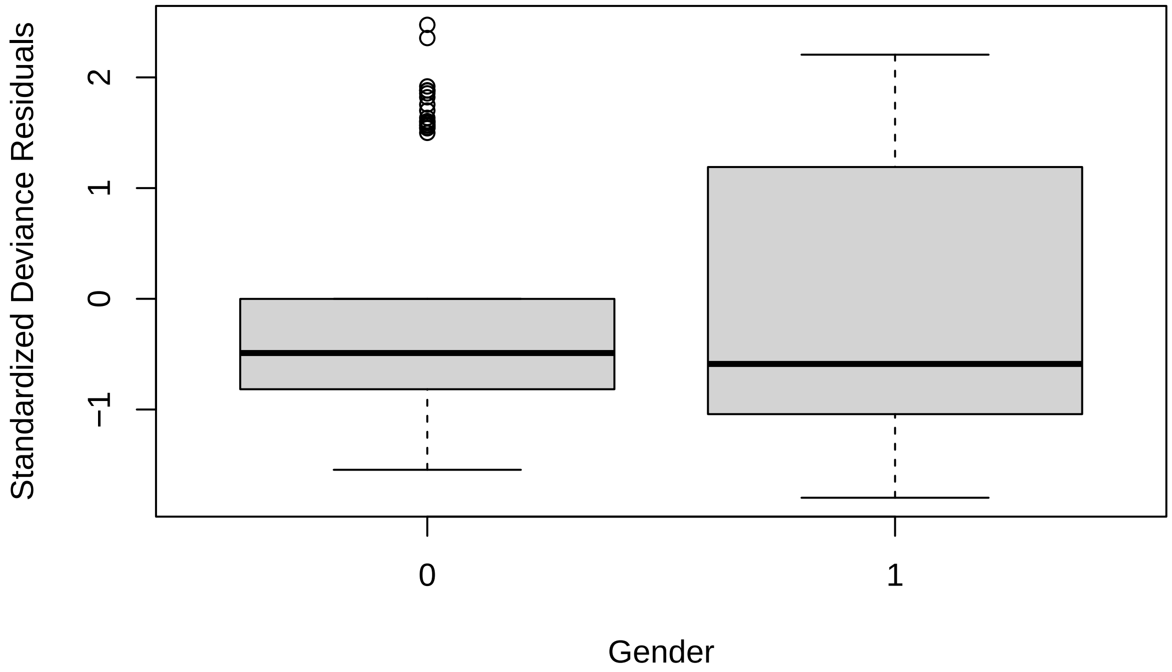
Three of the predictors in the full model have estimated coefficients that are statistically significant. In descending order of significance, these are PONV history, gender, and nonsmoker.

Plots of standardized deviance residuals

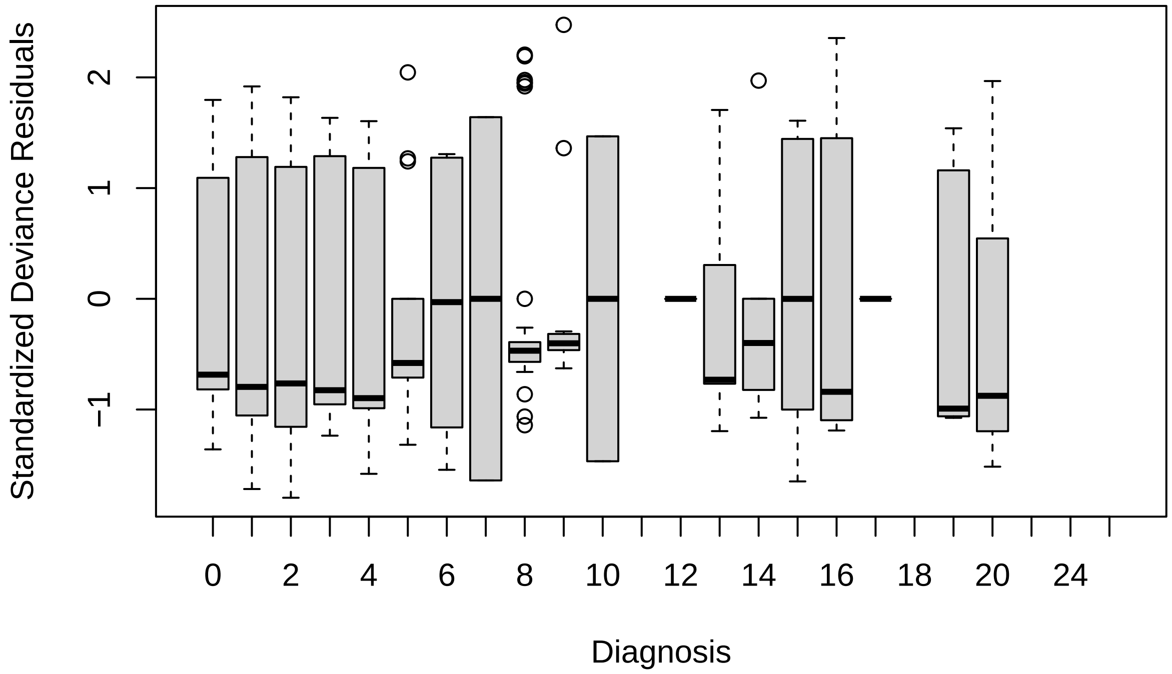
```
# Hat matrix
hval.glmFit1 <- influence(glmFit1)$hat

# Standardized deviance residuals
stanresDev.glmFit1 <- residuals(glmFit1)/sqrt(1-hval.glmFit1)

plot(Age, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Age")
```



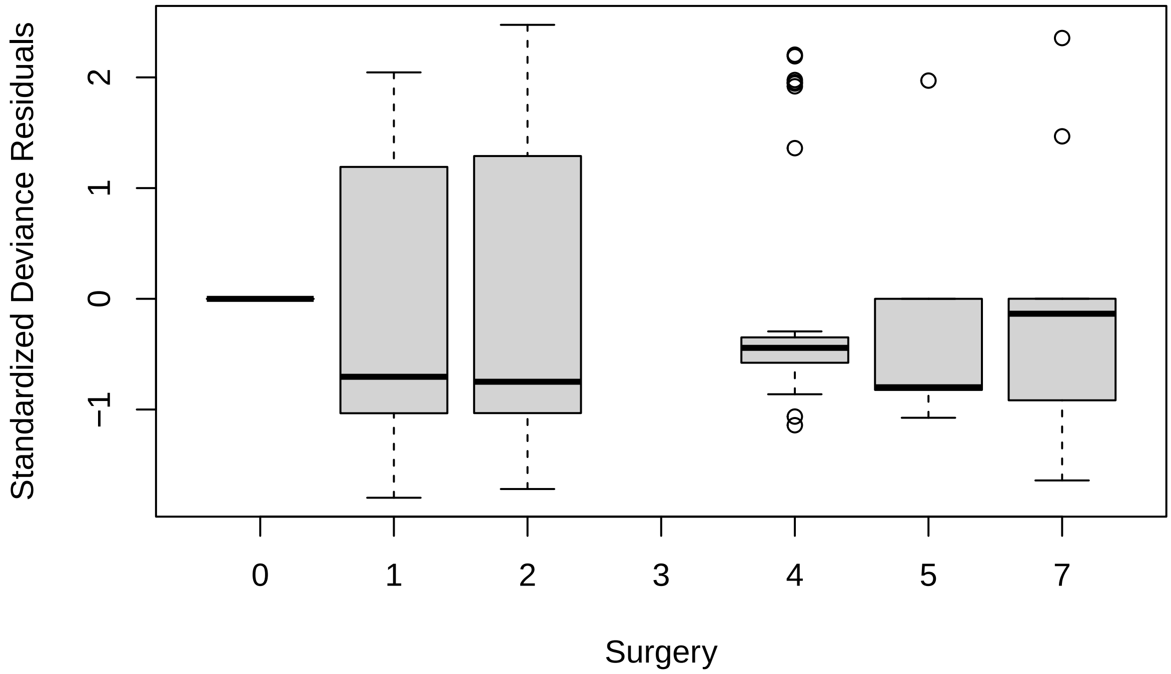
```
plot(Gender, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Gender")
```



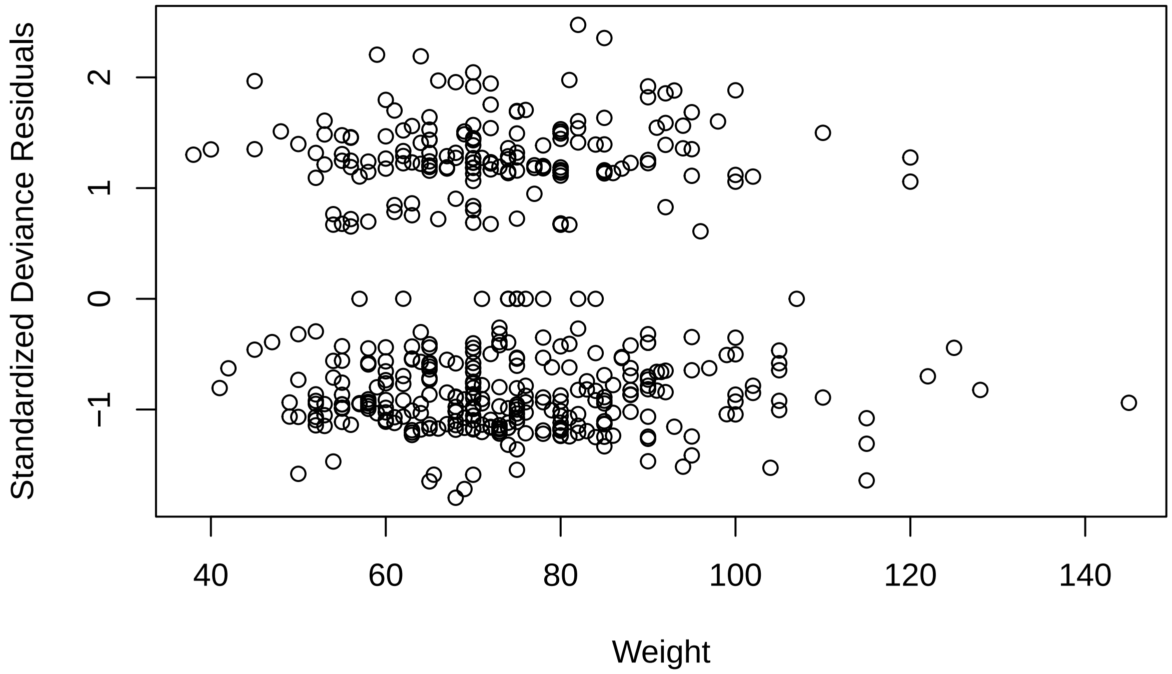
```
plot(Diagnosis, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Diagnosis")
```



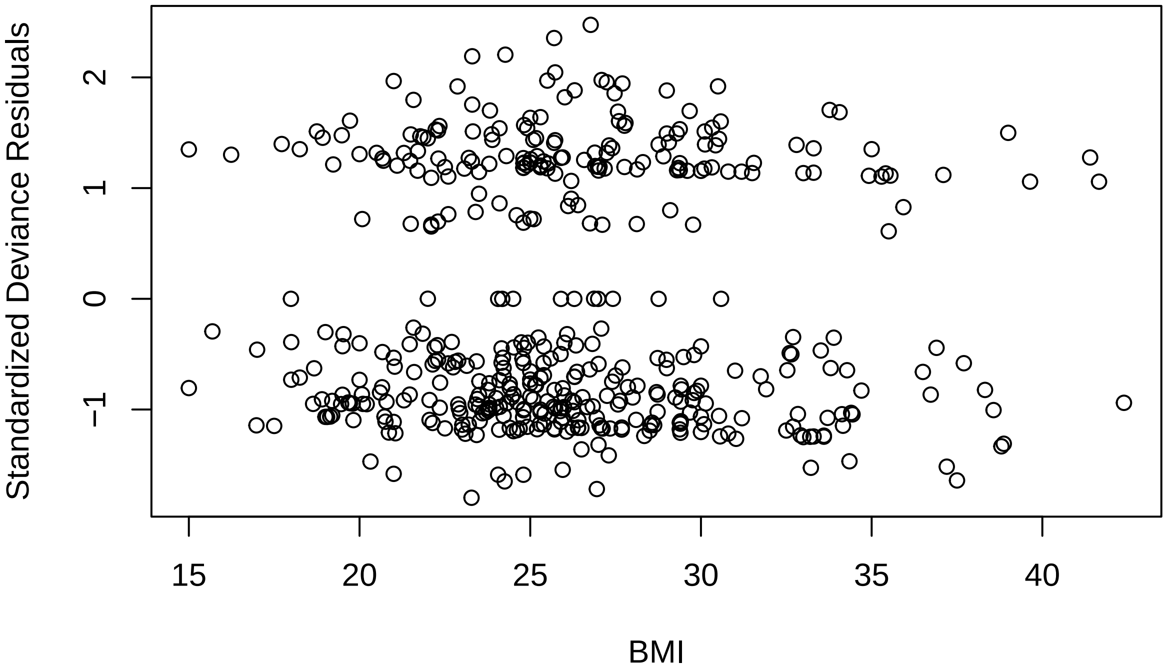
```
plot(Surgery, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Surgery")
```



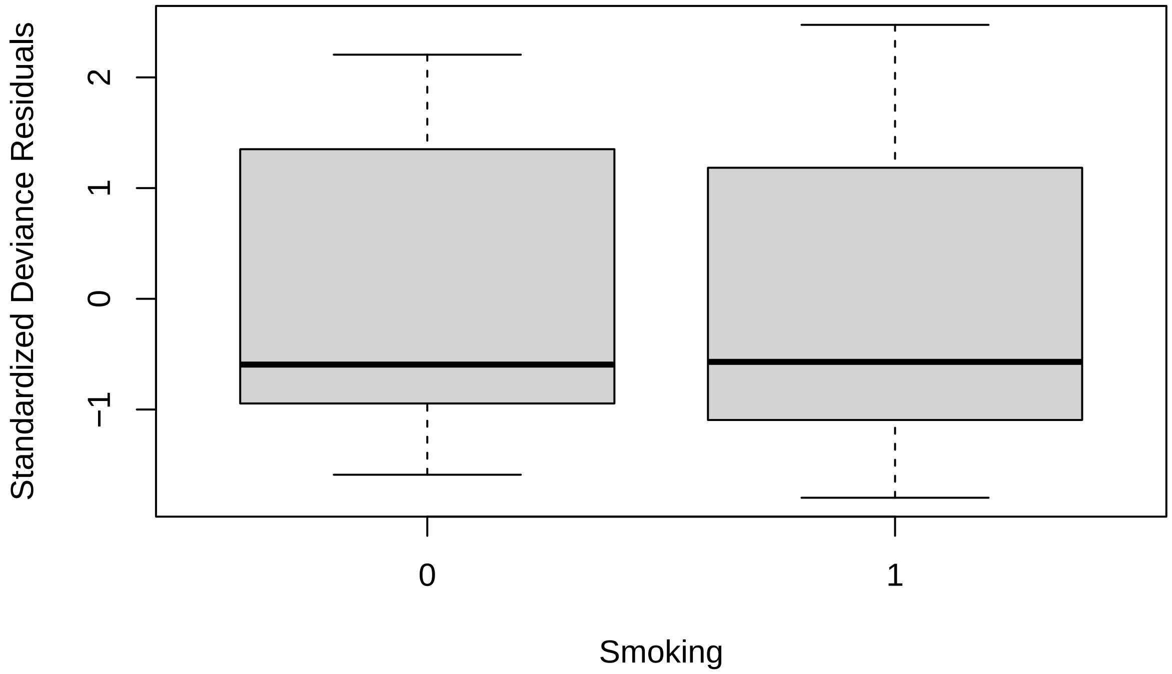
```
plot(Weight, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Weight")
```



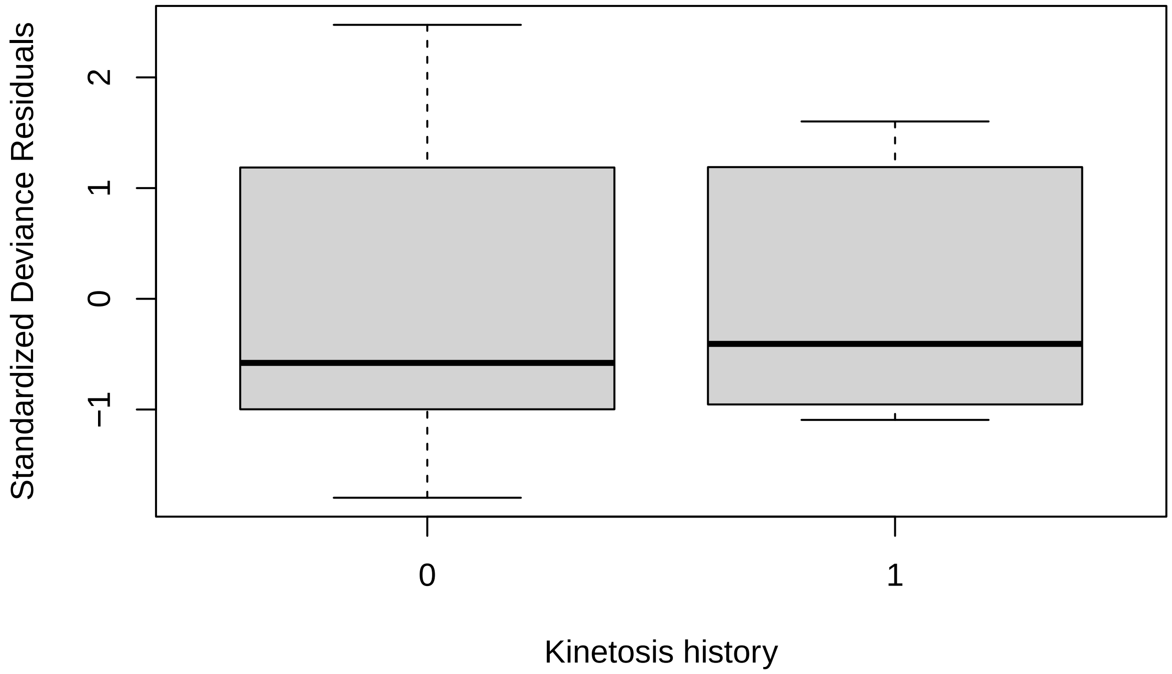
```
plot(BMI, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "BMI")
```



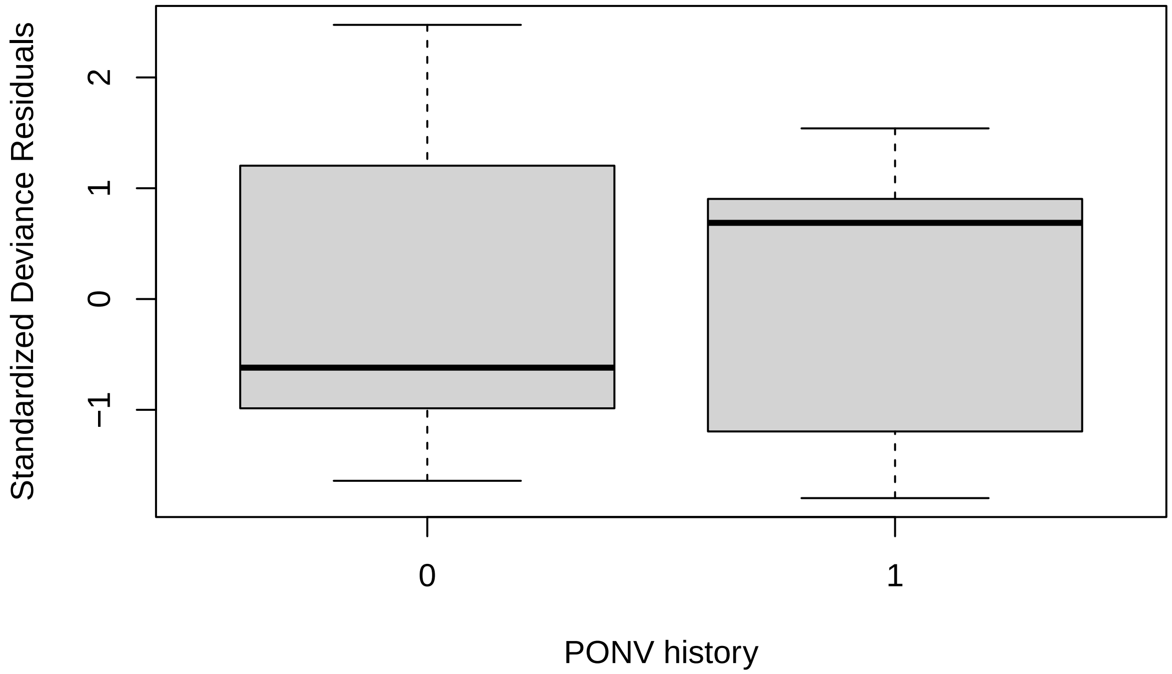
```
plot(Nonsmoker, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Smoking")
```



```
plot(KinetosisHistory, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "Kinetosis history")
```



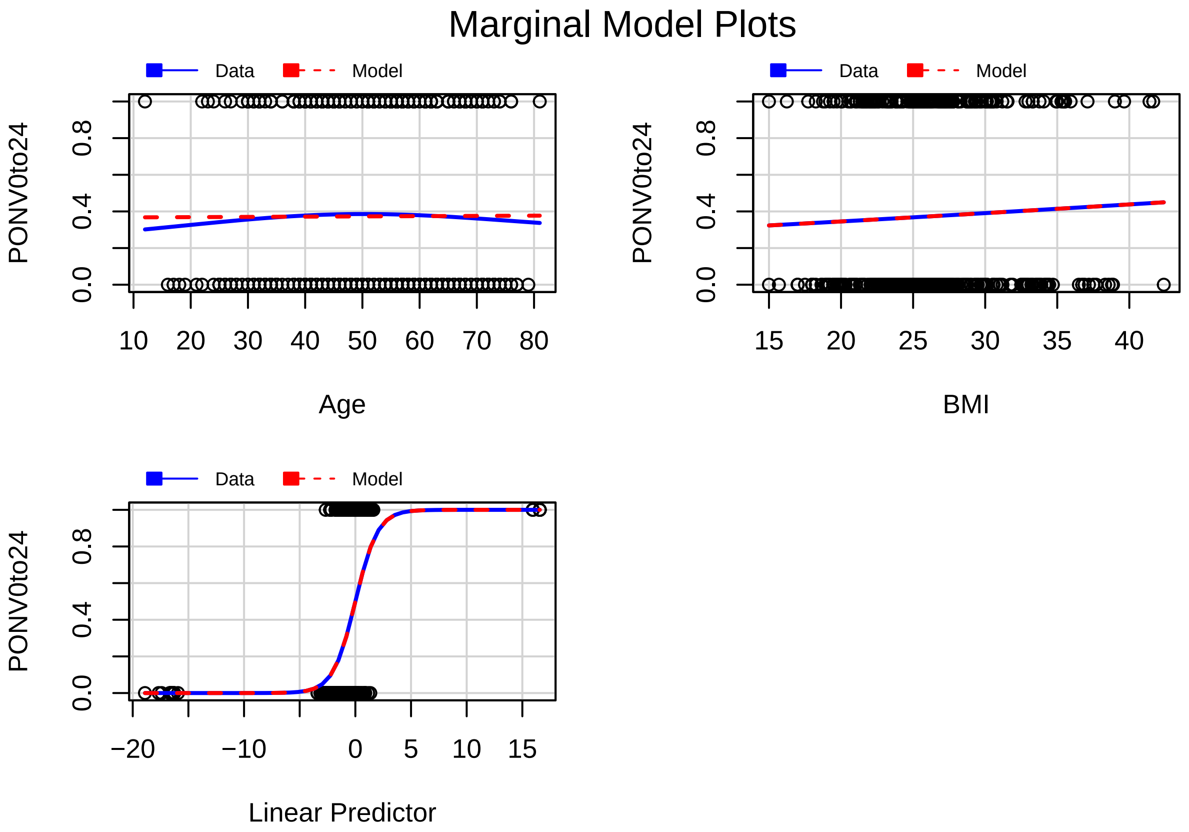
```
plot(PONVhistory, stanresDev.glmFit1,
     ylab = "Standardized Deviance Residuals", xlab = "PONV history")
```



Skewness is present in all of the predictors, most of which are right-skewed. This suggests that the log odds may depend on each skewed predictor through both a linear function and a log transformation. However, residual plots are difficult to interpret for binary data, so we will examine marginal model plots instead.

Marginal model plots for the continuous predictors

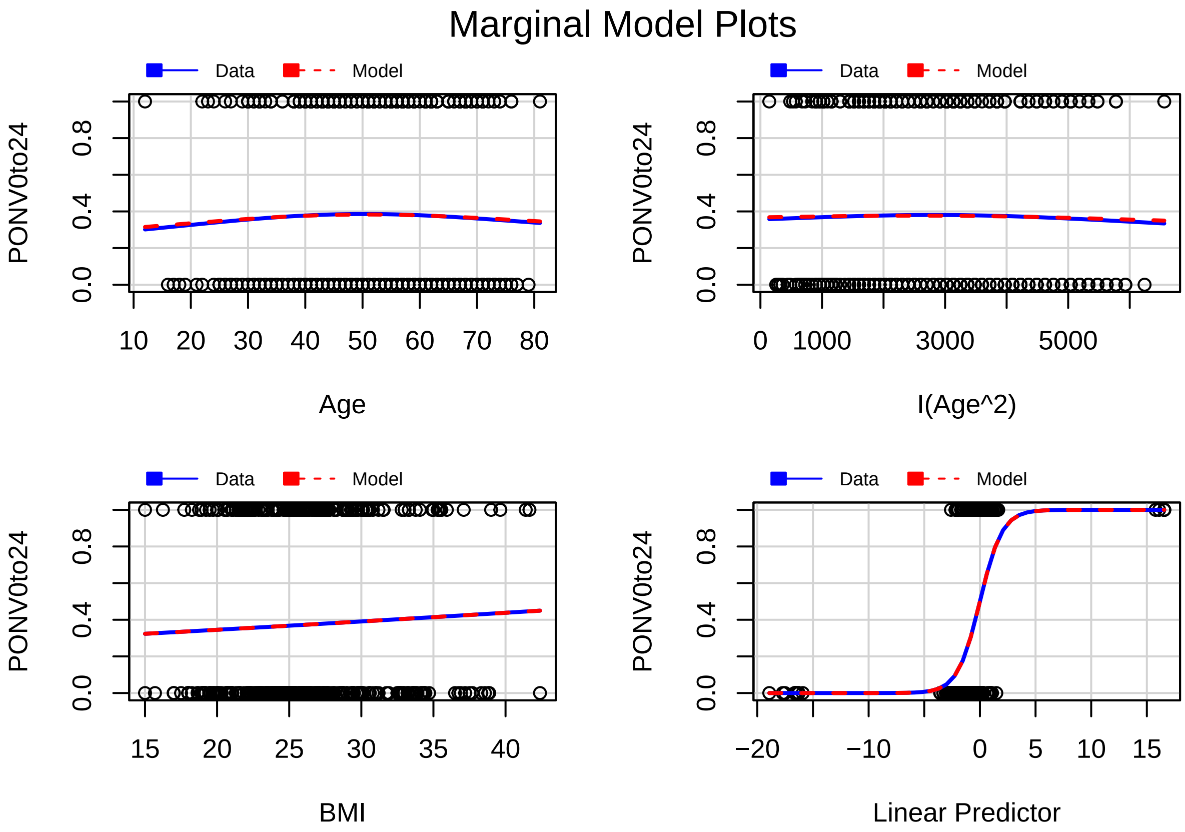
```
mmps(glmFit1) # marginal model plots displayed altogether in two-column format
```



```
#mmp(glmFit1,Age); mmp(glmFit1,BMI); mmp(glmFit1) # to view plots individually
```

There is reasonable agreement between the two fits in each of the marginal model plots for BMI and the linear predictor. Due to the lack of fit for Age, and the presence of parabolic curvature for the observed response, one possible approach is to consider adding a quadratic term for Age.

```
# Add a quadratic term for age to the model
glmFit2 <- glm(PONV0to24 ~ Age + I(Age^2) + Gender + Diagnosis + Surgery + BMI +
           Nonsmoker + KinetosisHistory + PONVhistory,
           family = binomial, data = ponv)
#summary(glmFit2)
mmps(glmFit2)
```

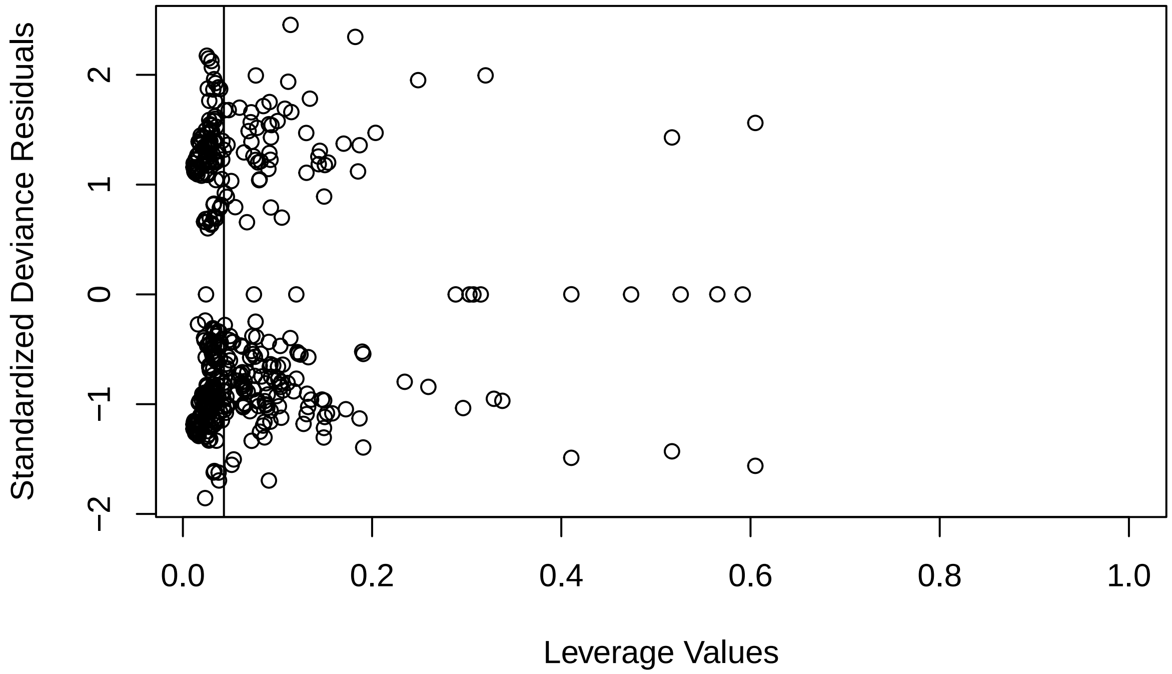


After adding a quadratic term for age, there is reasonable agreement between the two fits in each of the marginal model plots for Age, Age^2, BMI, and the linear predictor. This indicates that the current model is an adequate fit for the data.

Leverage values and standardized deviance residuals

As a final validity check, we examine leverage values and standardized deviance residuals.

```
hval.glmFit2 <- influence(glmFit2)$hat # hat matrix (leverage values)
p2 <- p + 1 # 10 predictors = 9 linear predictors plus quadratic term for Age
avgLev.glmFit2 <- (p2 + 1) / n  # average leverage
cutLev.glmFit2 <- 2 * avgLev.glmFit2  # cutoff for high leverage
stanresDev.glmFit2 <- residuals(glmFit2)/sqrt(1-hval.glmFit2) # standardized deviance residuals
plot(hval.glmFit2, stanresDev.glmFit2,
     ylab = "Standardized Deviance Residuals",
     xlab = "Leverage Values")
abline(v = cutLev.glmFit2)
identify(hval.glmFit2, stanresDev.glmFit2, labels = Patient)
```



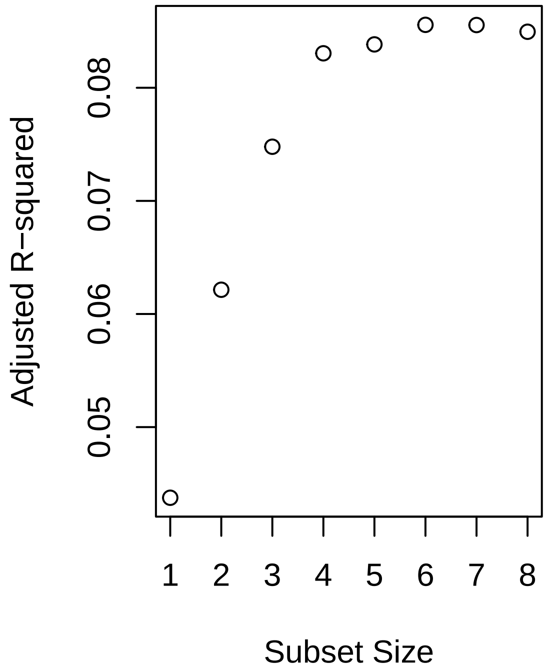
```
## integer(0)
```

A plot of leverage values and standardized deviance residuals reveals that none of the leverage points exceed 2.5 standard deviations. Six of the points exceed two standard deviations and should be investigated. However, since these points comprise only 1% of the 461 values in the data set, we will continue with the assumption that the current model is an adequate fit for the data. Therefore, we next proceed to variable selection.

Variable selection using all possible subsets

Plots of R^2_{adj} against subset size for the best subset of each size

```
Age2 <- Age^2
X <- cbind(Age, Age2, Gender, Diagnosis, Surgery, BMI, Nonsmoker,
           KinetosisHistory, PONVhistory)
b <- regsubsets(as.matrix(X), PONV0to24)
rs <- summary(b)
par(mfrow = c(1,2))
plot(1:8, rs$adjr2, xlab = "Subset Size", ylab = "Adjusted R-squared")
#subsets(b, statistic = c("adjr2"))
```



The plot of adjusted R^2 values shows the best predictor subsets to be as follows:

1 predictor: PONVhist
2 predictors: Surgery, PONVhist
3 predictors: Gender, Surgery, PONVhist
4 predictors: Gender, Surgery, Nonsmoker, PONVhist
5 predictors: Age^2, Gender, Surgery, Nonsmoker, PONVhist
6 predictors: Age^2, Gender, Surgery, BMI, Nonsmoker, PONVhist
7 predictors: Age, Age^2, Gender, Surgery, BMI, Nonsmoker, PONVhist
8 predictors: Age, Age^2, Gender, Surgery, BMI, Nonsmoker, KinetosisHist, PONVhist

The maximum value of R^2 corresponds to the predictor subset of size seven.

Viewing the results of the adjusted R^2 plot another way, the number of models that include each variable is as follows:

8 models: PONVhist
7 models: Surgery
6 models: Gender
5 models: Nonsmoker
4 models: Age^2
3 models: BMI
2 models: Age
1 model: KinetosisHist

Values of R^2_{adj}, AIC, AIC$_C$, and BIC for the best subset of each size

The predictor with the smallest p-value for its estimated coefficient is added to each subset to obtain the next subset.

```
# Subset size = 1
om1 <- glm(PONV0to24 ~ PONVhistory, family = binomial, data = ponv)
# Subset size = 2
om2 <- glm(PONV0to24 ~ PONVhistory + Gender, family = binomial, data = ponv)
# Subset size = 3
om3 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker, family = binomial,
          data = ponv)
# Subset size = 4
om4 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker + Diagnosis,
          family = binomial, data = ponv)
# Subset size = 5
om5 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker + Diagnosis + I(Age^2),
          family = binomial, data = ponv)
# Subset size = 6
om6 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker + Diagnosis + I(Age^2)
           + Age, family = binomial, data = ponv)
# Subset size = 7
om7 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker + Diagnosis + I(Age^2)
           + Age + BMI, family = binomial, data = ponv)
# Subset size = 8
om8 <- glm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker + Diagnosis + I(Age^2)
           + Age + BMI + KinetosisHistory, family = binomial, data = ponv)
# Subset size = 9
om9 <- glmFit2
```

Calculate AIC

```
extractAIC(om1, k = 2)
```

```
## [1]   2.000 592.694
```

```
extractAIC(om2, k = 2)
```

```
## [1]   3.0000 584.9411
```

```
extractAIC(om3, k = 2)
```

```
## [1]   4.0000 581.9543
```

```
extractAIC(om4, k = 2)
```

```
## [1]  27.0000 586.0426
```

```
extractAIC(om5, k = 2)
```

```
## [1]  28.0000 587.7882
```

```
extractAIC(om6, k = 2)
```

```
## [1]  29.0000 587.6944
```

```
extractAIC(om7, k = 2)
```

```
## [1]  30.0000 589.4248
```

```
extractAIC(om8, k = 2)
```

```
## [1]  31.0000 591.3129
```

```
extractAIC(om9, k = 2)
```

```
## [1]  37.0000 594.5913
```

The minimum value of AIC corresponds to the predictor subset of size three: PONVhistory, Gender, and Nonsmoker

Calculate AIC$_C$

```
npar.1 <- length(om1$coefficients) + 1
npar.2 <- length(om2$coefficients) + 1
npar.3 <- length(om3$coefficients) + 1
npar.4 <- length(om4$coefficients) + 1
npar.5 <- length(om5$coefficients) + 1
npar.6 <- length(om6$coefficients) + 1
npar.7 <- length(om7$coefficients) + 1
npar.8 <- length(om8$coefficients) + 1
npar.9 <- length(om9$coefficients) + 1

extractAIC(om1, k = 2) + 2 * npar.1 * (npar.1 + 1) / (n - npar.1 - 1)
```

```
## [1]   2.052516 592.746551
```

```
extractAIC(om2, k = 2) + 2 * npar.2 * (npar.2 + 1) / (n - npar.2 - 1)
```

```
## [1]   3.087719 585.028857
```

```
extractAIC(om3, k = 2) + 2 * npar.3 * (npar.3 + 1) / (n - npar.3 - 1)
```

```
## [1]   4.131868 582.086134
```

```
extractAIC(om4, k = 2) + 2 * npar.4 * (npar.4 + 1) / (n - npar.4 - 1)
```

```
## [1]  30.75926 589.80187
```

```
extractAIC(om5, k = 2) + 2 * npar.5 * (npar.5 + 1) / (n - npar.5 - 1)
```

```
## [1]  32.03712 591.82531
```

```
extractAIC(om6, k = 2) + 2 * npar.6 * (npar.6 + 1) / (n - npar.6 - 1)
```

```
## [1]  33.32558 592.01998
```

```
extractAIC(om7, k = 2) + 2 * npar.7 * (npar.7 + 1) / (n - npar.7 - 1)
```

```
## [1]  34.62471 594.04951
```

```
extractAIC(om8, k = 2) + 2 * npar.8 * (npar.8 + 1) / (n - npar.8 - 1)
```

```
## [1]  35.93458 596.24747
```

```
extractAIC(om9, k = 2) + 2 * npar.9 * (npar.9 + 1) / (n - npar.9 - 1)
```

```
## [1]  44.0237 601.6150
```

The minimum value of AIC$_C$ also corresponds to the predictor subset of size three: PONVhistory, Gender, and Nonsmoker.

Calculate BIC

```
extractAIC(om1, k = log(n))
```

```
## [1]   2.0000 600.9608
```

```
extractAIC(om2, k = log(n))
```

```
## [1]   3.0000 597.3413
```

```
extractAIC(om3, k = log(n))
```

```
## [1]   4.0000 598.4879
```

```
extractAIC(om4, k = log(n))
```

```
## [1]  27.0000 697.6444
```

```
extractAIC(om5, k = log(n))
```

```
## [1]  28.0000 703.5233
```

```
extractAIC(om6, k = log(n))
```

```
## [1]  29.0000 707.5629
```

```
extractAIC(om7, k = log(n))
```

```
## [1]  30.0000 713.4267
```

```
extractAIC(om8, k = log(n))
```

```
## [1]  31.0000 719.4482
```

```
extractAIC(om9, k = log(n))
```

```
## [1]  37.000 747.527
```

The minimum value of BIC corresponds to the predictor subset of size two: PONVhist and Gender.

Parsimonious model selection

AIC and AIC$_C$ each chose the predictor subset of size three. This subset consists of all three predictors having statistically significant coefficients in the full logistic regression model, both before and after adding the squared term for Age. Furthermore, the predictor subset of size three has a higher R^2 value than the predictor subset of size two. Therefore, we choose the predictor subset of size three as the parsimonious logistic regression model:

$Y = g(\beta_0 + \beta_1 PONVhist + \beta_2 Gender + \beta_3 Smoking + e)$

where $e \sim$ iid $N(0, 1)$.

As before, we use the logit function to model the binary response variable:

$g^{-1}(Y) = \log(\frac{\theta(Y)}{1-\theta(Y)}) = \beta_0 + \beta_1 PONVhist + \beta_2 Gender + \beta_3 Smoking + e$

where $\theta(Y) = \frac{\exp(Y)}{1+\exp(Y)} = \frac{1}{1+\exp(-Y)}$

```
glmFit3 <- glm(PONV0to24 ~ PONVhistory + Surgery + Gender,
               family = binomial, data = ponv)
summary(glmFit3)
```

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

```
## glm(formula = PONV0to24 ~ PONVhistory + Surgery + Gender, family = binomial,
##     data = ponv)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5782  -1.0297  -0.7133   1.3327   1.9836
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -16.323    727.699  -0.022  0.98210
## PONVhistory1    1.264      0.317   3.986 6.72e-05 ***
## Surgery1       15.208    727.699   0.021  0.98333
## Surgery2       15.084    727.699   0.021  0.98346
## Surgery3       31.132   1627.184   0.019  0.98474
## Surgery4       13.749    727.699   0.019  0.98493
## Surgery5       14.304    727.700   0.020  0.98432
## Surgery7       14.681    727.699   0.020  0.98390
## Gender1         0.757      0.287   2.638  0.00834 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 609.06  on 460  degrees of freedom
## Residual deviance: 557.95  on 452  degrees of freedom
## AIC: 575.95
##
## Number of Fisher Scoring iterations: 14
```

The lrm function is very similar to the glm function, with identical output for the regression coefficient estimates. The glm function also provides a summary of the deviance residuals. The lrm function also provides results of the model likelihood ratio test, as well as indices for discrimination and rank discrimination.

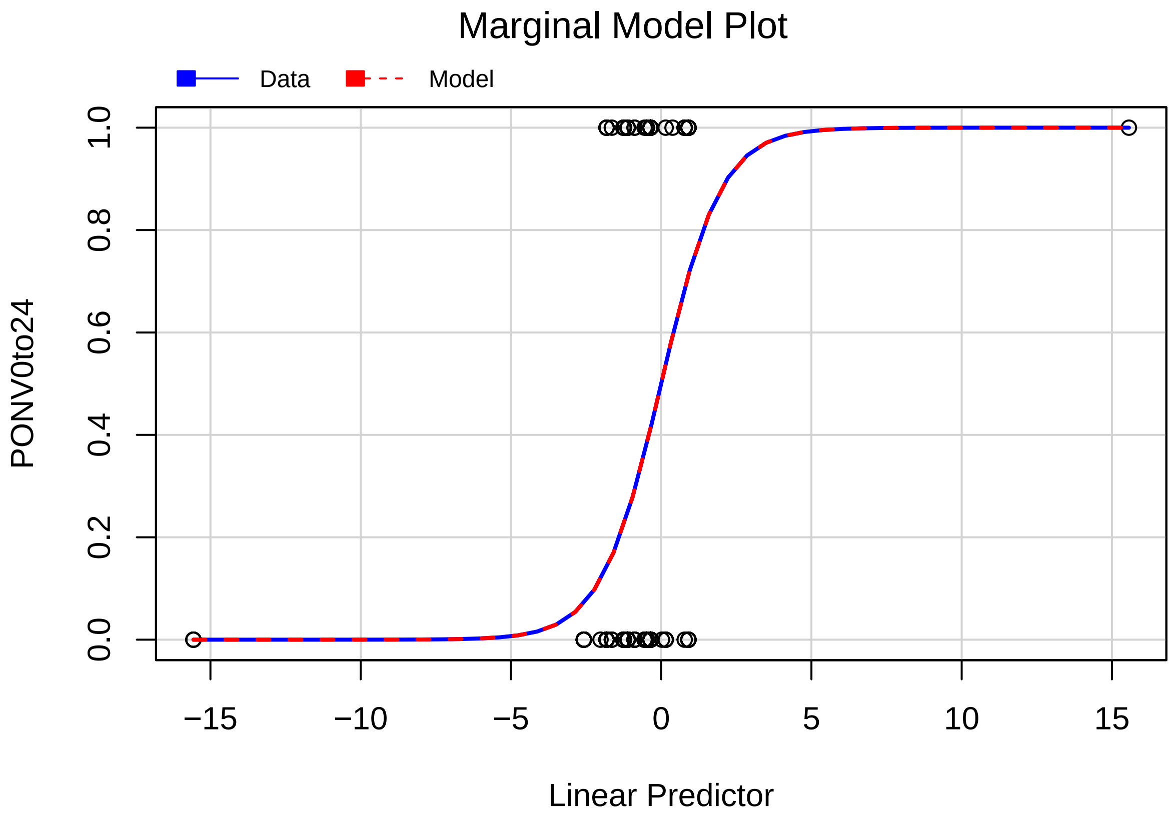
```
# lrm (Logistic Regression Model) function
rcsFit3 <- lrm(PONV0to24 ~ PONVhistory + Gender + Nonsmoker, data = ponv)
rcsFit3
```

```
## Logistic Regression Model
##
##  lrm(formula = PONV0to24 ~ PONVhistory + Gender + Nonsmoker, data = ponv)
##
##                         Model Likelihood    Discrimination    Rank Discrim.
##                               Ratio Test           Indexes          Indexes
##  Obs           461    LR chi2      35.11    R2       0.100    C       0.639
##   0            289    d.f.             3    g        0.618    Dxy     0.278
##   1            172    Pr(> chi2) <0.0001    gr       1.855    gamma   0.391
##  max |deriv| 9e-11                          gp       0.135    tau-a   0.130
##                                             Brier    0.216
##
##                 Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept     -1.6911 0.3027 -5.59  <0.0001
##  PONVhistory=1  1.3029 0.3132  4.16  <0.0001
##  Gender=1       0.8401 0.2803  3.00  0.0027
##  Nonsmoker=1    0.4766 0.2159  2.21  0.0272
##
```

The intercept and all three of the predictors in the chosen model have estimated coefficients that are statistically significant, consistent with the full model.

```
mmps(glmFit3) # marginal model plots for the chosen model
```

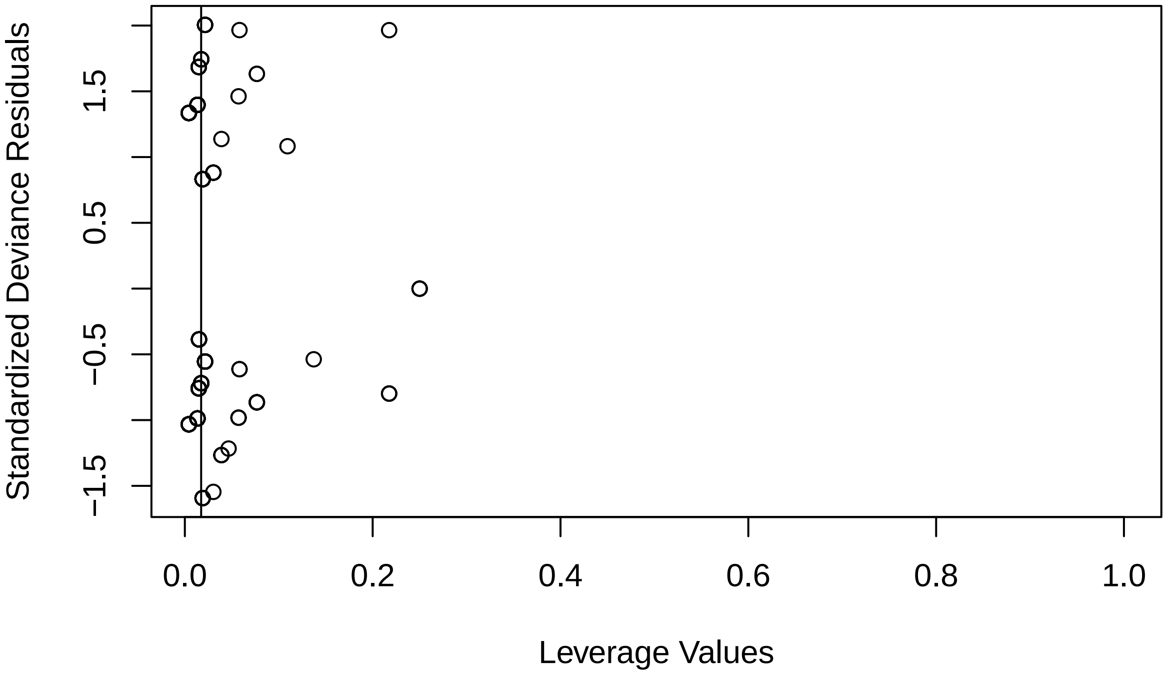
```
## Warning in mmps(glmFit3): Interactions and/or factors skipped
```



Since the chosen model has only factor predictor variables, a marginal model plot could only be obtained for the linear fit. There is reasonable agreement between the two fits in the marginal model plot for the linear fit. This indicates that the model is an adequate fit for the data.

We next look at leverage values versus standardized deviance residuals.

```
hval.glmFit3 <- influence(glmFit3)$hat # hat matrix (leverage values)
p3 <- 3 # number of predictors (PONVhist + Gender + Smoking)
avgLev.glmFit3 <- (p3 + 1) / n  # average leverage
cutLev.glmFit3 <- 2 * avgLev.glmFit3  # cutoff for high leverage
stanresDev.glmFit3 <- residuals(glmFit3)/sqrt(1-hval.glmFit3) # standardized deviance residuals
plot(hval.glmFit3, stanresDev.glmFit3,
     ylab = "Standardized Deviance Residuals",
     xlab = "Leverage Values")
abline(v = cutLev.glmFit3)
identify(hval.glmFit3, stanresDev.glmFit3, labels = Patient)
```



```
## integer(0)
```

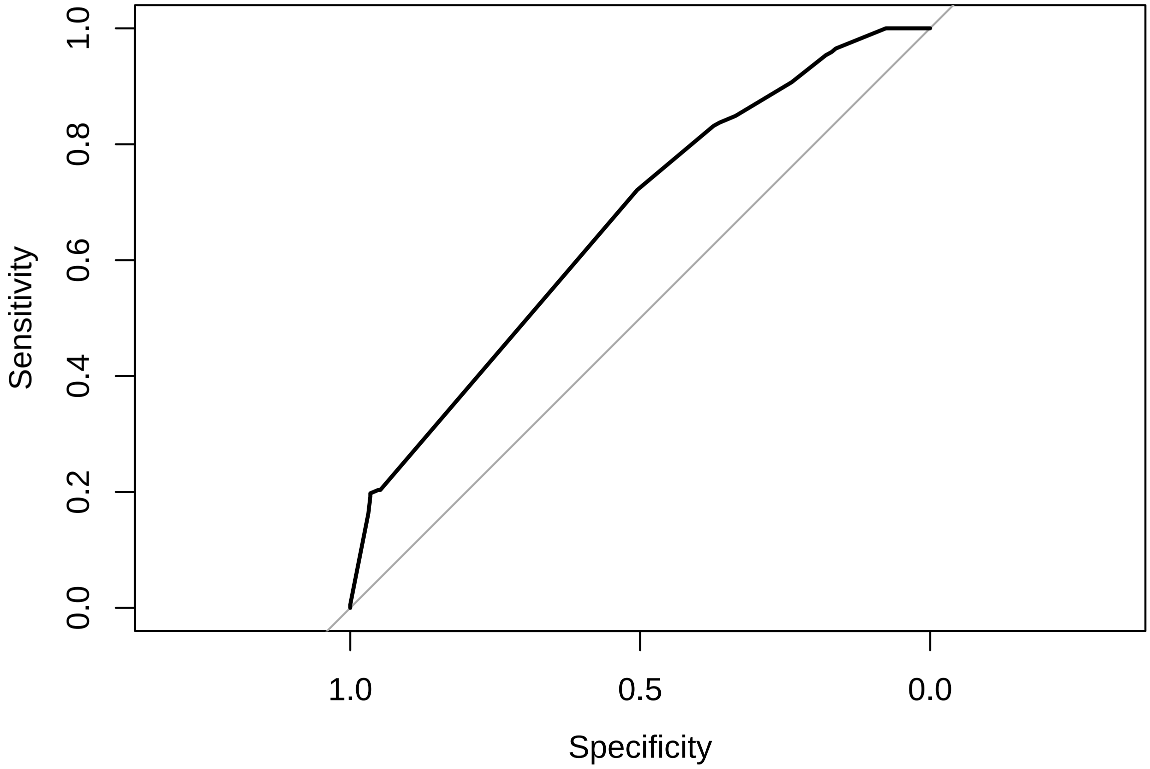
The plot of leverage values and standardized deviance residuals consists of points that are all within two standard deviations, which means there are no bad leverage points.

We next proceed to assess the predictive ability of this model using an ROC curve.

```
roc3 <- roc(PONV0to24, glmFit3$fitted.values, plot=TRUE); roc3
```

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

```
## Setting direction: controls < cases
```



```
##
## Call:
## roc.default(response = PONV0to24, predictor = glmFit3$fitted.values,     plot = TRUE)
##
## Data: glmFit3$fitted.values in 289 controls (PONV0to24 0) < 172 cases (PONV0to24 1).
## Area under the curve: 0.663
```

The area under the ROC curve (AUC) is 0.663. Since the ROC curve is a function of both sensitivity and specificity, the curve is insensitive to class imbalance.

Resampling techniques

k-fold cross-validation

We next perform logistic regression using five repeats of 10-fold cross-validation, to generate 50 different held-out sets for estimating model accuracy. With *k* chosen to be 10, each training set contains 90% of the entire data set, while each test set contains the other 10% of the data.

```
set.seed(1)
# Make syntactically valid names for the factor levels of the response variable
# Resampling specification is 5 repetitions of 10-fold cross-validation
logisticReg <- train(make.names(PONV0to24) ~ Age + I(Age^2) + Gender +
                       Diagnosis + Surgery + BMI +
                       Nonsmoker + KinetosisHistory + PONVhistory,
                     data = ponv,
                     method = "glm",
                     metric = "ROC",
                     trControl = trainControl(method = "repeatedcv",
                                              repeats = 5,
                                              classProbs = TRUE))
logisticReg # summary and results
```

```
## Generalized Linear Model
##
## 461 samples
##   8 predictor
##   2 classes: 'X0', 'X1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 416, 414, 415, 415, 414, 415, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.6415667  0.1722848
```

```
confusionMatrix(logisticReg) # confusion matrix
```

```
## Cross-Validated (10 fold, repeated 5 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction   X0   X1
##         X0 51.5 24.7
##         X1 11.1 12.6
##
##  Accuracy (average) : 0.6416
```

Repeated 10-fold cross-validation resulted in a logistic regression model with an accuracy of 64%, which has reasonable agreement with a baseline accuracy rate of 70%.

The choice of k to be 10 for k-fold cross-validation avoids the high bias of smaller values of k, as well as the computational burden of higher values of k. k-fold cross-validation generally has high variance compared to other methods. The potential issues with bias and variance become negligible for large training sets. Applying 10-fold cross-validation to our data set resulted in training sets each having a sample size between 414 and 416, which may be considered reasonably large. Furthermore, repeating the k-fold cross-validation procedure is known as an effective way to increase the precision of the estimates and still maintain a small bias.

```
summary(logisticReg)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8339  -0.9581  -0.5420   1.1239   2.3112
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.994e+01  1.168e+03  -0.017   0.9864
## Age                7.853e-02  5.507e-02   1.426   0.1539
## `I(Age^2)`        -8.170e-04  5.392e-04  -1.515   0.1298
## Gender1            8.152e-01  3.223e-01   2.529   0.0114 *  
## Diagnosis1         8.950e-01  6.322e-01   1.416   0.1569    
## Diagnosis2         1.077e+00  5.786e-01   1.862   0.0627 .  
## Diagnosis3         1.053e+00  7.991e-01   1.318   0.1875    
## Diagnosis4         1.189e+00  6.157e-01   1.932   0.0534 .  
## Diagnosis5        -1.127e-01  8.928e-01  -0.126   0.8996    
## Diagnosis6         1.204e+00  9.253e-01   1.301   0.1933    
## Diagnosis7         1.488e+00  1.688e+00   0.882   0.3780    
## Diagnosis8        -8.362e-01  1.218e+00  -0.687   0.4924    
## Diagnosis9        -8.259e-01  1.453e+00  -0.568   0.5697    
## Diagnosis10        2.031e+00  2.040e+00   0.996   0.3193    
## Diagnosis11       -1.348e+01  2.400e+03  -0.006   0.9955    
## Diagnosis12       -1.376e+01  1.696e+03  -0.008   0.9935    
## Diagnosis13       -4.460e-02  8.909e-01  -0.050   0.9601    
## Diagnosis14        1.822e+01  1.220e+03   0.015   0.9881    
## Diagnosis15        5.787e-01  8.139e-01   0.711   0.4771    
## Diagnosis16        9.297e-01  9.279e-01   1.002   0.3164    
## Diagnosis17       -1.398e+01  1.336e+03  -0.010   0.9916    
## Diagnosis18        1.680e+01  2.400e+03   0.007   0.9944    
## Diagnosis19        7.261e-01  9.649e-01   0.753   0.4517    
## Diagnosis20        7.744e-01  1.380e+00   0.561   0.5746    
## Diagnosis23        1.851e+01  2.400e+03   0.008   0.9938    
## Diagnosis24       -1.479e+01  2.400e+03  -0.006   0.9951    
## Diagnosis25        1.888e+01  2.400e+03   0.008   0.9937    
## Surgery1           1.535e+01  1.168e+03   0.013   0.9895    
## Surgery2           1.532e+01  1.168e+03   0.013   0.9895    
## Surgery3           3.246e+01  2.669e+03   0.012   0.9903    
## Surgery4           1.563e+01  1.168e+03   0.013   0.9893    
## Surgery5          -2.504e+00  1.689e+03  -0.001   0.9988    
## Surgery7           1.415e+01  1.168e+03   0.012   0.9903    
## BMI                1.406e-02  2.485e-02   0.566   0.5716    
## Nonsmoker1         5.910e-01  2.495e-01   2.368   0.0179 *  
## KinetosisHistory1  1.241e-02  4.976e-01   0.025   0.9801    
## PONVhistory1       1.409e+00  3.430e-01   4.106 4.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 609.06  on 460  degrees of freedom
## Residual deviance: 520.59  on 424  degrees of freedom
## AIC: 594.59
## 
## Number of Fisher Scoring iterations: 15
```

Three of the predictors in the full model have estimated coefficients that are statistically significant at the $\alpha = .05$ level or lower. In descending order of significance, these are PONV history, gender, and nonsmoker. These match the subset of predictors obtained from the model fitted using all possible subsets on the full data set.

The bootstrap

```
bootStrap <- train(make.names(PONV0to24) ~ Age + I(Age^2) + Gender +
                     Diagnosis + Surgery + BMI +
```

```r
                     Nonsmoker + KinetosisHistory + PONVhistory,
                   data = ponv,
                   method = "glm",
                   metric = "ROC",
                   trControl = trainControl(method = "boot",
                                            classProbs = TRUE))
bootStrap # summary and results
```

```
## Generalized Linear Model
##
## 461 samples
##   8 predictor
##   2 classes: 'X0', 'X1'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 461, 461, 461, 461, 461, 461, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.6293991  0.1629926
```

```r
confusionMatrix(bootStrap) # confusion matrix
```

```
## Bootstrapped (25 reps) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##           Reference
## Prediction   X0   X1
##         X0 49.2 23.7
##         X1 13.4 13.7
##
##  Accuracy (average) : 0.629
```

```r
summary(bootStrap) # coefficient estimates
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8339  -0.9581  -0.5420   1.1239   2.3112
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.994e+01  1.168e+03  -0.017   0.9864
## Age                7.853e-02  5.507e-02   1.426   0.1539
## `I(Age^2)`        -8.170e-04  5.392e-04  -1.515   0.1298
## Gender1            8.152e-01  3.223e-01   2.529   0.0114 *
## Diagnosis1         8.950e-01  6.322e-01   1.416   0.1569
## Diagnosis2         1.077e+00  5.786e-01   1.862   0.0627 .
## Diagnosis3         1.053e+00  7.991e-01   1.318   0.1875
## Diagnosis4         1.189e+00  6.157e-01   1.932   0.0534 .
## Diagnosis5        -1.127e-01  8.928e-01  -0.126   0.8996
## Diagnosis6         1.204e+00  9.253e-01   1.301   0.1933
## Diagnosis7         1.488e+00  1.688e+00   0.882   0.3780
## Diagnosis8        -8.362e-01  1.218e+00  -0.687   0.4924
## Diagnosis9        -8.259e-01  1.453e+00  -0.568   0.5697
## Diagnosis10        2.031e+00  2.040e+00   0.996   0.3193
## Diagnosis11       -1.348e+01  2.400e+03  -0.006   0.9955
## Diagnosis12       -1.376e+01  1.696e+03  -0.008   0.9935
## Diagnosis13       -4.460e-02  8.909e-01  -0.050   0.9601
## Diagnosis14        1.822e+01  1.220e+03   0.015   0.9881
## Diagnosis15        5.787e-01  8.139e-01   0.711   0.4771
## Diagnosis16        9.297e-01  9.279e-01   1.002   0.3164
## Diagnosis17       -1.398e+01  1.336e+03  -0.010   0.9916
## Diagnosis18        1.680e+01  2.400e+03   0.007   0.9944
## Diagnosis19        7.261e-01  9.649e-01   0.753   0.4517
## Diagnosis20        7.744e-01  1.380e+00   0.561   0.5746
## Diagnosis23        1.851e+01  2.400e+03   0.008   0.9938
## Diagnosis24       -1.479e+01  2.400e+03  -0.006   0.9951
## Diagnosis25        1.888e+01  2.400e+03   0.008   0.9937
## Surgery1           1.535e+01  1.168e+03   0.013   0.9895
## Surgery2           1.532e+01  1.168e+03   0.013   0.9895
## Surgery3           3.246e+01  2.669e+03   0.012   0.9903
## Surgery4           1.563e+01  1.168e+03   0.013   0.9893
## Surgery5          -2.504e+00  1.689e+03  -0.001   0.9988
## Surgery7           1.415e+01  1.168e+03   0.012   0.9903
## BMI                1.406e-02  2.485e-02   0.566   0.5716
## Nonsmoker1         5.910e-01  2.495e-01   2.368   0.0179 *
## KinetosisHistory1  1.241e-02  4.976e-01   0.025   0.9801
## PONVhistory1       1.409e+00  3.430e-01   4.106 4.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 609.06  on 460  degrees of freedom
## Residual deviance: 520.59  on 424  degrees of freedom
## AIC: 594.59
##
## Number of Fisher Scoring iterations: 15
```

Bootstrapping resulted in a logistic regression model with an accuracy of 63%, which is lower than that obtained from 10-fold cross-validation using 5 repeats.

References

Apfel, C. C., Kranke, P., Eberhart, L. H. J., Roos, A., and Roewer, N. (2002), “Comparison of Predictive Models for Postoperative Nausea and Vomiting,” *British Journal of Anaesthesia*, 88 (2), 234-40.

Eberhart, L. H. J., Hogel, J., Seeling, W., Staack, A.M., Geldner, G., and Georgieff, M. (2000), “Evaluation of Three Risk Scores to Predict Postoperative Nausea and Vomiting,” *Acta Anaesthesiologica Scandinavica*, 44, 480–488.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning* (2nd ed.), New York, NY: Springer Science+Business Media, LLC.

Kuhn, M., and Johnson, K. (2013), *Applied Predictive Modeling*, New York , NY: Springer Science+Business

Media, LLC.

Pampel, F. C. (2021), *Logistic Regression* (2nd ed.), Thousand Oaks, CA: SAGE Publications, Inc.

Sheather, S. J. (2009), *A Modern Approach to Regression with R*, New York, NY: Springer Science+Business Media, LLC.

Sinclair, D. R., Chung, F., and Mezei, G. (1999), "Can Postoperative Nausea and Vomiting Be Predicted?" *Anesthesiology*, 91, 109-118.

Thomas, R., Jones, N. A., and Strike, P. (2002), "The Value of Risk Scores for Predicting Postoperative Nausea and Vomiting when Used to Compare Patient Groups in a Randomised Controlled Trial," *Anaesthesia*, 57, 1119-1128.

van den Bosch, J.E., Kalkman, C. J., Vergouwe, Y., Van Klei, W. A., Bonsel, G. J., Grobbee, D. E., and Moons, K. G. M. (2005), "Assessing the Applicability of Scoring Systems for Predicting Postoperative Nausea and Vomiting," *Anaesthesia*, 60, 323-331.

Vidakovic, B. (2017), *Engineering Biostatistics*, Hoboken, NJ: John Wiley & Sons Ltd.