Predicting the incidence of postoperative nausea and vomiting

Ken Marciel, Project in Statistics (STAT 685), Texas A&M University
Spring 2022

The incidence of post-operative nausea and vomiting (PONV) is generally in the range of 20-40% (Apfel *et al*). This condition has negative effects on the health and well-being of patients, and is financially costly to healthcare providers. The tradeoff is that preventive therapy has negative side effects and financial costs. So, the challenge is to develop a scoring system that most accurately recommends prophylaxis for the patients at high risk of PONV. In other words, a predictive model that is neither to conservative nor too liberal in determining which patients should be prescribed prophylaxis. This balancing act has been described in the medical literature as the *prevent-or-cure dilemma*.

There are several well-documented models for predicting PONV, to help guide prudent administration of anti-emetic prophylaxis. These models are typically developed using logistic regression and stepwise backward elimination for variable selection. The most common measures of validity are discrimination and calibration. The most common measure of discriminating power is AUC, the area under the receiver operating characteristic curve (ROC). Calibration is most commonly assessed using the slope and squared correlation (R^2) for the line in a calibration plot. In the literature, AUC values range from 0.61 to 0.785 (Apfel *et al*, Sinclair *et al*), calibration slopes range from 0.3 to 1.71 (Apfel *et al*, Eberhart *et al*), and squared correlation ranges from 0.763 to 0.99 (Apfel *et al*).

This investigation analyzed a data set of 461 patients from anesthesiologist Jelena Velickovic, MD, in Belgrade, Serbia. The purpose was to develop a predictive model for PONV with performance comparable to or better than models previously published in the literature.

Methods

Software

Data analysis was performed using the R statistical computing software through the R Markdown interface, with the following additional packages installed to R.

```
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

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

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

Variables in data set

Next, the original 26 variables were properly encoded for data analysis. These 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

To develop a predictive model for PONV, I excluded the eight predictor variables corresponding to anesthetic and postoperative patient risk factors. Therefore, in the full model, I considered the remaining eight predictor variables corresponding to the preoperative patient risk factors:

x_1 = Age (integer)
x_2 = Gender (binary)
x_3 … x_{27} = Diagnosis (categorical with 26 levels)
x_{28} … 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.

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

I began by considering the following generalized linear model for the binary response variable:

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

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

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

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

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

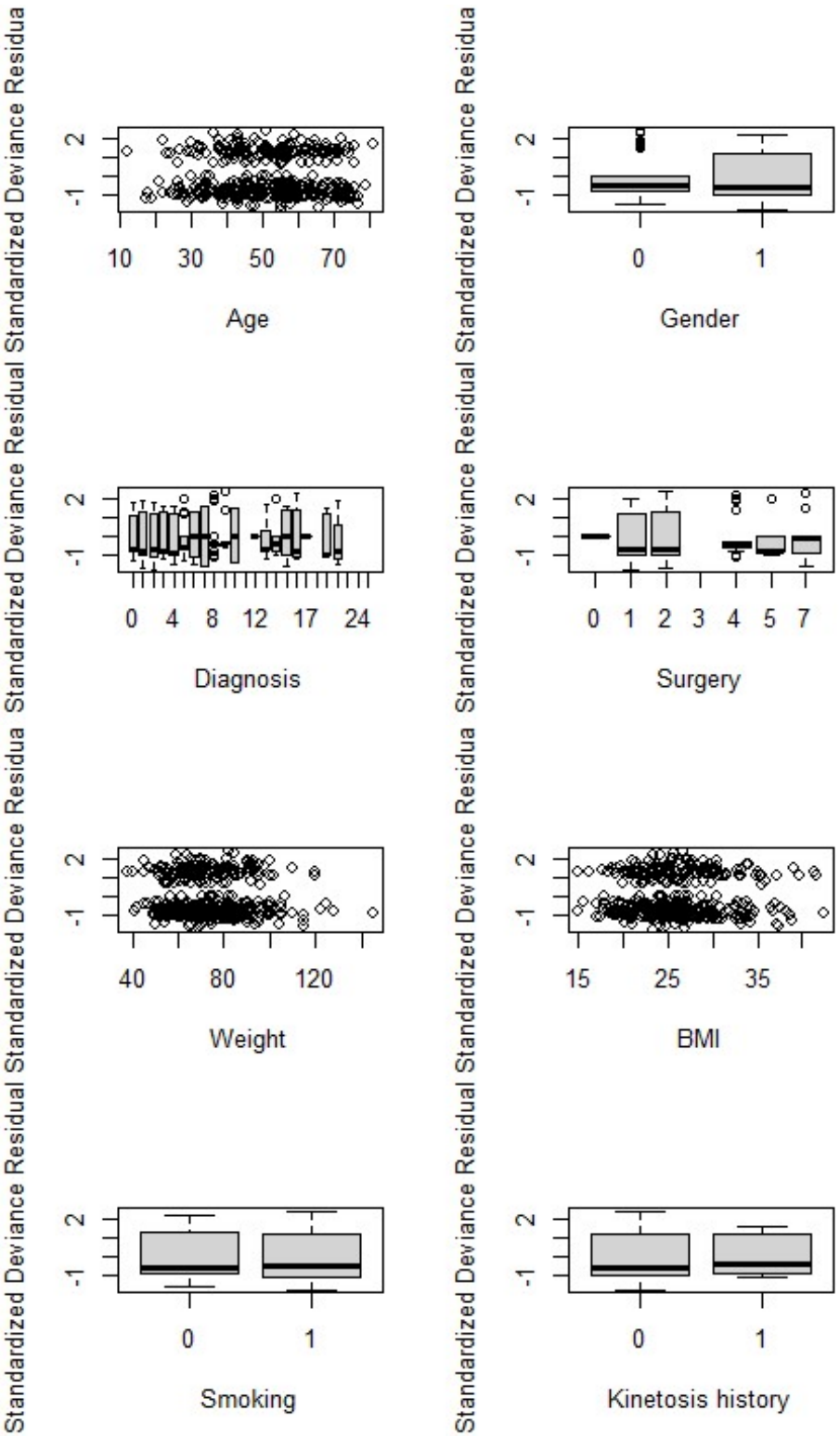
$$\theta(Y) = \frac{\exp(Y)}{1+\exp(Y)} = \frac{1}{1+\exp(-Y)}$$

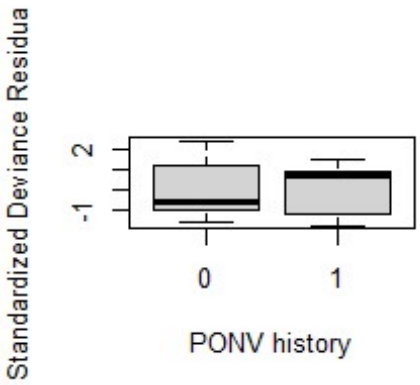
For this analysis, the θ of interest is the proportion of patients diagnosed with PONV within the 24 hours following surgery.

Exploratory data analysis

The logistic regression model was fitted using the generalized linear method of least squares. 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.

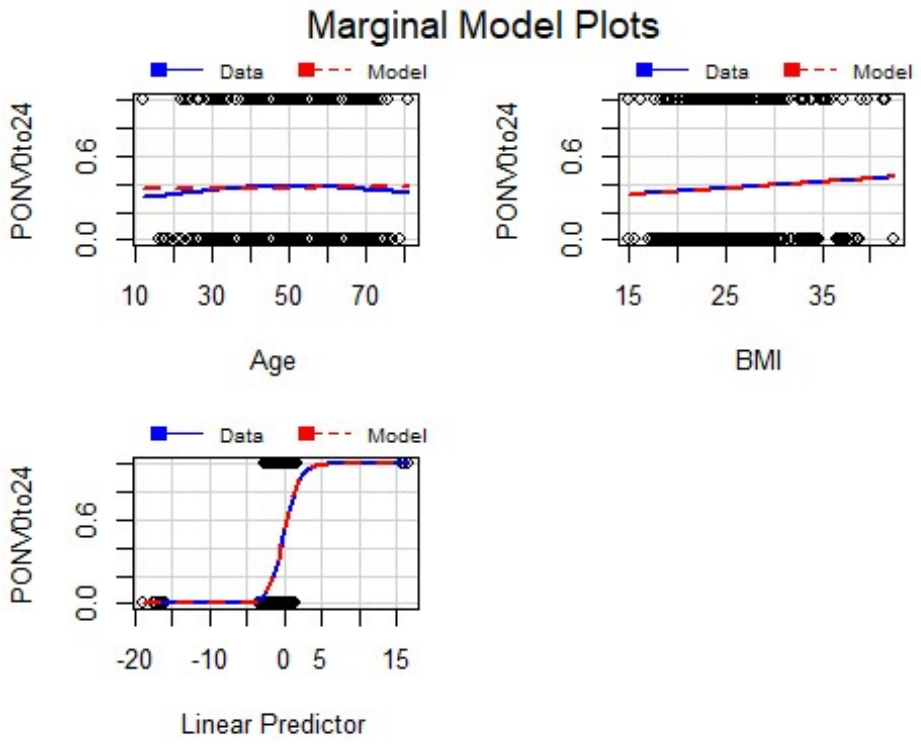
Plots of standardized deviance residuals


Standardized Deviance Residual
Standardized Deviance Residua
Standardized Deviance Residual
Standardized Deviance Residua
Standardized Deviance Residual
Standardized Deviance Residua
Standardized Deviance Residual
Standardized Deviance Residua
-1
2
10
30
50
70
Age
0
1
Gender
0
4
8
12
17
24
Diagnosis
0
1
2
3
4
5
7
Surgery
40
80
120
Weight
15
25
35
BMI
Smoking
Kinetosis history



Skewness is present in all 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 I examined marginal model plots instead.

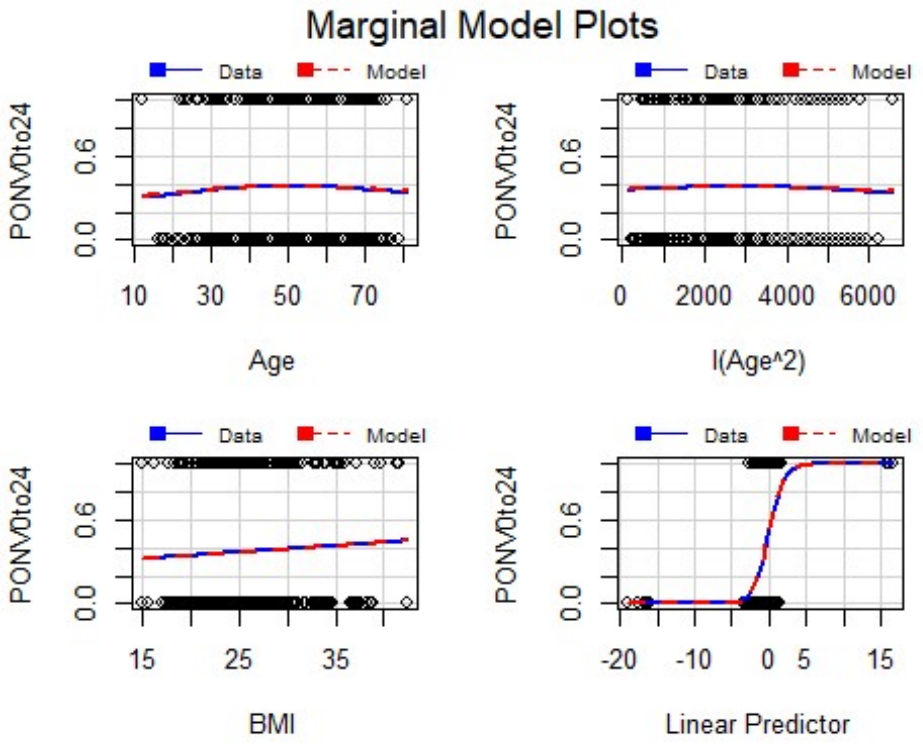
Marginal model plots for the continuous predictors



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.

$$g^{-1}(Y) = \log\left(\frac{\theta(Y)}{1-\theta(Y)}\right)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_{28} x_{28} + \beta_{29} x_{29} + \ldots + \beta_{35} x_{35} + \beta_{36} x_{36} + \beta_{37} x_{37} + \beta_{38} x_{38} + \beta_{39} x_{39} + e$$

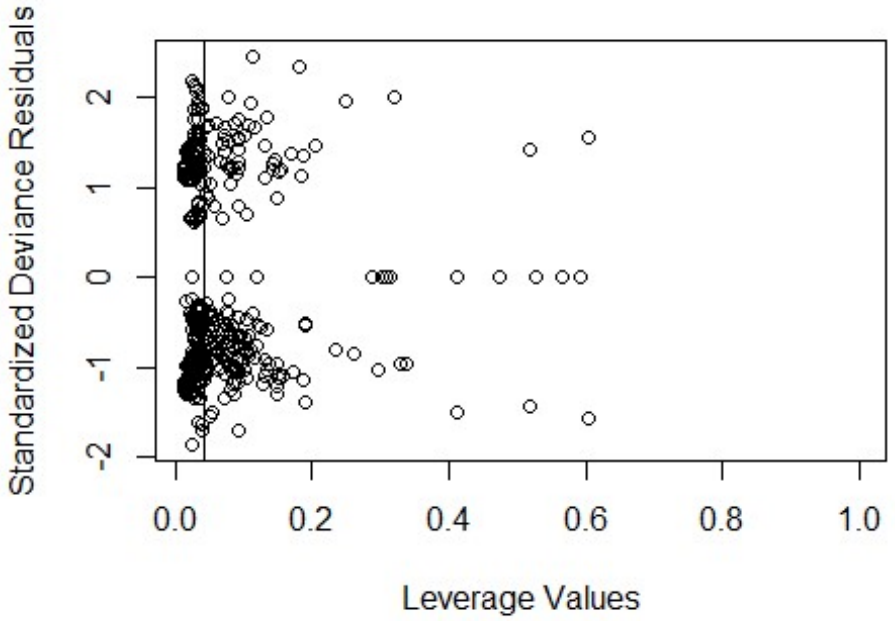
After adding a squared term for age to the full model, three of the predictors have estimated coefficients that are statistically significant. In descending order of significance, these are PONV history, gender, and nonsmoker. This is the same result obtained without the squared term for age.



After adding the quadratic term for age, there is reasonable agreement between the two fits (observed and predicted) 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

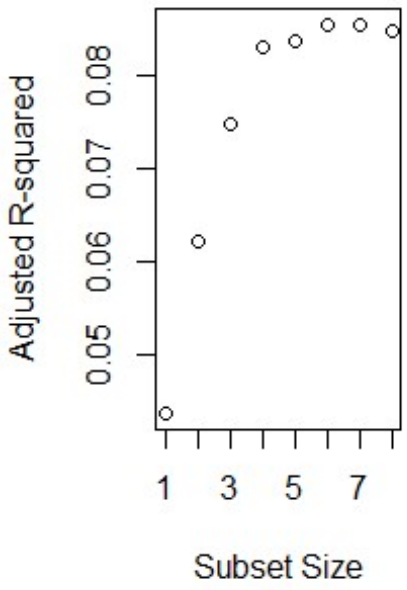
As a final validity check, I examined leverage values and standardized deviance residuals.



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, I continued with the assumption that the current model is an adequate fit for the data. Therefore, I next proceeded to variable selection.

Variable selection using all possible subsets

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



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

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

The maximum value of R^2 corresponds to the predictor subset of size seven. This is expected since R^2 increases (without penalty) as the number of predictors added to the model increases. This may lead to overfitting of the model to the data that it is trained on.

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 was added to each subset to obtain the next subset. The minimum value of AIC corresponds to the predictor subset of size four: PONVhistory, Surgery, Gender, and Nonsmoker. The minimum value of AIC_C also corresponds to the predictor subset of size four. The minimum value of BIC corresponds to the predictor subset of size two: PONVhistory and Surgery.

Parsimonious model selection

The minimum AIC and AIC_C values each correspond to the predictor subset of size four. 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. These are PONVhistory, Gender, and Nonsmoker. The fourth variable of the subset is Surgery. Furthermore, the predictor subset of size four has a higher R^2 value than the predictor subset of size two. Therefore, I chose the predictor subset of size three as the parsimonious logistic regression model:

$$Y = g(\beta_0 + \beta_1 PONVhistory + \beta_2 Gender + \beta_3 Nonsmoker + \beta_4 Surgery + e)$$

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

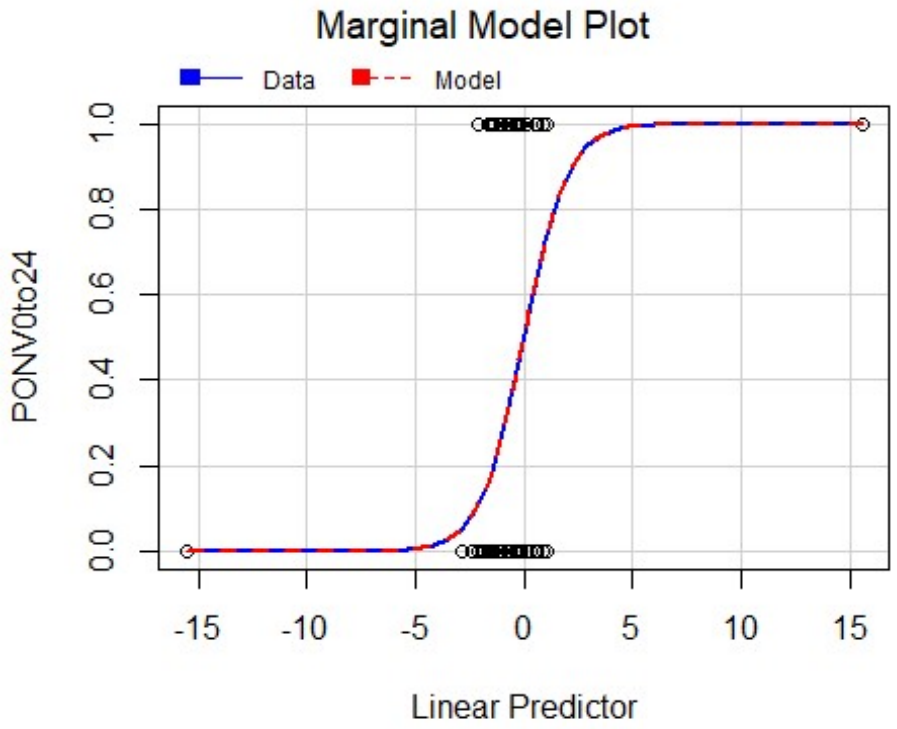
As before, I used the logit function to model the binary response variable:

$$g^{-1}(Y) = \log\left(\frac{\theta(Y)}{1-\theta(Y)}\right) = \beta_0 + \beta_1 PONVhist + \beta_2 Gender + \beta_3 Nonsmoker + \beta_4 Surgery + e$$

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

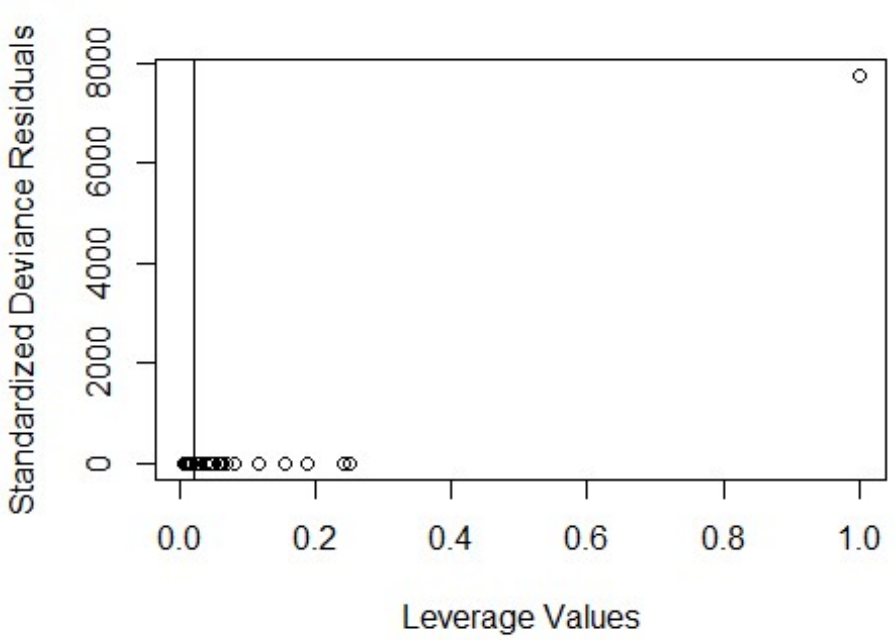
The chosen model has three estimated coefficients that are statistically significant, which are the same three that are statistically significant in the full model. In descending order of significance, these are PONVhistory, Gender, and Nonsmoker. The fourth predictor

in the model is Surgery, but none of the estimated coefficients for this factor predictor variable are statistically significant.



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 (actual and predicted) in the marginal model plot for the linear fit. This indicates that the model is an adequate fit for the data.

As a final validity check, I next looked at leverage values versus standardized deviance residuals.



The plot of leverage values and standardized deviance residuals consists of a single extreme leverage point, while the remaining points are all clustered closer to a standard deviation of zero. Recall that none of the coefficients for Surgery were statistically significant. Furthermore, all have high standard errors, with one being more than double the value of the others. Finally, the predictor subsets of three and four have comparable R^2. So, I next removed the Surgery variable to obtain the predictor subset of three as the parsimonious model:

$$Y = g(\beta_0 + \beta_1 PONVhistory + \beta_2 Gender + \beta_3 Nonsmoker + e)$$

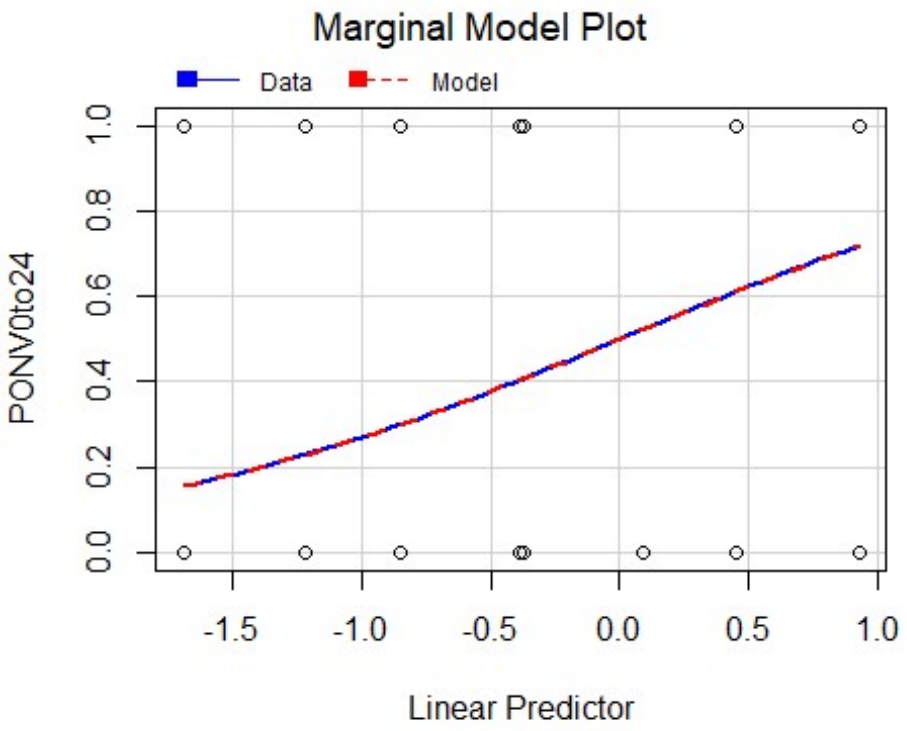
where $e \sim$ iid $N(0,1)$.

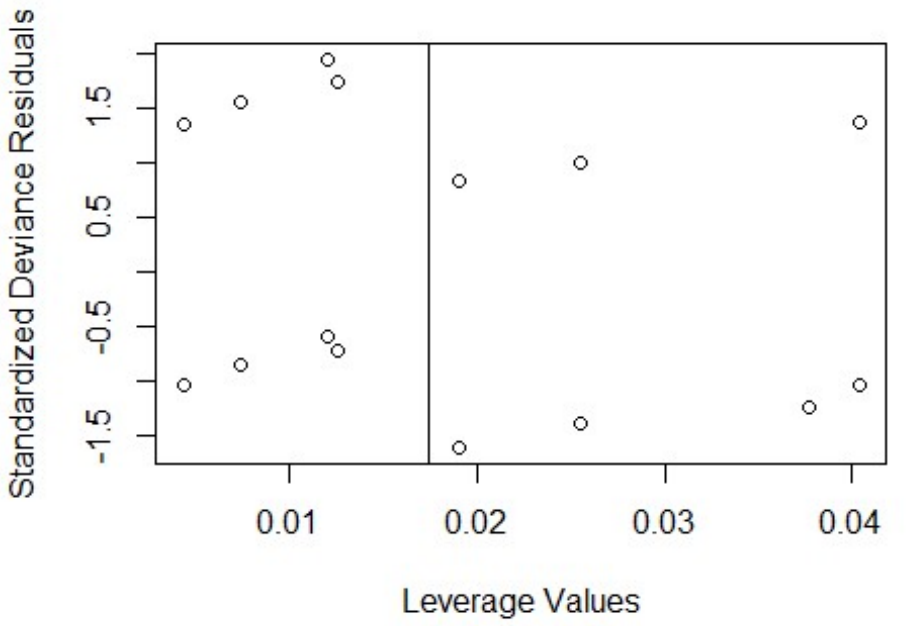
As before, we use the logit function to model the binary response variable:

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

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

```
##
## Call:
## glm(formula = PONV0to24 ~ PONVhistory + Gender + Nonsmoker, family =
binomial,
##     data = ponv)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.588  -1.023  -0.721   1.340   1.929
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.6911     0.3027  -5.587 2.32e-08 ***
## PONVhistory1   1.3029     0.3132   4.160 3.19e-05 ***
## Gender1        0.8401     0.2803   2.997  0.00273 **
## Nonsmoker1     0.4766     0.2159   2.208  0.02724 *
## ---
## 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: 573.95  on 457  degrees of freedom
## AIC: 581.95
##
## Number of Fisher Scoring iterations: 4
```

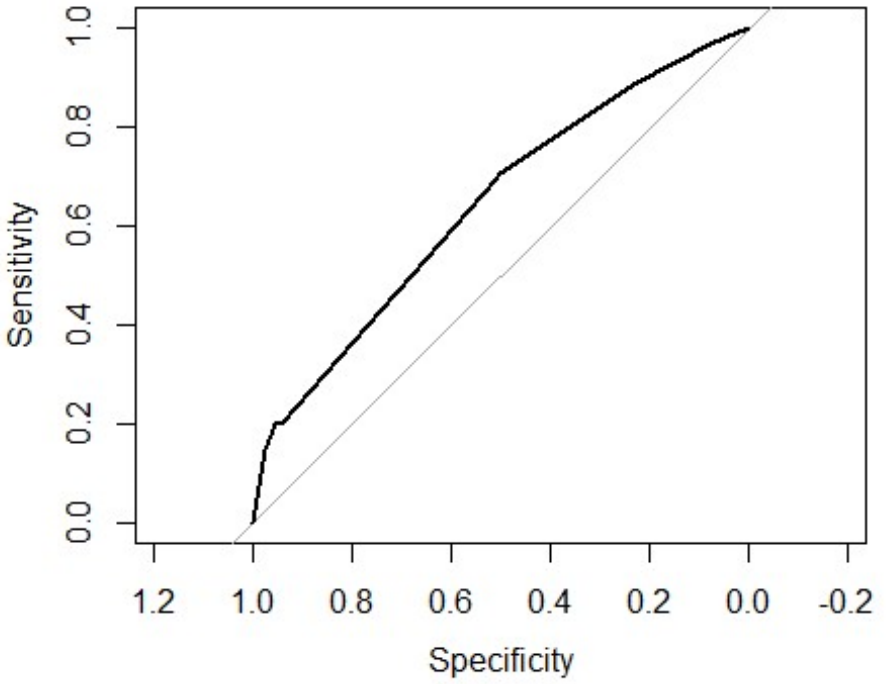


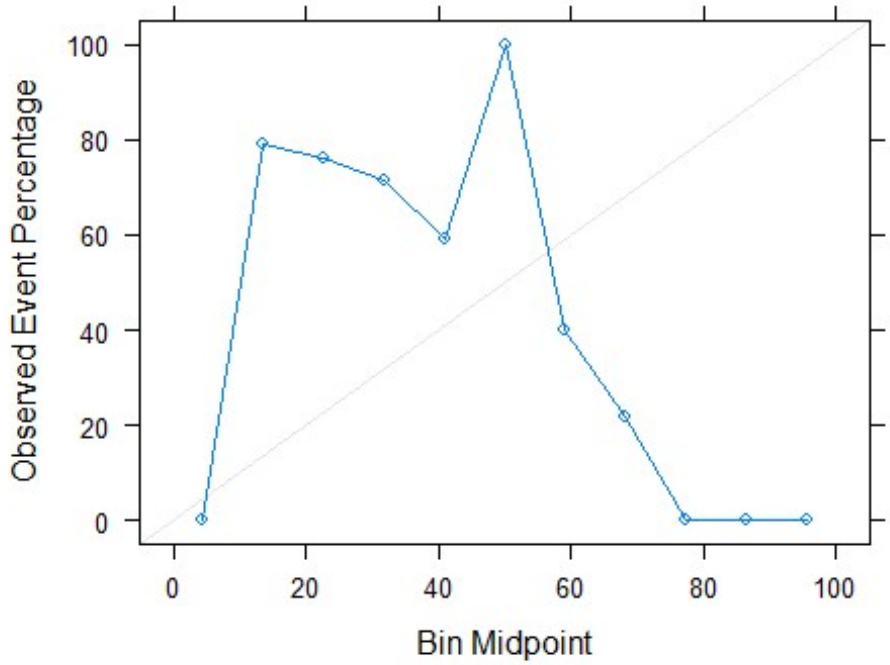


In the current model, the intercept and all three predictors are statistically significant. The marginal model plot shows reasonable agreement between the two fits (actual and predicted) for the linear fit. This indicates that the model is an adequate fit for

the data. 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. Next, I proceeded to assess the predictive ability of this model.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 272 138
##          1  17  34
##
##                Accuracy : 0.6638
##                  95% CI : (0.6186, 0.7068)
##     No Information Rate : 0.6269
##     P-Value [Acc > NIR] : 0.05527
##
##                   Kappa : 0.1619
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.9412
##             Specificity : 0.1977
##          Pos Pred Value : 0.6634
##          Neg Pred Value : 0.6667
##              Prevalence : 0.6269
##          Detection Rate : 0.5900
##    Detection Prevalence : 0.8894
##       Balanced Accuracy : 0.5694
##
##        'Positive' Class : 0
##
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```





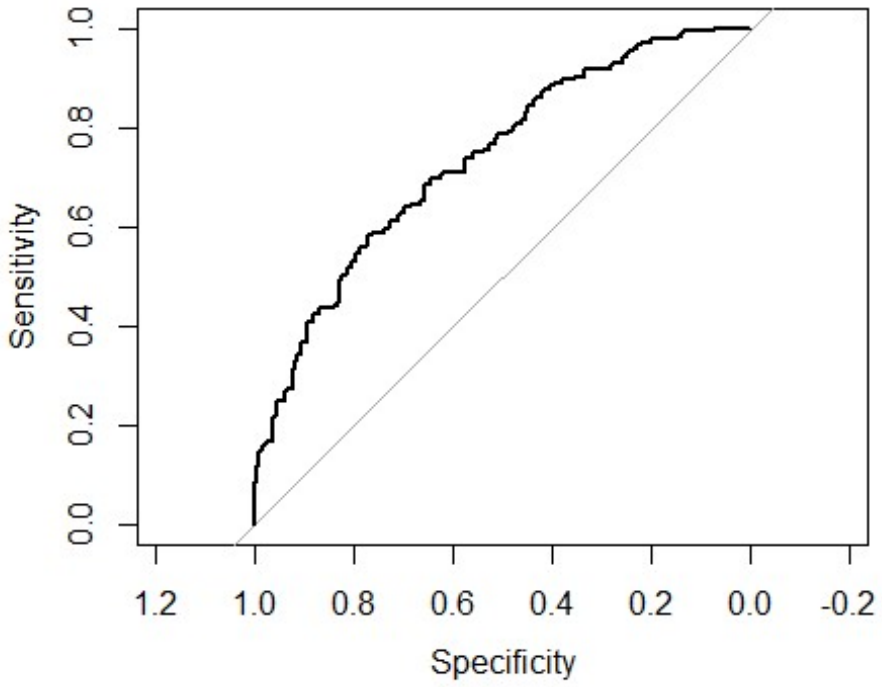
The area under the ROC curve (AUC) is 0.639. Since the ROC curve is a function of both sensitivity and specificity, the curve is insensitive to class imbalance. With a sensitivity of 0.94 and a specificity of 0.20, the model is good at classifying high risk

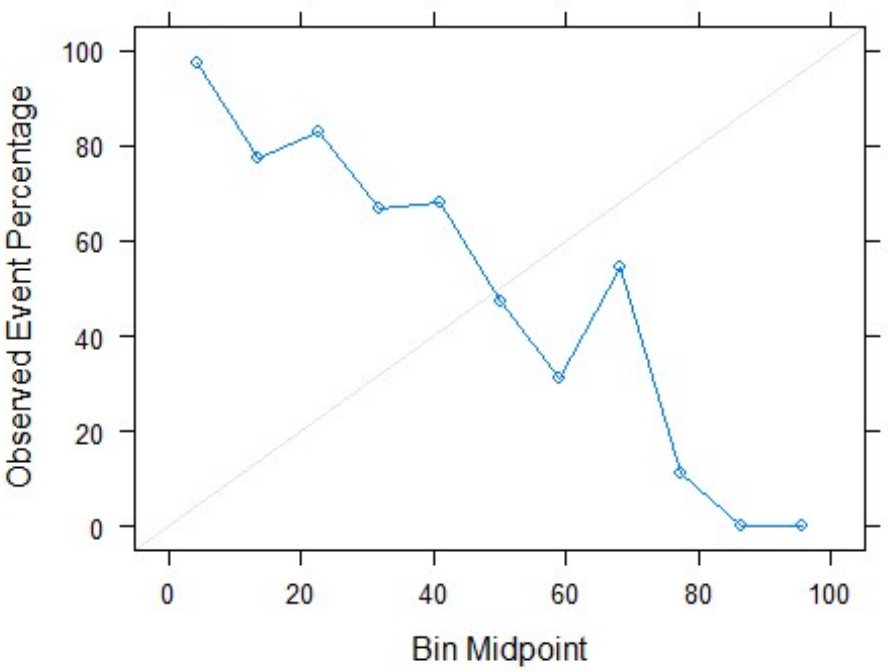
patients, but poor at classifying low risk patients. This is reflected in the calibration plot, which has a slope that underestimates patients with low PONV risk, and overestimates patients with high PONV risk.

Resampling techniques

k-*fold cross-validation*

We next perform logistic regression using five repeats of 10-fold cross-validation, to generate 50 different holdout 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.





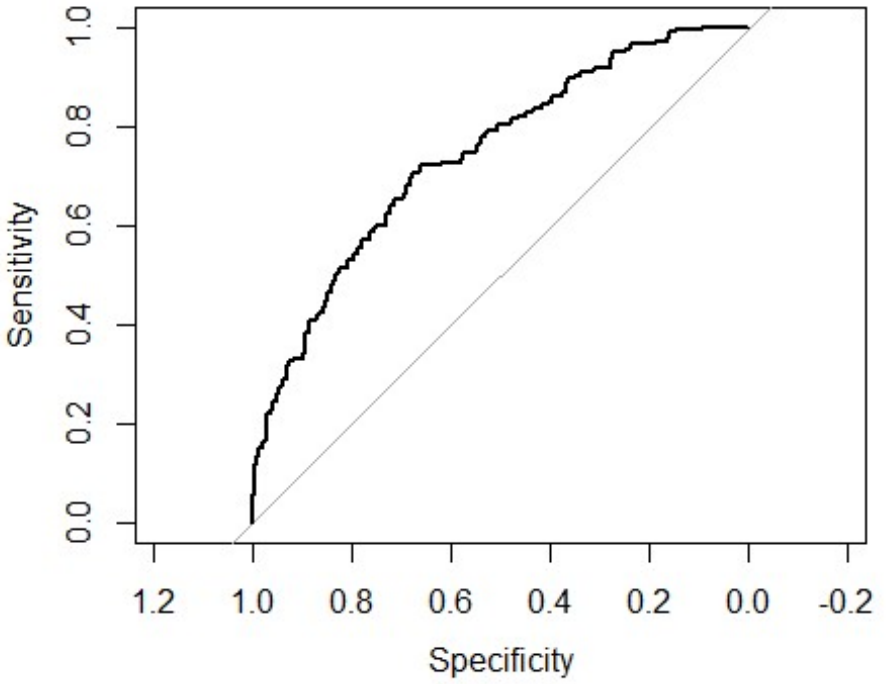
Repeated 10-fold cross-validation resulted in a logistic regression model with an AUC of 0.7358, which is an improvement over the AUC of 0.64 obtained for our baseline model developed from all possible subets. With a sensitivity of 0.31 and a specificity of 0.83, the model is poor at classifying high risk patients, but good at classifying low risk patients.

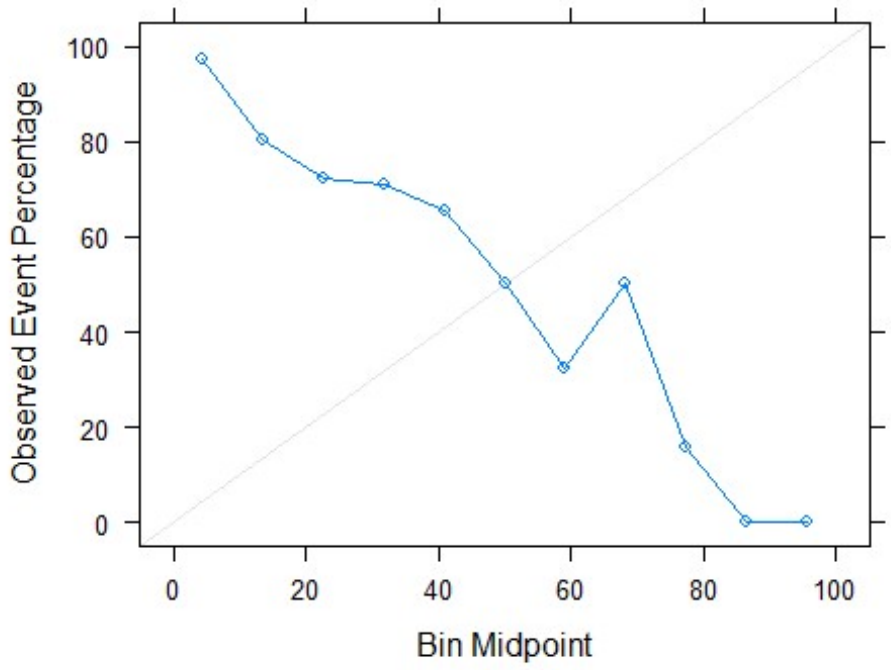
The choice of k to be 10 for k-fold cross-validation avoids the high bias of smaller values of k, as well as the computational burden of higher values of k. k-fold cross-validation generally has high variance compared to other methods. The potential issues with bias and variance become negligible for 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.

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

Next, I performed the bootstrap technique of resampling on the full logistic regression model. A random sample equal to the size of the data set was taken *with replacement*. This was repeated 25 times to fit the full logistic regression model.





Bootstrapping resulted in a logistic regression model with an AUC of 0.7408, which is an improvement over our two preceding models. With a sensitivity of 0.34 and a specificity of 0.78, this model provides the most reasonable balance of our three models.

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 preceding two models.

Predictions

To conclude, I made predictions of PONV for some examples of hypothetical patients. The logistic regression model obtained from all possible subsets consists of three predictors which are all binary variables. Since it lacks the numerous dummy variables of the two models trained with resampling techniques, I choose it as the parsimonious model to make predictions.

```
mean(predict(glmFit4, data.frame(PONVhistory="1", Gender="1",
Nonsmoker="1"),type="response"))
## [1] 0.7167777
mean(predict(glmFit4, data.frame(PONVhistory="1", Gender="1",
Nonsmoker="0"),type="response"))
## [1] 0.6110923
mean(predict(glmFit4, data.frame(PONVhistory="1", Gender="0",
Nonsmoker="1"),type="response"))
## [1] 0.5221021
mean(predict(glmFit4, data.frame(PONVhistory="1", Gender="0",
Nonsmoker="0"),type="response"))
## [1] 0.4041599
mean(predict(glmFit4, data.frame(PONVhistory="0", Gender="1",
Nonsmoker="1"),type="response"))
## [1] 0.407475
mean(predict(glmFit4, data.frame(PONVhistory="0", Gender="1",
Nonsmoker="0"),type="response"))
## [1] 0.2992145
mean(predict(glmFit4, data.frame(PONVhistory="0", Gender="0",
Nonsmoker="1"),type="response"))
## [1] 0.2289091
mean(predict(glmFit4, data.frame(PONVhistory="0", Gender="0",
Nonsmoker="0"),type="response"))
## [1] 0.15563
```

As expected, the more significant predictors that a patient has, the more likely the patient will have PONV. A patient with all three risk factors in the parsimonious model has a 72% probability of experiencing PONV, while a patient with none of the three risk factors has a 16% probability of experiencing PONV. Having previously determined that this is a valid predictive model, it may have practical application for patient populations with characteristics like the data set of this investigation. In choosing a threshold for prescription of prophylaxis, healthcare professionals could select one of the two hypothetical predictions in our example having a probability greater than 50%.

Acknowledgments

This semester-long investigation was conducted under the guidance of Professor Brani Vidakovic, PhD, and in collaboration with graduating classmates (MS in Statistics) Scott Tillet and Nguyet Vu, in Spring 2022, for the Project in Statistics (STAT 685) graduate course at Texas A&M University.

References

Literature

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.

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.

Textbooks

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.

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