Categorical Data Analysis Final Project

Sherman Selix

12/16/2013

Methods

Variables

This dataset comes from the American College of Surgeons National Surgical Quality Improvement Program. This particular dataset takes 3001 patients who underwent lung surgery and looks at 113 variables associated with mortality or morbidity. Some of these data are nominal, some are ordinal, and some are continuous. For the purposes of this study, the outcome variable of interest is "Death", which indicated whether the patient died within 30 days of surgery. All 3001 patients had information on mortality, with 72 deaths and 2929 survivors.

The goal of the study is to create a logistic regression model to look at the risk factors associated with this 30-day mortality data. For logistic regression, ordinal variables are treated as continuous. Nominal categorical variables must have at least ten observations of every category in order to fit the assumptions of logistic regression. In addition, many variables were missing observations for many of the patients. Logistic regression analysis can only be completed on patients that have observations for every variable in the model selection. Using several variables with a significant portion of the data missing would result in the omission of a very large portion of the patients, especially if the missing observations do not overlap between variables. The omitted patients could be different from the included ones in significant and unknown ways. For this purpose, variables are omitted if they lack observations for more than 10% of the patients. In this dataset, this meant omitting variables with more than 300 missing observations.

The data set was analyzed with SAS. The PROC FREQ procedure was used to examine the counts of observations within each category of categorical variables; if a categorical variable contained less than 10 observations within at least one category, it was omitted. It also examined the amount of missing observations for each variable; if a variable was missing more than 300 observations, it was also omitted. If a variable was a repeat of another variable, it was omitted. The results of the variable analysis are the following:

Variable Type	**Count**
Response	1
Nominal	46
Ordinal/Continuous	22
Omitted	44

Statistical Analysis

The response variable is 30-day mortality. To analyse the data, logistic regression was used to model the probability of 30-day mortality, π, by predictors, $x_1, x_2, x_3, \ldots, x_{p-1}$, such that:

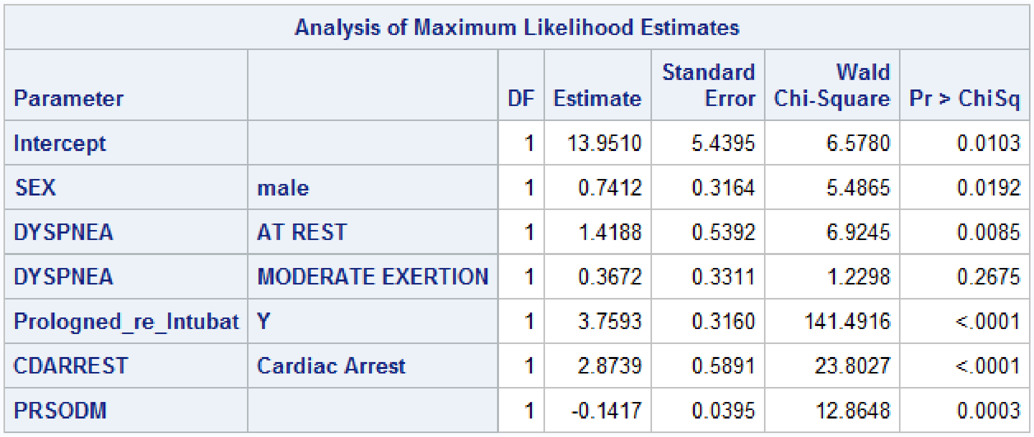
$$log(\frac{\pi(x_1,x_2,x_3,\ldots,x_{p-1})}{1-\pi(x_1,x_2,x_3,\ldots,x_{p-1})}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_{p-1} x_{p-1}$$

PROC LOGISTIC in SAS was used to find the logistic regression fit. The following steps were taken to find the proper regression model.

1. A saturated logistic regression model was entered into PROC LOGISTIC, with stepwise selection criteria. The α-to-entry was set at 0.1 and the α-to-exit was set at 0.15, as used in class.
2. The saturated model resulted in 1041 patients with an observation missing in at least one variable. Therefore, these patients were not included in the selection processes. This creates problems, as the final selection will use less variables and, hence, include some of these 1041 patients. The process of eliminating variables with more than 10% observations missing was intended to diminish this. As 1041 is less than half of the dataset, it was assumed to be adequate to find an adequate model fit.
3. The stepwise selection resulted in a model with seven predictors, five nominal and two ordinal/continuous.
4. This model was analysed on its own. With only these seven predictors in the model, 202 patients were not included from a missing observation in at least one variable. This means 839 additional patients were used to fit a logistic regression model between the saturated stepwise selection and the reduced model fitting.
5. To further refine the model, an α of 0.05 was chosen to eliminate predictors based on Wald Chi-Square Type 3 analysis. This is because the analysis is used not to acquire the best fit to predict 30-day mortality but to identify independent risk factors. In the reduced model, two variables exceed 0.05. They were removed one at a time.
6. To check for interaction terms, all ten potential interaction terms were tested one at a time with the five remaining predictors. None were significant.
7. Influence and lack of fit tests were conducted. The final model with five predictors was chosen as adequate.

<u>Results</u>

The final model is as follows:



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	13.9510	5.4395	6.5780	0.0103
SEX	male	1	0.7412	0.3164	5.4865	0.0192
DYSPNEA	AT REST	1	1.4188	0.5392	6.9245	0.0085
DYSPNEA	MODERATE EXERTION	1	0.3672	0.3311	1.2298	0.2675
Prologned_re_Intubat	Y	1	3.7593	0.3160	141.4916	<.0001
CDARREST	Cardiac Arrest	1	2.8739	0.5891	23.8027	<.0001
PRSODM		1	-0.1417	0.0395	12.8648	0.0003

$$Logit[Y = 1] = 13.951 + 0.741x_1 + 1.419x_{21} + 0.367x_{22} + 3.759x_3 + 2.874x_4 - 0.142x_5$$

Variables, followed by the significance of '1' for nominal variables:

$Y = 30 - Day\ Mortality,\ yes$

$x_1 = sex,\ male$

$x_{21} = Dyspnea,\ at\ rest$

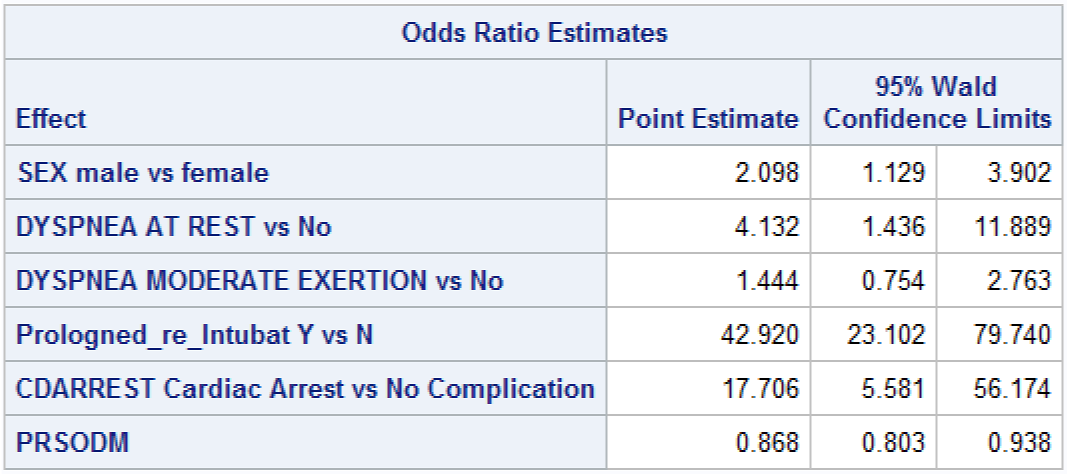
$x_{22} = Dyspnea,\ Moderate\ exertion$

$x_3 = Intubation\ required\ during\ operation,\ yes$

$x_4 = Whether\ the\ patient\ experienced\ cardiac\ arrest,\ yes$

$x_5 = Preoperative\ serum\ sodium,\ continuous$

Analysed as odds ratios, the selected predictors are as the following:



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SEX male vs female	2.098	1.129	3.902
DYSPNEA AT REST vs No	4.132	1.436	11.889
DYSPNEA MODERATE EXERTION vs No	1.444	0.754	2.763
Prologned_re_Intubat Y vs N	42.920	23.102	79.740
CDARREST Cardiac Arrest vs No Complication	17.706	5.581	56.174
PRSODM	0.868	0.803	0.938

Interpretations of odds ratios:

With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality among men are between 1.129 and 3.902 times the odds among women.

With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality among those with dyspnea "at rest" are between 1.436 and 11.889 times the odds among those with no dyspnea.

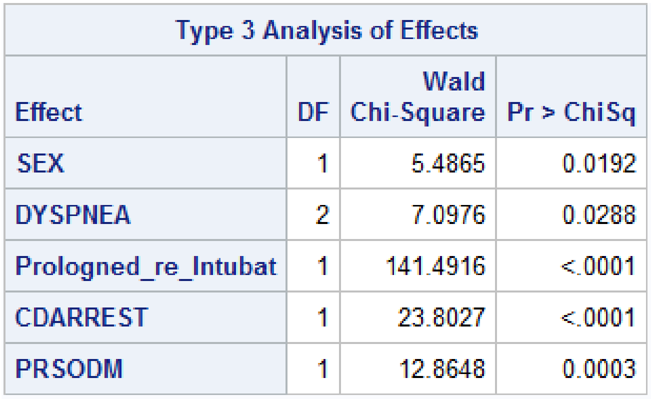
With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality among those with dyspnea "moderate exertion" are between 0.754 and 2.763 times the odds among those without dyspnea.

With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality among those who required intubation are between 23.102 and 79.740 times the odds among those who did not.

With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality among those who suffered cardiac arrest are between 5.581 and 56.174 times the odds among those who did not.

With a confidence coefficient of .95, it is estimated that the odds of 30 day mortality increase by a factor between 0.803 and 0.938 for every increase in one unit of preoperative sodium serum concentration.

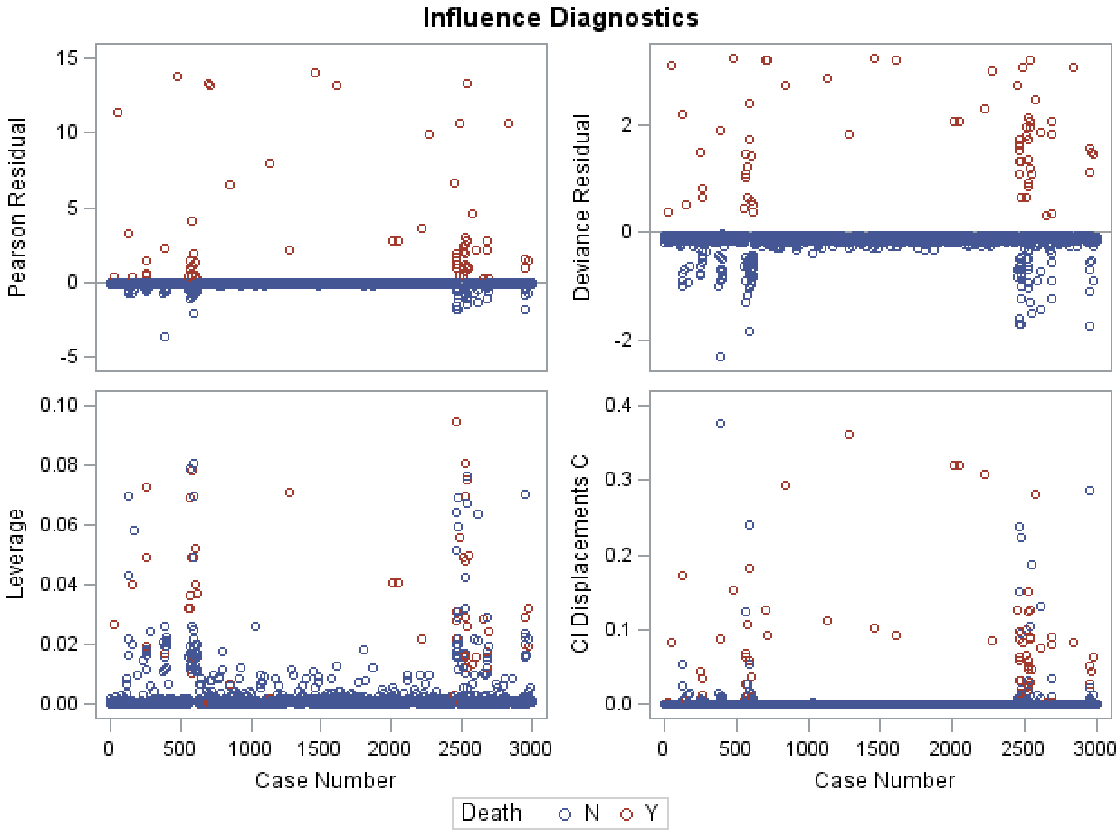
The following are the type 3 Wald-Chi Square statistics:



Type 3 Analysis of Effects

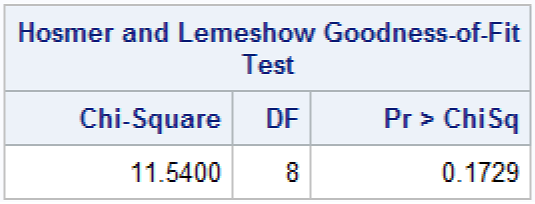
Effect	DF	Wald Chi-Square	Pr > ChiSq
SEX	1	5.4865	0.0192
DYSPNEA	2	7.0976	0.0288
Prologned_re_Intubat	1	141.4916	<.0001
CDARREST	1	23.8027	<.0001
PRSODM	1	12.8648	0.0003

Here are charts that look at measures of influence for certain patients on the data set:



The Pearson Residual plot shows several outliers, defined as an absolute Pearson residual statistic exceeding 3. In fact, some standardized residuals exceed 10. However, the amount of outliers pales in comparison to the size of the dataset overall. Therefore, the residuals may only pose a minor problem and not suggest lack of fit. Further tests must be conducted.

The following is the test for goodness of fit. We fail to reject our null hypothesis and assume the model adequately fits the data:



Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
11.5400	8	0.1729

Discussion

This study examined 30-day mortality associated with lung surgery. Data was collected on 3001 patients and looks at 113 potential risk factors or determinants of mortality and morbidity involved with lung surgery. Of these 3001 patients, 72 died within 30 days. The other variables were examined to constructed a logistic regression model to identify complications and risk factors associated with 30-day mortality.

The results of this logistic regression model are above. The results suggest that 30-day mortality is independently associated with the sex of the patient, whether the patient suffered from Dyspnea and what type they suffered from, whether intubation was required during the procedure, whether the patient suffered from cardiac arrest, and the preoperative sodium serum level of the patient. The point estimates for the odds ratios associated with intubation and cardiac arrest were 42.9 and 17.7, respectively. These could both be considered “overwhelming” or “dramatic” results, indicating a very strong, and perhaps causal association. These are the two most significant findings.

The most significant potential limitation of this study is the incomplete dataset. The missing observations for several variables led to the elimination of many potential risk factors from consideration in the regression model. Even with this elimination, when using standard selection criteria, the saturated model did not consider more than 1041 patients, while the chosen reduced model on its own did not consider 202 patients. This meant that the dataset used for selection criteria was essential not the same one used for final analysis. This suggests a superior model could have potentially been developed had the selection process only removed the patients lacking observations for the subset of variables under consideration, not all the patients lacking observations in the saturated model, a much larger number. Another potential limitation is that the pearson standard residuals for the chosen model are by and large positive for the patients that died, and some of these residuals are very large. This could mean that, while the model may adequately estimate when a patiently will live after 30 days, the model may systematically underestimate when a patient will die. This model is adequate to postulate on the odds ratios of independent risk factors of 30-day mortality, but it may not be an effective estimator to predict the probability of mortality based on these risk factors.