Diana Zabala

Estimating Likelihood of Hospital Readmission for Coronary Artery Bypass Graft (CABG) Surgery Patients

Model Results

I produced a logistic regression model to estimate the likelihood of hospital readmission for CABG patients in the dataset. I controlled for age, systolic blood pressure, and the number of comorbidities for each patient, as well as the ID of the hospital the patient was most recently admitted to and indicators for each type of comorbidity. The variables ind_diabwc (indicator for diabetes with complications) and ind_canc (indicator for cancer) were excluded to resolve multicollinearity concerns. The results are shown in **Table 1**.

Table 1. Logistic Regression Model Results

Term	Coefficient	Std. Error	Z-Statistic	P-value	Odds Ratio	% Change in Likelihood
(Intercept)	-4.77	0.97	-4.94	0.00		
Age	0.03	0.01	2.27	0.02	1.03	2.8%
Systolic BP	-0.01	0.00	-2.42	0.02	0.99	-0.7%
Total Number of Comorbidities	1.22	0.27	4.52	0.00	3.39	238.7%
Hospital ID	-0.04	0.04	-0.90	0.37	0.96	-3.8%
Myocardial Infarction	0.41	0.59	0.70	0.48	1.51	51.0%
Congestive Heart Failure	-14.45	699.11	-0.02	0.98	0.00	-100.0%
Peripheral Vascular Disease	0.31	0.56	0.56	0.58	1.37	36.7%
Cerebrovascular Disease	-0.36	0.47	-0.76	0.45	0.70	-30.0%
Dementia	1.05	0.87	1.21	0.23	2.86	185.7%
Chronic Pulmonary Disease	-0.44	0.43	-1.02	0.31	0.65	-35.5%
Rheumatoid Disease	-1.39	0.92	-1.51	0.13	0.25	-75.0%
Peptic Ulcer Disease	-0.61	0.66	-0.93	0.35	0.54	-45.8%
Mild Liver Disease	0.47	0.68	0.69	0.49	1.60	60.2%
Diabetes Without Complications	-1.00	0.94	-1.06	0.29	0.37	-63.2%
Hemiplegia/ Paraplegia	-1.26	1.08	-1.17	0.24	0.28	-71.7%
Renal Disease	-1.24	0.69	-1.79	0.07	0.29	-71.0%
Moderate/Severe Liver Disease	-15.75	1006.66	-0.02	0.99	0.00	-100.0%
Metastatic Cancer	0.15	0.63	0.24	0.81	1.16	16.2%

As highlighted above in orange, only age and the total number of comorbidities for the patient were significant at an alpha level of 0.05. These variables can be interpreted as follows:

- For each additional year of age, a patient's likelihood of being readmitted increases by about 3% on average, all else held constant.
- For each one unit increase in systolic blood pressure, a patient's likelihood of being readmitted decreases by about 1%, all else held constant.
- For each additional comorbidity a patient has, the likelihood of being readmitted becomes 3.4 times more likely on average, all else held constant.

Model Selection

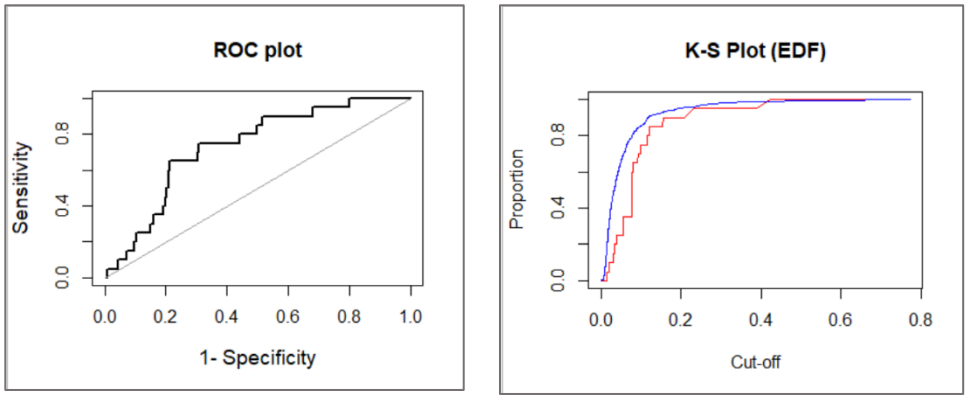
Readmission is a binary categorical variable, so the model needed to be able to predict a categorical dependent variable. This ruled out linear regression, which only makes predictions for continuous dependent variables.

I chose to develop a logistic regression model because it can estimate likelihood of occurrence for binary dependent variables. Although there are other models that can make predictions for a binary dependent variable, such as tree-based models or Naïve Bayes classifiers, I chose to use logistic regression due to the interpretability of the model's coefficients.

After resolving multicollinearity concerns, I trained a model data using a random sample of 70% of the data (n=1205). The results are shown above in **Table 1**.

I then assessed the model's performance on the held-out test data (n=516). The model had a Receiver Operator Curve (ROC) area under the curve (AUC) value of 73% indicates the rate at which the model assigns higher probabilities of readmission to those patients who were truly readmitted compared to those who were not. The model had a Kolmogorov–Smirnov (K-S) statistic of 0.442 corresponding to a probability cutoff value of 0.0551, which I used to classify patients as readmitting or not.

Figures 1a and 1b: (a) ROC plot; (b) K-S plot



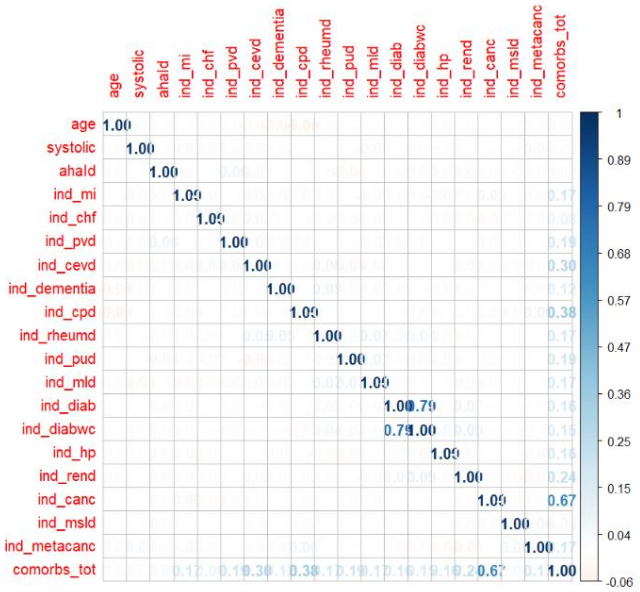
The classification using the K-S statistic to generate a cutoff value resulted in a sensitivity of 70% and a specificity of 69%. Because of high risks to patients of foregoing care, I adjusted the probability cutoff value to 0.0500 to obtain a sensitivity of 75% and a specificity of 66%. Lowering the threshold further would result in severe reductions of specificity, so no further adjustments were made.

Model Assumptions

Five binary logistic regression model assumptions need to be met to produce unbiased estimates. My assessment of how well these assumptions were met is as follows:

1. **Independence of observations**: I retained only one row for each unique patient ID to ensure independence of observations.
2. **Little to no multicollinearity:** I assessed this using the variance inflation factor (VIF) for the model as well as Spearman's rank correlation tests between predictor variables. When all variables were included in the model, there was strong multicollinearity according to the VIF function I was using in R. Thus, I created and plotted a matrix of Spearman's rank correlation coefficients between all the independent variables, as shown below in **Figure 1**.

   Figure 2. Matrix of Spearman Rank Test Coefficients



   The matrix shows that ind_diab (diabetes indicator) and ind_diabwc (diabetes with complications indicator) are strongly correlated, as are ind_cancer (cancer indicator) and comorbs_tot (total number of comorbidities). I chose to move forward with including ind_diab and comorbs_tot in the model while excluding ind_diabwc and ind_cancer. After removing those two variables, the remaining variables had a VIF of under 5.
3. **No influential outliers:** I assessed this using Cook's Distance metric to detect highly influential observations. I deemed an observation to be influential if the Cook's Distance metric exceeded 4 divided by the training sample size n=1205, or 0.0023. I also calculated the absolute standardized residuals and found none that exceeded 3 standard deviations, so there were no outliers and thus no influential outliers.
4. **Linear relationship between independent variables and the log-odds (logit) of readmission**: I assessed this assumption by fitting a generalized additive model (GAM) with the same variables as the logistic regression model and using ANOVA to determine whether the difference between the two models is statistically significant. The resulting chi-squared statistic was very small (-1.4257e-05), so we conclude no difference between the two models. This indicates that

a linear relationship between the logit and the predictor variables is a good fit for the data and the assumption of linearity is met.

5. **Large sample size**: According to Long (1997)[1], sample sizes of 500 observations or greater are adequate for most logistic regression modeling situations. The sample size for the training dataset is n=1205, so this assumption is met.

[1] Long, J.S. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: SAGE Publications, Inc.