Case Study

Zach White

October 10, 2017

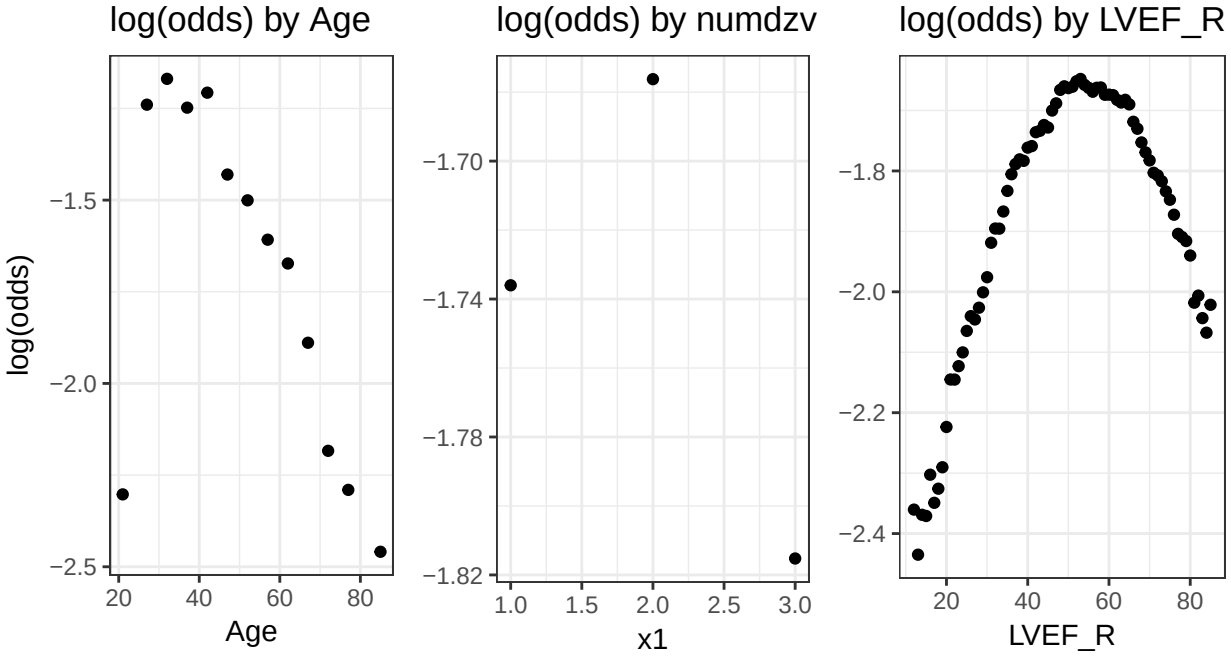
Introduction

Using data from the Duke Databank for Cardiovascular Diseases on patients undergoing a cardiac catheterization procedure, we seek to understand which factors are most predictive of a subsequent myocardial infarction (MI). We use generalized linear models with different link functions to analyze the relationships between explanatory variables and whether the patient experienced a MI.

Specifically, we will perform initial exploratory data analyses to understand the nature of certain variables and to assess the appropriateness of certain link functions. We decide to use all explanatory variables and some interaction terms, and we then fit our model using three different link functions to compare the estimates and predictive performance. After comparison, we decide to use the logit link function for our model. We analyze predictive performance through k-fold cross-validation, and we assess the appropriateness of the model through deviance residuals, pseudo-R^2, and Receiver Operating Characteristic (ROC) curves.

Exploratory Analysis

We first perform some exploratory data analysis. We are interested in one response variable, specifically, myocardial infarction (MI), which indicates whether the patient has a MI. This is clearly binary, and thus we want to use some generalized linear model for binary regression. There are 14 explanatory variables. Something of note with these data is that there are missing data, especially in the LVEF_R variable, which is continuous. Out of 11,201 missing values, 9,569 are from the LVEF_R variable. Although this might not be best in practice, we choose to discard these values because there are still 22,366 observations without missing values. In practice, we should be hesitant to discard so many observations because we are losing significant information. We could do multiple imputation and conduct our analysis on each of the imputed datasets, but for this case study, we will simply discard any observation with missing values.



Most of the explanatory variables are categorical variables with two levels, but there are two variables that are coded as categorical which could be considered continuous. The first of which is age (categorized as 1=18-24; 2=25-29; 3=30-34; 4=35-39; 5=40-44; 6=45-49; 7=50-54; 8=55-59; 9=60-64; 10=65-69; 11=70-74; 12=75-79; 13=>=80). We will test whether to use this as a factor or treat this as a continuous variable. The following plot shows the observed $\log(\frac{\hat{p}}{1-\hat{p}})$ for the midpoint of the age range of each of the levels. It looks quite linear, albeit the extremely low value for the first value. After this first value, it does look linear. We could treat this as a type of interaction term where we only use the age variable if it is not the youngest age group. We could also treat this as a polynomial term on age. The other possible variable that could be treated as a count instead of categorical variable is the number of deceased vessels found in the catheterization. The middle plot shows the observed log-odds, and it seems it is not necessary to use as a continuous variable, especially since there are only three different values (1,2, and 3).

The other continuous variable is LVEF_R, which represents the left ventricular ejection fraction (%) as measured during the catheterization. The plot on the right shows log-odds for binned values of LVEF_R. We use a moving log-odds for a range of 20% LVEF_R , and this is clearly quadratic. Thus, when modeling, we should make sure to add a second-order polynomial to this term. Without this, a simple linear term would not be able to capture trends in this way.

Methodology and Model Selection

We explore three different link functions for this binary regression model: logit, probit, and complementary log log. These can be described as the following:

$$\text{Logit: } log\big(\frac{\pi_i}{1-\pi}\big) = \boldsymbol{X}_i\boldsymbol{\beta}$$

$$\text{Probit: } Pr(Y_i = 1|\boldsymbol{X}_i) = \Phi[(\boldsymbol{X}_i^T\boldsymbol{\beta} - \tau)/\sigma] \text{ where we let}$$

$$Z_i = \boldsymbol{X}_i^T\boldsymbol{\beta} - \epsilon_i \text{ where } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\text{so } Y_i = \begin{cases} 1 \text{ if } Z_i \geq \tau \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Complementary log log: } \log[-\log(1-\pi(x))] = \boldsymbol{X}_i\boldsymbol{\beta}$$

The interpretation of the coefficients in a logistic regression model are quite simple. The β coefficients are the odds-ratio between that and the baseline conditioning on the values of other predictors staying fixed. The interpretation of the coefficients for a probit model is generally associated with changes in E(Y^*), which is the expectation of the latent variable. So the β represents the increase in the expectation of the latent variable associated with an increase of 1 in the covariate. The interpretation of the complementary log log model is a little bit more complex.

In choosing our final model, we use analysis of deviance to test which model is better. We can do this as long as the models are nested. We first check basic models without interactions. We find that although it is better to use age as a factor, for the interpretability of coefficients and interactions, we choose to turn it into a continuous random variable by choosing the middle values of the window. We also compare the continuous age with the full model with a second degree polynomial added for age. This is clearly better than the model without the polynomial. We then test out all possible interactions and choose the ones that seem significant and also intuitive. We choose the following interactions: AGE_G:HXSMOKE, RACE_G:HXSMOKE, RACE_G:HXCHF, HXSMOKE:LVEF_R, HXCHF:HXMI, HXDIAB:NUMDZV, HXMI:LVEF_R, HXMI:NUMDZV, and LVEF_R:NUMDZV. Each of these were significant under each of the link functions, and also many of them are make intuitive sense according to our current understanding of cardiovascular health.

Results

We now report both the finalized coefficients and certain measures of goodness of fit for this binary regression. To assess the model fit, we use ROC curves to find AUC, k-fold cross-validation and pseudo-R^2. After analyzing the these measures, it seems clear that these different link functions produces extremely similar models. For this reason, we just report the coefficients, standard error, and confidence intervals for the coefficients under the generalized linear model with the logit link function in the following table.

	logit.est	logit.se	lower	upper	odds
(Intercept)	-2.21	0.14	-2.49	-1.94	1.096000e-01
RACE_G2	-0.24	0.10	-0.43	-0.06	7.847000e-01
RACE_G3	-0.22	0.16	-0.54	0.07	8.017000e-01
HXSMOKE1	0.07	0.05	-0.03	0.16	1.067500e+00
CHFSEV1	1.89	0.74	0.65	3.72	6.591200e+00
CHFSEV2	1.74	0.74	0.52	3.57	5.674200e+00
CHFSEV3	1.81	0.74	0.58	3.64	6.089800e+00
CHFSEV4	1.56	0.76	0.28	3.42	4.751600e+00
HXANGINA1	0.25	0.13	-0.01	0.52	1.283200e+00
HXCEREB1	0.15	0.07	0.02	0.28	1.163900e+00
HXCHF1	-1.85	0.72	-3.66	-0.68	1.576000e-01
HXCOPD1	-0.04	0.10	-0.24	0.16	9.645000e-01
HXDIAB1	0.31	0.08	0.16	0.47	1.369000e+00
HXHTN1	0.03	0.04	-0.05	0.11	1.032100e+00
HXHYL1	-0.15	0.04	-0.22	-0.07	8.647000e-01
HXMI1	0.50	0.14	0.24	0.78	1.654000e+00
NUMDZV2	-0.08	0.15	-0.36	0.21	9.273000e-01
NUMDZV3	-0.32	0.15	-0.63	-0.02	7.255000e-01
poly(AGE_G, 2)1	-33.26	5.06	-43.13	-23.30	0.000000e+00
poly(AGE_G, 2)2	-14.22	4.92	-24.04	-4.74	0.000000e+00
poly(LVEF_R, 2)1	-23.74	8.77	-40.75	-6.36	0.000000e+00
poly(LVEF_R, 2)2	-16.49	8.25	-32.93	-0.57	0.000000e+00
HXSMOKE1:poly(AGE_G, 2)1	-24.29	6.66	-37.42	-11.30	0.000000e+00
HXSMOKE1:poly(AGE_G, 2)2	3.00	6.38	-9.44	15.57	2.017970e+01
RACE_G2:HXSMOKE1	0.24	0.12	0.02	0.47	1.273900e+00
RACE_G3:HXSMOKE1	0.36	0.19	-0.01	0.74	1.436400e+00
RACE_G2:HXCHF1	0.10	0.17	-0.24	0.43	1.106600e+00
RACE_G3:HXCHF1	0.66	0.26	0.13	1.17	1.940900e+00
HXSMOKE1:poly(LVEF_R, 2)1	18.24	6.47	5.53	30.89	8.352826e+07
HXSMOKE1:poly(LVEF_R, 2)2	1.24	6.70	-11.82	14.47	3.469300e+00
CHFSEV1:HXMI1	-1.85	0.84	-3.81	-0.37	1.568000e-01
CHFSEV2:HXMI1	-1.32	0.83	-3.27	0.15	2.667000e-01
CHFSEV3:HXMI1	-1.37	0.84	-3.33	0.11	2.534000e-01
CHFSEV4:HXMI1	-1.41	0.86	-3.40	0.14	2.446000e-01
HXANGINA1:HXMI1	-0.40	0.14	-0.67	-0.14	6.698000e-01
HXANGINA1:NUMDZV2	0.25	0.13	0.00	0.51	1.286000e+00
HXANGINA1:NUMDZV3	0.40	0.14	0.13	0.68	1.489600e+00
HXCHF1:HXMI1	1.32	0.80	-0.09	3.23	3.737100e+00
HXDIAB1:NUMDZV2	-0.13	0.11	-0.35	0.09	8.822000e-01
HXDIAB1:NUMDZV3	-0.23	0.11	-0.44	-0.02	7.954000e-01
HXMI1:poly(LVEF_R, 2)1	15.49	7.37	0.90	29.81	5.340473e+06
HXMI1:poly(LVEF_R, 2)2	-4.09	7.00	-17.74	9.72	1.680000e-02
HXMI1:NUMDZV2	0.01	0.10	-0.19	0.21	1.009200e+00
HXMI1:NUMDZV3	0.20	0.10	0.01	0.40	1.227000e+00
NUMDZV2:poly(LVEF_R, 2)1	18.41	8.45	1.82	34.97	9.850604e+07

	logit.est	logit.se	lower	upper	odds
NUMDZV3:poly(LVEF_R, 2)1	32.24	7.99	16.51	47.85	1.006371e+14
NUMDZV2:poly(LVEF_R, 2)2	9.73	8.44	-6.82	26.30	1.687607e+04
NUMDZV3:poly(LVEF_R, 2)2	5.28	7.95	-10.25	20.92	1.958387e+02

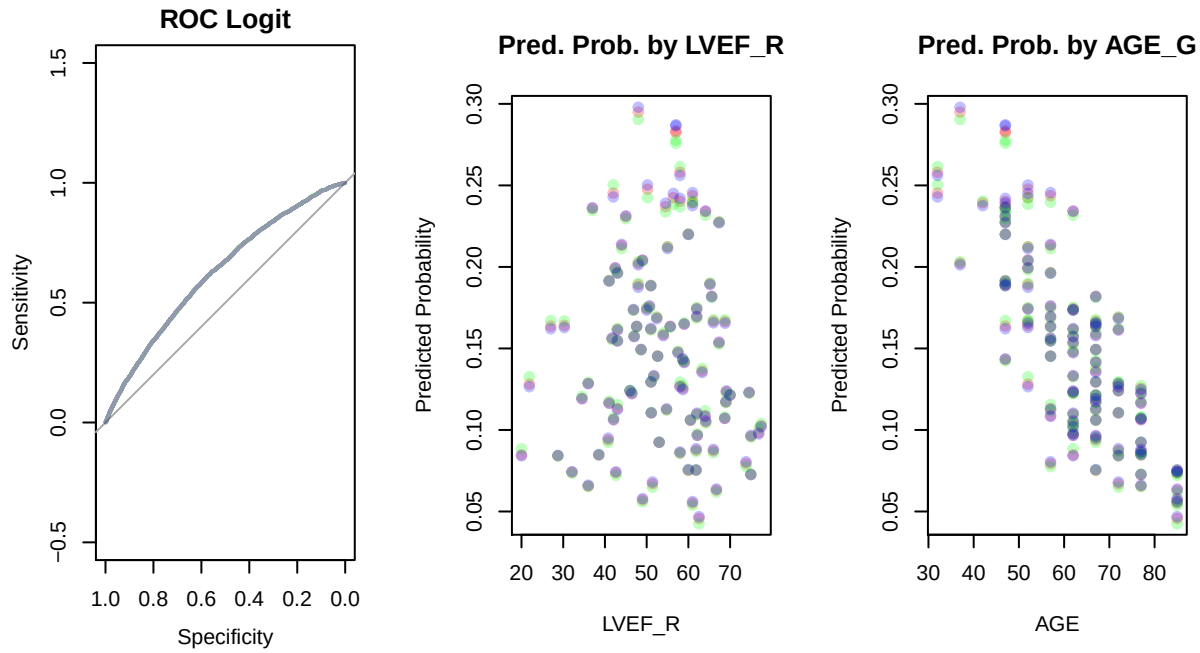
If we analyze this table, it is clear that many of these predictors we have chose are significant under the Wald-test. The e^{β_0}, which is the exponentiated intercept, as the odds of our baseline individual, which is someone who doesn't exhibit any of the characteristics described by our categorical variables. In our case, the baseline probability of MI is the following: $\frac{\exp\{-2.21\}}{1+\exp\{-2.24\}} = 0.219$, which is quite high considering the prevalence of the of MI in our working dataset is 14.8%. Among other reasons, I think the reason this is happening is because the baseline for our design would be someone aged zero, and since we are using a quadratic for age, this would be high. This also demonstrates a flaw in using age as a continuous covariate. There are only 11 observations of individuals in the lowest age division, which indicates that there is fairly low prevalence of cardiac catheterization in younger individuals.

Due to the design of our model, some β coefficients are quite easy to interpret, but others are not. For example, the coefficient for history of cerebrovascular disease is significant but is not seen in any interaction terms. Thus we can say that holding all others constant, $e^{\beta_{\text{HXCEREB}}}$ represents the odds ratio between someone with history of cerebrovasucular disease and without. The effect of other variables is more difficult to interpret due to polynomials and interaction terms. For example, we show how to interpret the effect of one year increase in age according to our model.

$$
\begin{aligned}
&\frac{\exp\{\beta_{\text{AGE}}(x+1) + \beta_{\text{AGE}^2}(x+1)^2 + \alpha_{\text{AGE:HXSMOKE}(x+1)(x_{\text{smoke}})}\}}{\exp\{\beta_{\text{AGE}}x + \beta_{\text{AGE}^2}x^2 + \alpha_{\text{AGE:HXSMOKE}(x)(x_{\text{smoke}})}\}} \\
&= \frac{e^{\beta_{\text{AGE}}x} e^{\beta_{\text{AGE}}} e^{\beta\text{AGE}^2 x^2} e^{\beta\text{AGE}^2 2x} e^{\beta\text{AGE}^2} e^{\alpha_{\text{AGE:HXSMOKE}}(x)(x_{\text{smoke}})} e^{\alpha_{\text{AGE:HXSMOKE}}(x_{\text{smoke}})}}{e^{\beta_{\text{AGE}}x} e^{\beta\text{AGE}^2 x^2} e^{\alpha_{\text{AGE:HXSMOKE}(x)(x_{\text{smoke}})}}} \\
&= e^{\beta_{\text{AGE}}} e^{\beta\text{AGE}^2 2x} e^{\beta\text{AGE}^2} e^{\alpha_{\text{AGE:HXSMOKE}}(x_{\text{smoke}})} \\
&= e^{\beta_{\text{AGE}} + \beta\text{AGE}^2 2x + \beta\text{AGE}^2 \alpha_{\text{AGE:HXSMOKE}}(x_{\text{smoke}})}
\end{aligned}
$$

Thus, the interpretation of a one unit increase in age is not nearly as simple as the coefficient for history of cerebrovascular disease because it not only depends on the current value of age, but it also depends on the value the smoking status of the individual.

To assess whether this model is actually a good fit of the data, we analysis ROC curves to find AUC and we also find pseudo-R^2. To begin, we analyze the Receiver Operating Characteristic (ROC) curve for this. This curve is a plot of the sensitivity across values of π_0 (sensitivity is described as $Pr(\hat{y} = 1|y = 1)$). In our case, the optimal $\pi_0 = 0.1505$, and we find that the area under the curve is 0.6252, which shows the concordance index. The AUC for this model is not good. The following plot on the left shows our ROC curve compared with a 1-1 line. Ideally, in this plot, we would like the OC curve to be near 1 for sensitivity and 1 for specificity, but it is clear that is not achieved in this plot. When we find the pseudo-R^2, which is defined as $1 - \frac{L(\hat{\mu};y)}{L(\hat{\mu}_0;y)} = 0.0296$. This is also very bad. Both measures indicate that this model may not be very well suited for these data. These measures were very similar for both the probit and complementary log log link functions. The following plots on the right show probability of MI predictions for 100 randomly selected individuals from the dataset. They show the predictions for each of the log-link functions with both age and LVEF_R on the x-axis. It is clear analyzing these plots that the link-functions produce quite similar results, especially near where the bulk of the data are where the predictions are nearly perfectly overlapped.



A goal in this analysis is to understand how well this model can predict MI. To do this we perform k-fold cross-validation. Although we have an extremely large dataset and could probably use a more folds, we choose to use 10-fold cross validation, which is where we split the data into 10 mutually exclusive subsets. We then hold out each of the folds and fit the model on the remaining folds and validate the results based on observed values in the folds. We can test accuracy in this way, and we find that our accuracy is 0.85. Although this sounds high, it is quite poor because the prevalence of MI in the dataset is approximately 0.15. Cohen's κ normalizes accuracy by a baseline of random chance/expectation. In other words, it measures how well we do beyond that expected by chance alone. According to our cross-validation, our $\kappa = 0$, which is incredibly bad. This means that essentially our predictions are not any better than using random chance and the null model of prevalence of MI. These results validating the model show some interesting features about fitting a model to this data.

Conclusions

We used analysis of deviance to understand which of the models is most appropriate. We decided to treat age as a continuous variable and add a second-order polynomial term also. We also added a second-order polynomial term in this. Treating this variables as such showed improvement in the amount of deviance explained. We included the following interactions in our model that were significant under a Wald-test: AGE_G:HXSMOKE, RACE_G:HXSMOKE, RACE_G:HXCHF, HXSMOKE:LVEF_R, HXCHF:HXMI, HXDIAB:NUMDZV, HXMI:LVEF_R, HXMI:NUMDZV, and LVEF_R:NUMDZV. Although some of the coefficients were different under the three different link functions we used, the results of the models were very similar. Thus, we decided to use the log(odds) link function.

Although this model outperformed other models we tested it against using analysis of deviance, there are very clear limitations. The AUC of our ROC curve is not very high, which indicates that our concordance index is not good. Also, the pseudo-R^2 describes that our model is only describing 0.0296 of the deviance in this data. Finally, the results from cross-validation describe our predictive ability like the null model. This lack of predictive ability might be because MI is hard to predict. Even understanding certain key factors about an individual, we are predictive an even in a human. Humans are complex organisms, and it can be very difficult to predict events involving them, even a medical one. Thus, even though we might have the leading predictors of MI, the act of prediction is still extremely difficult.

However, even though this model is not good, there are things that we can draw from this analysis. There

are explanatory variables described in this dataset that seem to influence the response. Race, severity of congestive heart failure, history of congestive heart failure, history of MI, history of diabetes, age, and lvef_r seem to have some effect on the probability of someone experiencing MI. We saw this throughout all the models we tested, but the extent of which might be difficult to quantify for certain since the model we use doesn't seem appropriate for the data. There are things that we could have considered or changed. We could have used a General Additive Models since we experienced non-linearities in both LVEF_R and age. We also could have treated the age variable differently and maybe done some hierarchical structure with age since the first age group was so much different than the others. We also could have handled the missing data differently, which might have changed our results since we discarded so much data. Although we did explore the influence and residual diagnostics, we did not display them in this code or report. We could have explored these ideas more because there might have been extremely influential or anomalous cases, but we chose not to explore them as deeply simply because of the size of the data. We justified that extreme cases may not influence as much when we have approximately 22,000 observations