Analysis on UCLA Stress Echocardiography Data

Ni Erchang

July 14, 2023

Background

Due to the difficulty of elevating heart rate through exercise in elderly patients, the use of dobutamine stress echocardiography(DSE) for risk stratification has become increasingly common. This research report mainly builds upon Janine Krivokapic's remarkable work[1]. The purpose of the study is to use a drug called "dobutamine" to induce stress on the heart before taking relevant measurements, and to identify which measurements are most helpful in predicting cardiac events.

To achieve the objectives of the study, we employ four variables in our study to imply the "cardiac events", they are defined as follows:

- myocardial infarction (MI)
- revascularization by percutaneous transluminal coronary angioplasty (PTCA)
- coronary artery bypass grafting surgery (CABG)
- cardiac death

Exploratory Data Analysis

The dataset consists of the medical records of 558 consecutive patients who underwent DSE, and includes information about the cardiac events that took place within 12 months following the procedure. The average age of the patients was 68 ± 12 years, with 338 women and 220 men. More detailed description of these 558 patients is given in Janine Krivokapic's work[1].

Independent variables are history of hypertension(hxofht), history of diabetes mellitus(hxofdm); history of heart attack(hxofmi); history of angioplasty(hxofptca); history of bypass surgery(hxofcabg); age; gender; peak dose of dobutamine(dose); rest and peak dobutamine heart rate(bhr and pkhr); rest and peak blood pressure(basebp and sbp); rest and peak double product(basedp and dp); maximum heart rate(maxhr); percent of achieved maximum predicted heart rate(%mphr(b)); maximum blood pressure(mbp); double product on maximum dobutamine dose(dpmaxdo); dobutamine dose at which maximum double product occured(dobdose); rest and peak dobutamine ejection fraction(baseef and dobef); presence of induced chest pain(chestpain); negative, equivocal or ischemic electrocardiogram(ecg); rest wall-motion abnormality(restwma) and positive stress echocardiogram(posse). The dependent variable in our study is the accurance of any of the heart events(any.event).

Characteristics	N	(%)	With Heart Event		Without Heart Event	
			N	(%)	N	(%)
Men	220	39	43	20	117	80
Women	338	61	46	14	292	86
HxCABG	88	16	20	23	68	77
HxPTCA	41	7	9	22	32	78
HxMI	154	28	41	27	113	73
HxHT	393	70	73	19	320	81
HxDM	206	37	44	21	162	79
Age>Avg	302	63	53	18	249	82
Age<Avg	256	34	36	14	220	86
bp>Avg	259	46	32	12	227	88
bp<Avg	299	54	57	19	242	81
posSE	136	24	46	34	90	66
ecgMI	71	13	23	32	48	68
CigHeavy	122	22	24	20	98	80
restwma	257	46	15	6	242	94
dobEF>Avg	298	53	31	10	267	90

Table 1. Characteristics of Patients With and Without a Cardiac Event

89 of the 558 patients observed heart events in the following 12 months, with an event rate of 16%. Table 1 shows A preliminary analysis of patients with and without a heart event.

First of all, in order to understand the linear relationship between variables, we generate Pearson correlation coefficients between each pair of variables. If we set the correlation coefficient from 0.60 to 0.79 as a strong correlation, and 0.80 to 1.00 as a very strong correlation, we can identify those variable pairs that have very strong relationships are pkhr and maxhr (0.954), pkhr and pctMphr (0.866), sbp and mbp (0.898), dp and dpmaxdo (0.944), maxhr and pctMphr (0.911) and baseEF and dobEF (0.900).

The original dataset contains 31 variables. We consider *any.event* as the target variable we want to predict. We will not include *newMI*, *newPTCA*, *newCABG*, and *death* in our analysis, as they are the same as the outcome variable *any.event*. Including these variables in the analysis can cause issues with multicollinearity and overfitting, and make it difficult to interpret the result.

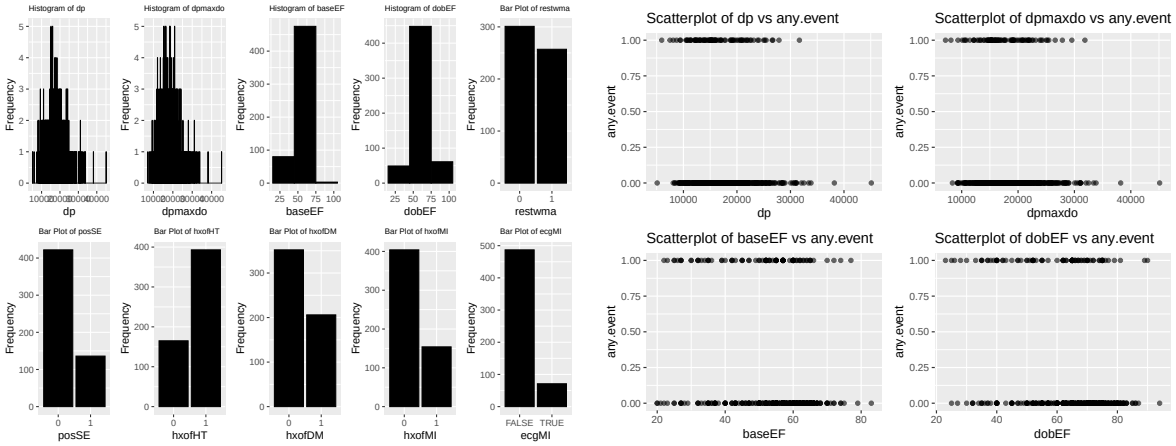
In order to reduce the huge amount of predictor variables, we first perform univariable logistic regression to determine the most important ones. We consider predictor variables with $Pr < 0.05$ as important and keep them for future model fitting. According to the coefficients obtained from univariable logistic regression in Table 2, the following predictor variables can be considered as significant predictors.

Figure 1(a) shows the dataset is well-structured, dp and dpmaxdo roughly obey normal dis-

Variable	p-value
dp	2.68×10^{-2}
dpmaxdo	3.93×10^{-2}
baseEF	2.86×10^{-7}
dobEF	2.11×10^{-9}
restwma	2.50×10^{-8}
posSE	5.29×10^{-10}
hxofHT	1.02×10^{-2}
hxofDM	8.17×10^{-3}
hxofMI	3.32×10^{-5}
ecgMI	1.21×10^{-2}

Table 2. Significant Predictor Variables With $Pr < 0.05$

tribution with a few outliers. Meanwhile, the scatter plot in Figure 1(b) shows the distribution of *dpmaxdo* and *dpmaxdo* values appears to be skewed towards lower values for 'any.event' equals 1, indicate potential negative relationship between them and heart evens.



(a) univariate Graphical Visualisation (b) Bivariate Graphical Visualisation

Figure 1. Univariate & Bivariate Graphical Visualisation

In conclusion, we identify 10 significant predictor variables, namely *dp*, *dpmaxdo*, *baseEF*, *dobEF*, *restwma*, *posSE*, *hxofHT*, *hxofDM*, *hxofMI*, and *ecgMI*. Where *ecgMI* represents a specific category 'MI' of the *ecg* variable. These variables will be used for future model fitting and analysis. Moreover, our EDA has shown that some of these significant variables we generate are strongly correlated, such as *dp* and *dpmaxdo*, and *baseEF* and *dobEF*. We should consider the potential multicollinearity issue when fitting the model in the next model fitting stage.

Model Fitting

To begin with, we conduct a multivariable logistic regression analysis using significant predictors identified from the above Exploratory Data Analysis stage to investigate the relationship between the variables and the likelihood of experiencing an event (namely *any.event*). The significant predictors we choose are dp, dobEF, restwma, posSE, hxofHT, hxofDM, hxofMI, and ecgMI, where ecgMI indicates the categorical variable ecg's state. According to Work[1], we can discover that patients with a positive SE and an ischemic ECG had a 42% event rate, and patients with a rest WMA, positive ECG for ischemia, and positive SE for ischemia had a cardiac event rate of 41%, which are both significantly higher than other combinations. Thus it's reasonable to introduce interaction terms *restwma:ecgMI* and *restwma:ecgMI:posSE*.

$$\begin{aligned} HeartEvent = {} & \beta_0 + \beta_1 * dp + \beta_2 * dobEF + \beta_3 * restwma + \beta_4 * posSE \\ & + \beta_5 * hxofHT + \beta_6 * hxofDM + \beta_7 * hxofMI + \beta_8 * ecgMI \\ & + \beta_9 * restwma : ecgMI + b10 * restwma : ecgMI : posSE + e \end{aligned}$$

Here, we choose R-square and AIC to help us determine which arguments to use in our model. A higher *R-squared* value indicates a more significant percentage of a dependent variable is explained by the independent variables in a regression model, this number is commonly stated between 0 and 1[2]. And the *Akaike Information Criterion (AIC)* is a model selection criterion with the goal to find the model that maximizes the empirical likelihood while penalizing excessive complexity [3].

However, the inclusion of all the above predictor variables and interaction terms produces a relatively low R-squared value of 0.17, the full model's residual deviance is 399.73 with 545 degrees of freedom, and the AIC is 425.73, there might exist a better model to improve the performance of our model.

To conduct further variable selection and automatically reduce the number of predictors, we apply a *stepwise variable selection* method based on the AIC criterion, by adding or removing predictor variables in both directions based on their statistical significance [4]. We can obtain the following model:

$$\begin{aligned} HeartEvent = {} & \beta_0 + \beta_1 * dp + \beta_2 * dobEF + \beta_3 * restwma + \beta_4 * posSE \\ & + \beta_5 * hxofHT + \beta_6 * hxofDM + \beta_7 * hxofMI + \beta_8 * ecgMI + e \end{aligned}$$

with an *AIC* of 419.7, which shows an improvement from the previous model, thus means we should focus on *dp*, *dobEF*, *restwma*, *posSE*, *hxofHT*, *hxofDM*, *hxofMI*, and *ecgMI* as predictors. The interaction terms are insignificant in this model, and their inclusion does not substantially improve the fit.

Table 3 shows the coefficients, standard errors, z-values, and p-values for each predictor variable in the stepwise logistic regression model.

Several predictor variables are statistically significant. Specifically, higher values of *restwma* (OR = 0.50, $p = 0.049$), higher values of *posSE* (OR = 2.59, $p < 0.001$), the presence of *hxofHT* (OR = 2.16, $p = 0.017$), and the presence of *ecgMI* (OR = 2.21, $p = 0.014$) are associated with an increased likelihood of experiencing heart events. While higher values of *dobEF* (OR = 0.97, $p = 0.005$) are associated with a decreased likelihood of heart events.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.21	0.84	-0.25	0.805
dp	-0.00004	0.00003	-1.63	0.104
dobEF	-0.03	0.01	-2.79	0.005**
restwma	-0.70	0.36	-1.97	0.049*
posSE	0.95	0.28	3.42	0.001 ***
hxofHT	0.77	0.33	2.38	0.017 *
hxofDM	0.43	0.26	1.65	0.098
hxofMI	0.45	0.28	1.62	0.104
ecgMI	0.80	0.32	2.46	0.014 *

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3. Logistic Regression Results for Predicting any.event (Stepwise Model)

According to the coefficients presented, we generate our final logistic regression formula:

$$P(HeartEvent = 1) = \frac{1}{1 + e^{-(logit)}} \tag{1}$$

$$\begin{aligned} \hat{logit} = &-0.21 - 0.00004 \cdot dp - 0.03 \cdot dobEF - 0.70 \cdot restwma + 0.95 \cdot posSE \\ &+ 0.77 \cdot hxofHT + 0.43 \cdot hxofDM + 0.45 \cdot hxofMI + 0.80 \cdot ecgMI \end{aligned} \tag{2}$$

Model Assessment

Logistic Regression Assumptions Diagnostics

In the logistic regression we hold several assumptions[5], some of them are:

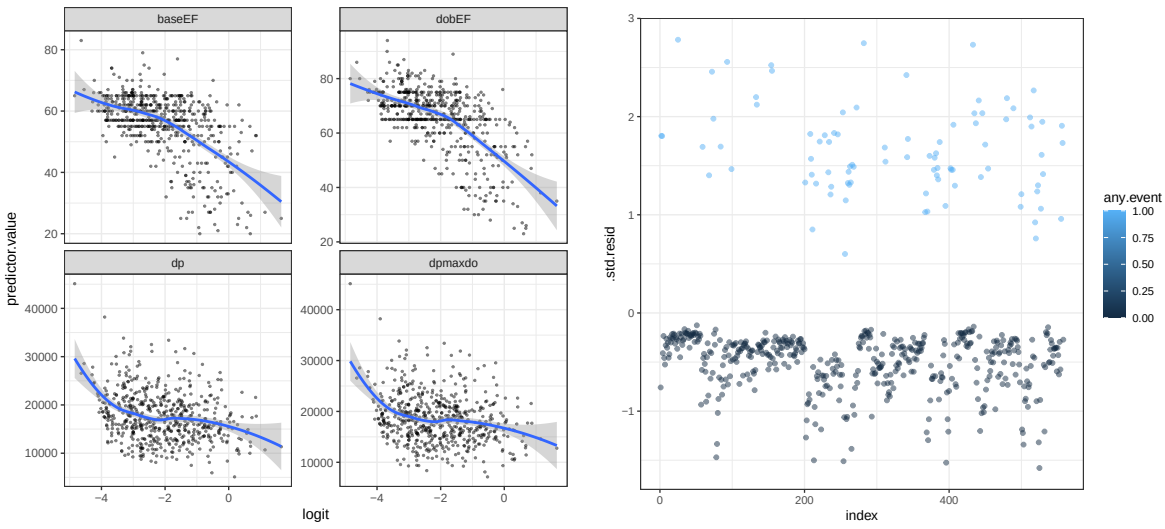
- Binary outcome
- Independence of Observations
- Linear relationship between outcome's logit and predictors

- No influential values or outliers
- No high multicollinearity among predictors

First of all, it's quite straightforward to confirm the binary outcome: HeartEvent is binary because it only contains two situations: happen or not. As we analyzed in EDA stage, the data is from a study conducted by UCLA Department of Physiology and the dataset consist of 558 individual patients records, so the assumption of independence is automatically met.

To prove the linearity assumption, we visually inspect the scatter plot of logit versus predictor in Figure 2(a), to inspect the linear relationship between our logit of the outcome and those continuous predictor variables, here we remove qualitative variables in the assessment. The outcome showes us relatively smoothed scatter plots, we could say that baseEF, dobEF, dp and dpmaxdo are all quite linearly associated with the heart event in the logit scale.

To test if there exist any extreme individual data, we extract the model results, compute the standardized residuals and the Cook's distance. A variable with a larger Cook's distance will present a stronger influence on the model. We find the top 3 observations with the highest Cook's distance values and create a scatter plot Figure 2(b). The points with absolute standardized residuals greater than 3 will be highlighted, and there is no potential outlier according to our scatter plot.



(a) Scatterplots of Logit vs. Predictors (b) Standardized Residuals of the Model

Figure 2. Logistic Regression Assumptions Diagnostics

Finally, to prove there is no high multicollinearity among predictors, we calculate the variance inflation factors as Table 4 shows. We take the *variance inflation factor (VIF)* values exceeding 5 as a suggestion of collinearity, VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity [6], and our VIF results indicate that there is no high collinearity.

dp	dobEF	restwma	posSE	hxofHT	hxofDM	hxofMI	ecgMI
1.05	1.24	1.35	1.17	1.04	1.04	1.14	1.06

Table 4. The Variance Inflation Factors of The Varibles

Model Fit and Accuracy

To evaluate the performance of our model, we first choose the *Hosmer-Lemeshow* test to evaluate our model fitting, by calculating if the observed event rates match the expected event rates in population subgroups [7]. Our null hypothesis is that the model fits the data well, our test results show a p-value of 0.219, which is greater than the significance level of 0.05, thus according to the Hosmer-Lemeshow test we fail to reject the null hypothesis, suggesting that there is no evidence of lack of fit for our model.

After that, we construct a *receiver operating characteristic (ROC) curve* for our model, the ROC curve is the plot of the true positive rate against the false positive rate, at various threshold settings. Figure 3 shows the ROC curve. We then calculate the area under the curve (AUC) to measure our model's predictive accuracy. With an AUC of 0.798, our model showes a good ability to predict the occurrence of heart events.

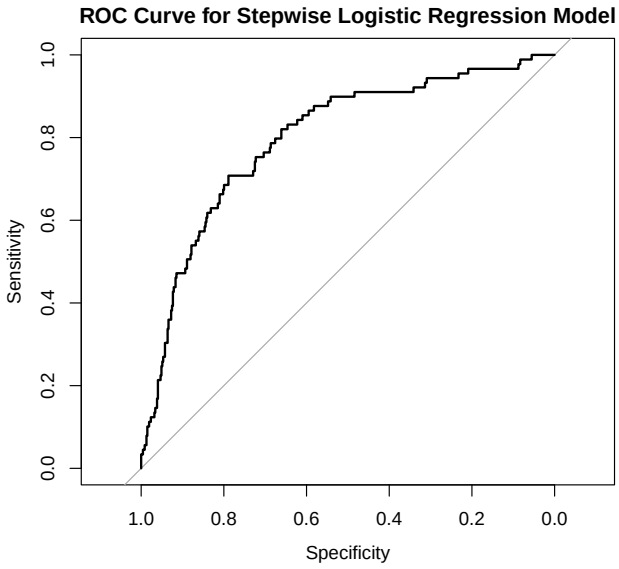


Figure 3. ROC Curve for Stepwise Logistic Regression Model

Conclusions

In conclusion, we develop a logistic regression model to predict the likelihood of heart events and to distinguish the most predictive factors. We first conduct single-variable logistic regression

to select relatively significant predict variables, which are *dp*, *dpmaxdo*, *baseEF*, *dobEF*, *restwma*, *posSE*, *hxofHT*, *hxofDM*, *hxofMI*, and *ecgMI*. To refine our model, we use stepwise variable selection to determine the most significant factors among them and generate the final model.

In order to perform the model assessment, we first draw scatter plots of logit vs. predictors and visually ensure there is a linear relationship between the outcome's logit and predictors. Then, we utilize Cook's distance and prove there are no outliers. Moreover, we use the VIF value to test for the absence of high collinearity. Finally, both the Hosmer-Lemeshow test and AUC value show that our model has a great ability to predict.

Reference

[1] A. Garfinkel, V. F. Froelicher, K. M. Kent, and B. G. Pollock, "Prognostic value of dobutamine stress echocardiography in predicting cardiac events in patients with known or suspected coronary artery disease," *Journal of the American College of Cardiology*, vol. 33, no. 3, pp. 708–716, 1999.

[2] J. Fernando, *R-squared: Definition, calculation formula, uses, and limitations*, Investopedia, Accessed: 05 July 2023, 2023. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp#toc-what-r-squared-can-tell-you.

[3] J. E. Cavanaugh and A. A. Neath, "The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Computational Statistics*, vol. 3, no. 11, 2019. DOI: 10.1002/wics.1460.

[4] A. Kassambara, *Stepwise logistic regression essentials in r*, STHDA, Accessed: 05 July 2023, 2018. [Online]. Available: http://www.sthda.com/english/articles/36-classification-methods-essentials/150-stepwise-logistic-regression-essentials-in-r/.

[5] A. Kassambara, *Logistic regression assumptions and diagnostics in r*, STHDA, Accessed: May 3, 2023, 2018. [Online]. Available: http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/#linearity-assumption.

[6] T. I. Team, *Variance inflation factor (vif)*, May 2023. [Online]. Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asp.

[7] S. Glen, *Hosmer-lemeshow test: Definition*, Dec. 2020. [Online]. Available: https://www.statisticshowto.com/hosmer-lemeshow-test/.