Article

Malebogo Pulenyane and Tlhalitshi Volition Montshiwa*

A Regression Model for Predicting the Likelihood of Reporting a Crime Based on the Victim's Demographic Variables and Their Perceptions Towards the Police

https://doi.org/10.1515/spp-2020-0003
Received March 16, 2020; accepted September 24, 2020

Abstract: Despite the growing criminal activities in South Africa, many victims still do not report the crimes, therefore there was a need to understand the determinants of the likelihood of reporting a crime in the country. Binary logistic regression is a supervised machine learning algorithm that can assist in predicting the likelihood of reporting a crime but the selection of relevant variables to add in the model varies from one author to the other. Selection of theoretically sound and statistically relevant independent variables is key to achieving parsimonious multivariate models. This study sought to test the efficiency of some commonly used variable selection methods for logistic regression models in order to identify the most relevant determinants of the likelihood of reporting a crime of housebreaking. The study used 17 candidate variables such as the victims' demographic variables and their perceptions on the police. The multivariate model fitted using stepwise selection was found to be a best fit for the data based on the lowest AIC, the highest classification accuracy rate and the highest Area under the Receiver Operating Characteristic curve. The model fitted using the Hosmer-Lemeshow (H-L) algorithm was the worst fit for the data. The study revealed a limitation of the stepwise selection method which is that this method may select different independent variables for each unique set of randomly selected observations of the same dataset. The study established a multivariate logistic regression model to predict the likelihood of a victim reporting a crime of housebreaking and the determinants thereof.

***Corresponding author: Tlhalitshi Volition Montshiwa**, PhD, Department of Business Statistics & Operations Research, North West University, Mahikeng, South Africa,
E-mail: volition.montshiwa@nwu.ac.za, https://orcid.org/0000-0003-3168-3441
Malebogo Pulenyane, Department of Business Statistics & Operations Research, North West University, Mahikeng, South Africa, E-mail: mpulenyane@gmail.com, https://orcid.org/0000-0002-0196-8799

Key words: crime, housebreaking, logistic regression, reporting

1 Introduction

With the rise in criminal activities in South Africa, reporting crimes is imperative but the dataset considered in this study shows evidence that there are still many unreported cases of housebreaking. As such, there was a need to understand the variables which influence the victims not to report crimes, hence the interest in the likelihood of reporting a crime in this study. According to Statistics South Africa (Stats SA)'s Victims of Crime Survey (VOCS) 2015/16 released on 14 February 2016, the reason provided by most households as to why they made the choice not to report the crime was that they believed the police could not or would not do anything. However, this perception of victims on the competency of the police is not the only variable that is available in the VOCS of 2015/2016. As such, this study generally sought to utilise both the victims' demographic variables and the different variables describing the victims' perceptions on the police to determine how such variables influence the victim's likelihood of reporting a crime of housebreaking in South Africa.

The aim was to seek for a model that can best predict the likelihood of reporting a crime or that can best explain the variation in the likelihood of reporting a crime but using only the most relevant variables from a pool of variables in the dataset, that is, a parsimonious model. Statistical modelling seeks to fit a multivariate model that can produce results that are representative of the data and this is achieved by only retaining significant variables in the final model (Hosmer and Lemeshow 2004). The correct selection of variables is fundamental when fitting logistic regression models (LRMs). It aids in identifying the variables that are theoretically and empirically related to the dependent variable and assists researchers in achieving parsimonious models, that is, models which explain a lot of variation with few variables. Purposeful selection of variables can also help to avoid over-determined models, that is, models having more variables than observations.

There are several methods that have been used by previous studies in selecting appropriate variables for multivariate logistic regression models. Hacke et al. (2004) used Spearman's rank correlation to test for the correlation between the dependent variable and the other continuous independent variables and they also used likelihood-ratio test (LRT) to check the relationship between the dependent variable and the categorical independent variables. Similar to the study by Hacke et al. (2004), Larkin et al. (2010) used the chi-square test of association to check the relationship between categorical or binary independent variables and the dependent variable but for continuous independent variables, the authors implemented the Mann-Whitney's *U*-test and *t*-tests. The independent variables showing significant correlations and associations with the dependent variable were then included in the multivariate model.

Shervin et al. (2009) implemented analysis of variance (ANOVA) to test the relationship between continuous independent variables and the dependent variable whereas the chi-square test of association was used to determine the association between categorical independent variables and the dependent variable. Bivariate Cox proportional hazards regression models were initially built using each candidate variable and the dependent variable. All variables considered theoretically sound as well as the variables found to be significant in each of the Cox regression models were included in the multivariate model.

Some studies such as the one conducted by North et al. (2011) implemented a combination of approaches to select variables for their multivariate model. These authors implemented the chi-square test of association as a pre-selection method to identify all the categorical independent variables with a bivariate association with the dependent variable. Moisey et al. (2013) used a similar approach but they also used the Mann-Whitney *U*-test to test the significance of the continuous variables and the Fisher's test (depending on the expected frequencies) to confirm the significance of the relationship between categorical independent variables and the dependent variable. Variables found to be significantly associated with the dependent variable were included in the multivariate model and stepwise regression was then used to select the final variables. Chen et al. (2009) used a similar approach but due to the nature of their data, they also used the *t*-test to compare the continuous independent variables to the dependent variable.

In another study which used a combination of variable selection methods, Austin and Tu (2004) used the chi-square test of association to determine the statistical significance of the association between the categorical independent variables and the dependent variable and they used bivariate logistic regression models to determine the statistical significance of the association between each of the continuous independent variable and the dependent variable. The variables showing significant bivariate relationships with the dependent variable were included in the multivariate model and backward, forward and stepwise selection methods were used in fitting the final multivariate model.

Other authors such as Bursac et al. (2008) and Bohlke et al. (2009) implemented a purposeful selection algorithm that automates the purposeful selection of variables to be included in the multivariate logistic regression model. The method they used is similar to that proposed by Hosmer and Lemeshow (2000) and is also referred to as the H-L Algorithm in this paper. Each variable was checked for statistical significance using a certain *p*-value cut-off point such as 0.25 and all the independent variables showing a reasonable association with the dependent variable were entered into the multivariate model. The stepwise selection method was then used and variables that did not meet the level of significance to stay in the model were removed.

The parameter estimates of the remaining variables were evaluated and once a significant change of say, at least 15% in an excluded variable or coefficient of other variables was observed, the variable was entered back into the model. The authors explain that this process was followed until there were no significant variables left to be added to the model and all variables included had a significant association to the dependent variable.

From the previous studies interrogated in this paper, it was noticed that purposeful variable selection is usually guided by theoretical soundness of the relationship between the variables and the dependent variable. Also, the method used also depends on the measurement scale of the variables. For example, continuous variables are selected using an approach which differs from how the categorical variables are selected. It was also noticed that most of these previous studies did not compare the efficiency of the different variable selection approaches; they were focusing more on the application of such. As such, the most appropriate approach to purposefully selecting variables for parsimonious multivariate logistic regression models remains unknown. Therefore, this paper sought to address this gap in literature. Another methodological contribution of this study is that it compares the different variable selection approaches under more than one sample so as to observe whether the results will be consistent across the samples or not, which is something that has never been done by previous studies focusing on this topic.

It was also noticed that most of the previous studies reviewed in this paper used purposeful selection of variables when fitting clinical data. Examples of such studies include the studies by Hosmer and Lemeshow (2000), Bohlke et al. (2009), Fanelli et al. (2011), Randall et al. (2013), and Folkerson et al. (2015). As such, little is known about purposeful selection of variables for data from other disciplines than clinical studies, including in crime statistics. Therefore, this paper also sought to extend the scope of previous studies by exploring purposeful variable selection of variables for a different dataset (crime statistics) which to the best of the researchers' knowledge has never been considered in previous studies on this topic. As such, this led the study to considering crime statistics as the area of interest.

2 Methods

2.1 The General Binary Logistic Model

Binary logistic regression is used in this study to fit the final multivariate model for predicting the likelihood of reporting a crime of housebreaking after using several purposeful variable selection methods to select the appropriate independent variables. The study sought to estimate parameters for the general model:

$$\log(\text{Reporting housebreaking}\,(1 = \text{yes},\ 0 = \text{no})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_j, \quad (1)$$

where β_i, $i = 0, 1, 2, \ldots, \beta_i$ are parameter estimates and X_j, $j = 0, 1, 2, \ldots, n$ are the independent variables. Mathematically, when the LRM involves only one explanatory variable, say X_1 such that X_1 takes on only two values of Y_i , so that $i = (0, 1)$ where 0 is not reported and 1 is reported, an LRM for this data would correspond to:

$$\log\left\{\frac{\pi(X_1 = 1)}{1 - \pi(X_1 = 1)}\right\} = \beta_0 + \beta_1, \quad (2)$$

where:

- X_1: Equal to 1 when it represents Reported housebreaking.
- β_0: Represents the logarithm of the odds of response for Not Reporting housebreaking.
- β_1: Represents the logarithm of the odds of response for Reporting housebreaking.

2.2 Model 1: Variables Selected Using the Stepwise Selection Method

As Bursac et al. (2008) and Olusegun, Dikko, and Gulumbe (2015) explain, stepwise selection is a combination of the forward and backward methods. The authors add that this is done such that variables are included and removed from the model in such a way that each forward selection process may be followed by one or more backward elimination steps until no other variable can be added into or removed from the model. Variables are checked for significance and entered into the model using forward selection and once entered, the variables are tested for significance and if found not to be significant, they are removed using the backward elimination process (Murtaugh 2009).

According to Olusegun, Dikko, and Gulumbe (2015), the stepwise selection process uses the F statistic for its selection. The F statistic is found by comparing the Mean Square of the Regressors (MSR) and the Mean Square of the Error (MSE) and it is defined as follows (Olusegun, Dikko, and Gulumbe 2015):

$$F^* = \frac{MSR(X_k)}{MSE(X_k)}, \quad (3)$$

where X_k is the k potential independent variables. In this study, those are Police Visibility, Police Accessibility, Police response, Service level and Satisfaction level to mention a few. The p-value is used to identify the variables to be included into or excluded from the model. The current study uses the F-test at a significance level of 0.05 as recommended by Murtaugh (2009). The model can be written as:

$$\log \pi_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i, \tag{4}$$

where π_i are the log odds of reporting housebreaking and X_i are all the independent variables, associated with reporting a crime of housebreaking.

2.3 Model 2: Variables Selected Using the Chi-Square Test of Association

The chi-square statistic is defined by:

$$\chi^2 = \sum_r \sum_c \frac{(P_{rc} - \hat{\pi}_{rc})}{\hat{\pi}_{rc}}, \tag{5}$$

where P_{rc} is the observed proportion from the data (observed count) and $\hat{\pi}_{rc}$ is the expected proportion (Heeringa, West, and Berglund 2010). The chi-square test is used to check association between the likelihood of reporting a housebreaking crime and each of the independent variables. All the pairs of variables used in these chi-square tests are categorical, therefore, the statistic used is appropriate for this measurement scale. Chi-square tests with a p-value less than the significance level of 0.05 was used in selecting variables for the multivariate model.

2.4 Model 3: Several Bivariate Regression Models Used in Identifying Significant Variables for the Multivariate Model

In this approach, the bivariate relationship between reporting a crime of housebreaking and each of the dependent variables was confirmed by performing bivariate binary logistic regression models of the form:

$$\log \pi_i = \beta_0 + \beta_1 X_1, \tag{6}$$

where: log π_i are the log odds of reporting a crime of housebreaking, β_0 is the intercept and X_1 the independent variable.

The F statistic similar to the one in Eq. (3) is used to test the significance of the overall model. If the model is found to be significant overall, the Wald chi-square test is then used to test whether the variable is also significant in the mode. The Wald test is given by (Heeringa, West, and Berglund 2010):

$$F_{wald} = Q_{wald} \times \frac{df - (R-1)(C-1)) + 1}{(R-1)(C-1)df} \sim F_{(R-1)(C-1)+1} \text{ under } H_0, \tag{7}$$

where C are the columns, R the rows and df the degrees of freedom. The statistical significance of the overall model and the variable are tested at 5% level of significance. Only when both the overall model and the variable are significant, the variable was considered for inclusion in the multivariate model.

2.5 Model 4: Variables selected using the H-L algorithm

In this model, this paper used an automated purposeful selection algorithm proposed by Hosmer and Lemeshow (2000) for building the multivariate logistic regression of the form:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}. \tag{8}$$

with a logit transformation of $\hat{\pi}(X_i)$ given as:

$$\begin{aligned} g(x) &= \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= \beta_0 + \beta_1 x \end{aligned} \tag{9}$$

To fit the model with an outcome $y = 0, 1$, the unknown parameters β_0 and β_1 should firstly be estimated. This is achieved by constructing a likelihood function defined as

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{10}$$

The obtained maximum likelihood function produces values for the estimates of the unknown parameters which maximise the probability of obtaining the observed set of data. The estimate of β_i is expressed as $\widehat{\beta}_i$ and $\hat{\pi}(X_i)$ is the maximum likelihood estimate of $\pi(X_i)$. A p-value of 0.05 is used to check statistical significance of the variables.

A univariate analysis is performed on each variable using the likelihood ratio test and the Wald statistic the variables are selected for the multivariable analysis. All the variables from the univariate analysis that reported a p-value of 0.25 are candidates for the model along with other unknown variables considered to be important pertaining to the subject matter. This step results in a model containing all the selected variables. Verification of the importance of each selected variable included in the multivariable model in step 2 is performed. This is achieved by examining the Wald statistic for each selected variable and comparing the estimated coefficient to the coefficient from the model with only the variable. A new model is fitted including only the contributing or significant variables, those that do not contribute to the model are removed. This verification process results in a *preliminary main effects model.*

The significance of the variables retained in the preliminary main effects model is checked closely to insure that they should remain in the model. Once the variable is identified as significant, the accurate parameter relationship or scale of the variables is obtained in the model refinement stage. This step results in a *main effects model.* Interactions among the variables in the main effects model are checked after insuring that each variable is scaled correctly. Interaction variables are created and included in the model one at a time and their significance is assessed using a likelihood ratio test. This results in a *preliminary final model.*

2.6 Comparison of the Models

The following comparison criteria were used in this study to compare the four models under each of the five samples:

2.6.1 The Akaike information criterion (AIC)

The model with the lowest AIC is considered to be the best model and the AIC is defined as follows:

$$AIC = -2\log L\left(\widehat{\theta}\right) + 2k, \tag{11}$$

where θ is the set (vector) of model parameters

$L\,(\widehat{\theta})$ is the likelihood of the candidate model given the data when evaluated at the maximum likelihood estimate of θ and k is the number of estimated parameters in the candidate model (Fabozzi et al. 2014; Hilbe 2011).

2.6.2 The Bayesian information criterion (BIC)

As is for AIC, the model with the lowest BIC also referred to as the Schwarz Information Criterion or Schwarz Bayesian Information criterion is considered to be the best model (Fabozzi et al. 2014). The BIC is a model selection method set on a Bayesian context but based on information theory and is computed as:

$$BIC = -2\log L(\widehat{\theta}) + k\log n, \tag{12}$$

where its terms are the same as those described for AIC above and n is the number of observations (Fabozzi et al. 2014).

2.6.3 Area under the Receiver Operating Characteristic (ROC) curve

The ROC curve originates from signal detection theory. It plots the probability of detecting a true signal and false signal for an entire range of possible cut points (Sarkar and Midi 2010). The ROC curve is used to test the accuracy of the fitted model. The model fit is measured using the area under the curve (AUC). According to Sarkar and Midi (2010) an ideal curve has an area of 1, with the worst case scenario being 0.5 and the accuracy of the ROC test is dependent on the level at which the test is able to separate the group being tested into those with or without the criteria in the model. A larger AUC indicates better predictability of the model (Park 2013). The ROC is formed by plotting the true positive rate (TPR) against the false positive rate (FPR):

$$TPR = \frac{TP}{TP + FN}, \tag{13}$$

$$FPR = \frac{FP}{FP + TN}, \tag{14}$$

or

$$FPR = (1 - \text{specificity}), \tag{15}$$

where *TP* is the true positive result, *FP* is the false positive result, *TN* true negative results and *FN* false negative (Hajian-Tilaki 2013).

2.6.4 Classification Rate

A classification table evaluates the predictive accuracy of a LRM, by cross-classifying observed responses and predicted values at a specified cut-off point (cut-off value specified by the user) (Schlotzhauer 1993). Classification using bias-adjusted predicted probabilities estimates bias caused by using all observations to fit the model since each observation will influence the model used to classify itself (classification table-analysis-model). Here observations are adjusted and classified according to the cut-off(s) specified. Park (2013) indicates that a model with a better fit is indicated by classification table resulting with higher sensitivity and specificity,

$$Sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

$$Specificity = \frac{TN}{FP + TN}, \tag{17}$$

where *TP* is the true positive result, *FP* is the false positive result also known as a Type I error, *TN* the true negative results and *FN* false negative results also known as a Type II error. The higher the classification rate, the better the model.

2.7 Evaluation of the Final Multivariate Model

The most preferred model from the five samples was applied to the whole dataset and the multivariate model founded was further evaluated using statistics AIC, BIC, AUC, Sensitivity and Specificity which are explained by Eq. (11) though to (17). In addition, the study used the following statistics to evaluate the multivariate model:

2.7.1 Wald Test for Overall Significance of the Multivariate Model

According to Park (2013) a chi-square statistic defined as:

$$\chi^2 = \sum_{i=1}^{n} r_i^2 \tag{18}$$

can be formed based on residuals $y_i - \hat{y}_i$ to test the fit of the logistic regression model, where the standardized residuals can be defined as:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1 - \hat{y}_i)}} \tag{19}$$

and the standard deviation of the residuals is following a chi-square distribution with $n - (k + 1)$ degrees of freedom to test the overall model fit. The current study uses the chi-square of the Wald test to assess the model fit by checking if any of the coefficients included in the model are not equal to zero. If all the coefficients in the model are equal to zero, then the model is considered to be insignificant.

2.7.2 The Hosmer and Lemeshow Goodness-of-Fit Test

Using the Pearson chi-square statistic calculated from the 2-by-*g* table of the observed and the expected frequencies (where *g* is the number of groups) the Hosmer and Lemeshow statistic can be obtained. The test checks the similarities between the observed proportions of the events and the predicted probabilities of occurrence in subgroups of the overall model (Park 2013; Sarkar and Midi 2010). The value of the statistic is written as:

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{\left(o_k - n'_k \bar{\pi}_k\right)^2}{n'_k \bar{\pi}_k \left(1 - \bar{\pi}_k\right)}, \tag{20}$$

where n'_k is the number of observations in the kth group, o_k is the sum of event outcomes in the kth group and $\bar{\pi}_k$ is the average estimated probability of an event outcome for the kth group (Sarkar and Midi 2010). A chi-square distribution with $(g - n)$ degrees of freedom is compared to the H-L statistic (the default value for n is 2) and a large value of the H-L statistic with a small p-value indicate a lack of fit of the model to data while a small value with a large p-value indicate a good fit (Park 2013; Sarkar and Midi 2010).

2.7.3 Wald Test for Significance of Individual Parameters

The Wald statistic given as:

$$W_j = \frac{\beta_j^2}{SE^2_{\beta j}} \tag{21}$$

is defined as the ratio of the square of the regression coefficient to the square of the standard error of the coefficient under the null hypothesis $H_0{:}\beta_i = 0$ $(i = 1, 2, \ldots, n)$ the Wald statistic is distributed as a chi-square with a one degree of freedom (Park 2013; Sarkar and Midi 2010).

2.7.4 Odds Ratios

The relationship between the coefficients and the odds ratio for a logistic regression model with a binary independent variable coded 1 and 0 is (Hosmer and Lemeshow 2000)

$$OR = e^{\beta_i}. \tag{22}$$

The odds ratio compares two odds relative to different events (odds of an event is the proportion of the probability of occurrence against non-occurrence (Park 2013)). It measures the association between an experience and an outcome and for any two events A and B the corresponding odds of A occurring relative to B occurring can be given as:

$$OR\{AvsB\} = \frac{odds\{A\}}{odds\{B\}} = \frac{P_A}{(1-P_A)} \bigg/ \frac{P_B}{(1-P_B)} \tag{23}$$

An odds ratio equal to one indicates no association, odds ratio greater than one indicates higher odds of outcome and odds ratio less than one indicate lower odds of outcome.

3 Results

3.1 Data

This study used data from the VOCS 2015/16 collected by StatsSA. VOCS is only representative of non-institutionalised and non-military persons or households in South Africa, that is, information is sought from a sample of all private households in all nine provinces of South Africa and residents in workers' hostels. The data collected provides information on the experiences and perceptions of crime of the South African households and victims of crime. The survey also focuses on the respondent's views of accessibility of police services, their response rate to crime and the criminal justice system.

This paper focused on the likelihood of reporting of a crime of housebreaking by victims based on their perception of the police as well as their demographic. The data set comprised 1061 observations. The likelihood of reporting a crime of housebreaking was used as the dependent variable (0 meant No, and 1 meant Yes) and 17 other variables were used as candidate independent variables for the multivariate model. From the dataset, 54.9% of the subject reported the crime of burglary whereas the remaining 45.1% did not. Appendix 1 shows the descriptive statistics for the variable Age (which is continuous) and Appendix 2 presents frequencies for each independent variable (all of which are categorical). These candidate variables included the Gender of the victim, Marital status of the victim, Average time to police station, Satisfaction with police and Whether the victim has visited the police station in three years to mention a few. The data was analysed using the Statistical Analysis System (SAS) version 9.4, and was prepared, cleaned and captured in the Statistical Package for Social Sciences (SPSS) version 25 prior to being analysed.

3.2 Sampling and Replication

To ensure the validity and replication of results as well as statistical rigour, the study randomly selected five samples with a variable-to-observation ratio of 1:10

from the 1061 observations. This variable-to-observation ratio is deemed sufficient for a multivariate logistic regression model by Hosmer David and Stanley (2000), Peng et al. (2002), and Menard (2002). Since there are 18 variables in this study, each sample comprised 180 observations. This sampling and replication was used in confirming whether the purposeful variable selection methods under study are able to select the same independent variables for each sample, and to confirm whether the same variable selection method will be identified as the most efficient across all the samples. These samples were treated as training datasets. The various purposeful variable section methods were used to fit multivariate models for each sample and were compared using some comparison criteria. This was done to ensure that the rightful purposeful variable selection model is chosen prior to applying it to the whole dataset in order to fit the final multivariate logistic regression model for the study.

3.3 Model Comparison and Selection

Using a ranking order in which 1 = lowest AIC and 4 = highest AIC (The lower the better), Model 1 was the most preferred model since it had the lowest AIC in Samples 2,3,4 and 5 and has the second lowest AIC for Sample 1 according to the results in Table 1. The results in Table 1 also show that the second best model was Model 3, followed by model 2 and Model 4 was the least preferred model based on the AIC.

Using a ranking order in which 1 = lowest BIC and 4 = highest BIC (The lower the better), it was found that for three samples (Samples 1, 4 and 5), Model 4 had the lowest BIC and also had the second lowest BIC for Sample 3 as shown in Table 2. As such, based on the BIC, Model 4 is the preferred model. The results in Table 2 also show that Model 3 had the second highest BIC for two samples (Samples 4 and 5) and also had the third lowest BIC for sample 1 and 3. As such, Model 3 is the second

Table 1: AIC.

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	AIC	Rank	AIC	Rank	AIC	Rank	AIC	Rank	AIC	Rank
Model 1	222.686	2	222.378	1	221.907	1	215.622	1	208.628	1
Model 2	227.642	3	228.583	3	236.781	3	222.779	3	216.409	3
Model 3	230.361	4	226.522	2	226.151	2	219.567	2	210.124	2
Model 4	221.732	1	233.423	4	237.008	4	225.371	4	222.217	4

Table 2: BIC.

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	BIC	Rank	BIC	Rank	BIC	Rank	BIC	Rank	SBI	Rank
Model 1	251.422	4	244.729	2	241.065	1	241.165	3	237.365	3
Model 2	240.414	2	247.741	4	281.482	4	251.516	4	245.146	4
Model 3	246.326	3	239.294	1	248.501	3	238.724	2	229.282	2
Model 4	237.696	1	246.195	3	243.394	2	238.142	1	228.603	1

most preferred model. In addition, Table 2 further shows that the BIC results are inconclusive for Model 1 since there is no trend in the ranking of this model whereas Model 2 is the least preferred model having the highest BIC for Samples 2, 3, 4 to 5.

Using a ranking order in which 1 = highest AUC and 4 = lowest AUC (The higher the better), Model 1 was the most preferred model since it had the highest AUC for four samples (Samples 1, 2, 4 and 5), and also has the second highest AUC for Sample 3 according to the results in Table 3. Model 2 was the second most preferred model since it had the highest AUC for Sample 3, second highest for Samples 2 and 4, and third highest AUC for Samples 1 and 5 as shown in Table3. The table also shows that relative to Model 2, Model 3 is the third most preferred model, and Model 4 had the lowest AUC for three samples, therefore it was deemed the least preferred.

Using a ranking order in which 1 = highest Classification Rate and 4 = lowest Classification Rate (The higher the better), Model 1 was the most preferred model since it had the highest classification rate for all the five samples. Model 2 had the second highest classification rate for Samples 2 and 4, and the third highest for the remaining samples, therefore it was deemed the second most preferred model as shown in Table 4. The table also shows that Model 3 is the third most preferred

Table 3: AUC.

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank
Model 1	0.7713	1	0.7532	1	0.7617	2	0.7889	1	0.7954	1
Model 2	0.7222	3	0.7144	2	0.7661	1	0.7808	2	0.7673	3
Model 3	0.7188	4	0.7008	3	0.7469	3	0.7710	3	0.7705	2
Model 4	0.7401	2	0.6660	4	0.6414	4	0.7338	4	0.6475	4

Table 4: Classification rate.

	Sample 1		Sample 2		Sample 3		Sample 4		Sample 5	
	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank
Model 1	58.9	1	60.0	1	61.1	1	64.4	1	65.0	1
Model 2	54.4	3	56.7	2	58.9	3	61.7	2	62.8	3
Model 3	53.3	4	55.6	4	60.0	2	61.1	3	63.9	2
Model 4	58.3	2	56.7	3	55.6	4	53.3	4	62.2	4

model since unlike Model 4 which had the lowest classification rate for Samples 3, 4 and 5, Model 3 had the lowest classification rate for Samples 1 and 2, it also had the second highest classification rate for Samples 2 and 5 whereas Model 4 only had a second highest classification rate for Sample 1.

Since three of the model comparison criteria namely: AIC, AUC and Classification Rate favoured Model 1, which was fitted using stepwise selection, the study concluded that stepwise selection gives the best fit of the data. In Table 5, this paper examined the variables that were selected by the stepwise selection (Model 1) across the five samples. Table 5 shows that although stepwise selection outperformed the other three competing purposeful variable selection approaches, it selected different variables across the five samples, and this lack of consistency is deemed to be a limitation of this purposeful variable selection method.

Table 5: Variables selected by stepwise selection method (Model 1).

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Population group (race) of persons, average time to police station, gender, age, satisfaction with police, trust in metro/traffic police, children approach police officer, visited the police station in three years	Population group (race) of persons, visited the police station in three years, official contact with police, police response in emergency, satisfaction with police, police officers on duty	Population group (race) of persons, visited the police station in three years, official contact with police, police officers on duty, specialised police operations	Population group (race) of persons, marital status, education level attained, visited the police station in three years, official contact with police, police response in emergency, satisfaction with police	Age, nearest police station, average time to police station, visited the police station in three years, police officers on duty, specialised police operations, children approach police officer, SAPS, trust in metro/ traffic police

Although it has a limitation of inconsistent selection of variables across samples, stepwise selection was most preferred purposeful variable selection approach and was used on the complete dataset to fit a multivariate logistic regression for predicting the likelihood of a victim reporting a crime of housebreaking. To address the stepwise selection's limitation of selecting varying variables for different observations of the same dataset, the researchers used Bootstrapping with 1000 replications to select the optimum multivariate model for predicting the likelihood of a victim reporting a crime of housebreaking. The model was further assessed for overall significance, significance of individual parameters, model fit, and prediction ability, the results are presented in Section 3.4.

3.4 A Multivariate Logistic Regression Model to Predict the Likelihood of Reporting a Crime of Housebreaking

The Wald chi-square test tests whether the overall multivariate LRM is significant. The test is significant at 5% (Chi-Square = 136.639, $df = 11$, p-value < 0.05) which indicates that at least one of the coefficients of the model is not equal to zero, meaning that the model is significant overall.

Table 6 lists all the variables selected by stepwise selection for entry into the final multivariate model at a significance level of 0.3 and the significance level for

Table 6: Summary of the variables entered in the model.

Step	Effect entered	DF	Score chi-square	P_r > Chi-square	Variable label
1	Race	1	95.9406	<0.0001	Population group of the persons in the household
2	Q64ContPolice	1	24.4519	<0.0001	Official contact with police
3	Q61NEARPOLSTA	1	44.8877	<0.0001	Nearest police station
4	Q62Time	1	12.1950	0.0005	Average time to police station
5	Q68Satisfied	1	8.3096	0.0039	Satisfaction with police
6	Q63POLSTA3YRS	1	6.9141	0.0086	Visited the police station in three years
7	Q12MSTATUS	1	3.3526	0.0671	Marital status of the persons in the household
8	Q671Time	1	3.2487	0.0715	Police response in emergency
9	Q13HIEDU	1	2.3526	0.1251	Educational attainment of the persons in the household
10	Q614APPROACH	1	2.3466	0.1256	Children approach police officer
11	Gender	1	1.3057	0.2532	Gender of persons in the household

Table 7: Model fit Statistics.

Criterion	Intercept-only	Intercept and variables
AIC	1462.450	1293.465
BIC	1467.417	1353.068

the variable to stay in the model was 0.35 (SAS defaults). The table shows that 11 of the 17 candidate variables were retained in the final model.

Table 7 shows the results of the tests performed on the final model yielded by stepwise selection method after all the variables are selected against the intercept-only model. The model with the intercept and variables fits the data well compared to the intercept-only model since it has the lowest AIC and BIC as shown in Table 7.

The Hosmer and Lemeshow test indicates that there is not enough evidence that the model is not a good fit since the *p*-value of the test is insignificant at 5% (Chi-Square = 10.9819, *df* = 8, *p*-value = 0.203), therefore the results of the Hosmer and Lemeshow test confirm those in Table 7 that the multivariate LRM established in this study fits the data well.

Table 8 shows that the following variables are significant (*p*-values < 0.05) determinants of the likelihood of reporting a crime of housebreaking: Population group of the persons in the household (Race), Average time to police station (Q62Time), Visited the police station in three years (Q63POLSTA3YRS), Official contact with police (Q64ContPolice), and Satisfaction with police (Q68Satisfied). To determine the effect of these significant variables on the likelihood of reporting a crime of housebreaking, the study considered the odds ratios in Table 9.

Table 8: Test for significance of individual parameters.

Parameter	DF	Estimate	Standard error	Wald chi-square	P_r > Chi-square
Intercept	1	−1.7279	1.3226	1.7069	0.1914
Gender	1	−0.1672	0.1464	1.3046	0.2534
Race	1	0.6057**	0.0906	44.6792	<0.0001
Q12MSTATUS	1	−0.0362	0.0278	1.7023	0.1920
Q13HIEDU	1	0.0134	0.00823	2.6620	0.1028
Q61NEARPOLSTA	1	2.5771	1.3209	3.8064	0.0511
Q62Time	1	−0.2017*	0.0904	4.9842	0.0256
Q63POLSTA3YRS	1	−0.2464**	0.0951	6.7089	0.0096
Q64ContPolice	1	−0.6469**	0.1406	21.1602	<0.0001
Q671Time	1	−0.1083	0.0625	3.0058	0.0830
Q68Satisfied	1	0.5034**	0.1542	10.6507	0.0011
Q614APPROACH	1	−0.1256	0.0805	2.4380	0.1184

Table 9: Odds ratios for significant determinants of the likelihood of reporting a crime of housebreaking.

Variables in the equation								
	B	**S.E.**	**Wald**	***df***	**Sig.**	**Exp (*B*)**	**95% C.I. for EXP (*B*)**	
							Lower	**Upper**
Population group of the persons in the household(1)	1.506**	0.307	24.002	1	0.000	4.508	2.468	8.234
Population group of the persons in the household(2)	0.983**	0.371	7.007	1	0.008	2.671	1.291	5.529
Population group of the persons in the household(3)	−0.233	0.637	0.134	1	0.715	0.792	0.227	2.762
Average time to police station(1)	−0.905	0.877	1.065	1	0.302	0.405	0.073	2.257
Average time to police station(2)	−0.931	0.885	1.106	1	0.293	0.394	0.070	2.235
Average time to police station(3)	−0.238	0.945	0.064	1	0.801	0.788	0.124	5.026
Average time to police station(4)	0.190	1.450	0.017	1	0.896	1.210	0.071	20.733
Visited the police station in three years(1)	−0.523	0.684	0.583	1	0.445	0.593	0.155	2.267
Visited the police station in three years(2)	0.356	0.689	0.267	1	0.605	1.427	0.370	5.503
Official contact with police(1)	−21.329	22131.684	0.000	1	0.999	0.000	0.000	
Official contact with police(2)	−20.818	22131.684	0.000	1	0.999	0.000	0.000	
Satisfaction with police(1)	0.473**	0.167	8.023	1	0.005	1.605	1.157	2.227
Constant	−19.882	40193.441	0.000	1	1.000	0.000		

Only variables which are significantly contributing to the variation in the likelihood of reporting a crime of housebreaking as shown in Table 8 are presented in Table 9. Table 9 shows that the odds of a victim reporting a crime of housebreaking is 4.508 times greater for victims in the population group (race) of 1 (Black African) as opposed to those in the other population groups, and is 2.671 times greater for victims in the population group (race) of 2 (Coloured) as opposed to those in the other population groups. The table also shows that the odds of reporting a crime of housebreaking is 1.605 times greater for victims who indicated their satisfaction with the police as 1 (Yes) as opposed to those who indicated that they are not satisfied with the police. All other categories are either insignificant at 5% level of significance (p-values > 0.05) or did not converge (no statistics in the upper confidence interval).

Table 10 shows the confusion matrix describing the performance of the classifier. The results in the table show that the rate of correctly predicting an event of reporting a crime (also known as sensitivity) using the multivariate LRM established in this study is 72.9%. Table 10 also shows that the chance of the model correctly predicting a non-event that is, not reporting the crime (also known as specificity) is 63.2%. The table further indicates a possible 36.8% chance of Type I error (false positives) and 27.1% chance of Type II error, which is the chance of incorrectly predicting the reported crimes as being non-reported (false negatives).

Figure 1 shows that the model with all the stepwise selected variables (step 11) has more predictive ability for the likelihood of reporting a crime of housebreaking when compared to the models fitted at each stage of the stepwise selection process. Generally, the AUC in Figure 1 is 73.73% and this indicates that the stepwise selection method resulted in a good model since it is capable of correctly distinguishing between a victims' chance of reporting a crime of housebreaking and the chance of not reporting the crime.

4 Discussion and Conclusions

This paper sought to identify the most efficient variable selection method for fitting a parsimonious model to predict the likelihood of reporting a crime of housebreaking through a comparison of the H-L algorithm, the bivariate logistic regression, stepwise selection and the chi-square test of association. This was done not only to identify the best variable selection method but to also identify the contributing factors to a victim's likelihood of reporting a crime of housebreaking. The stepwise selection method was identified as the most efficient compared to the other three selection methods, across all the five samples. However, it is worth noting that given the same variables and the same number of observations, the

Table 10: Confusion matrix.

		Predicted group	
		Yes	**No**
Housebreaking/burglary – reporting crime to the police	Yes	72.9%	27.1%
	No	36.8%	63.2%
Total		56.6%	43.4%

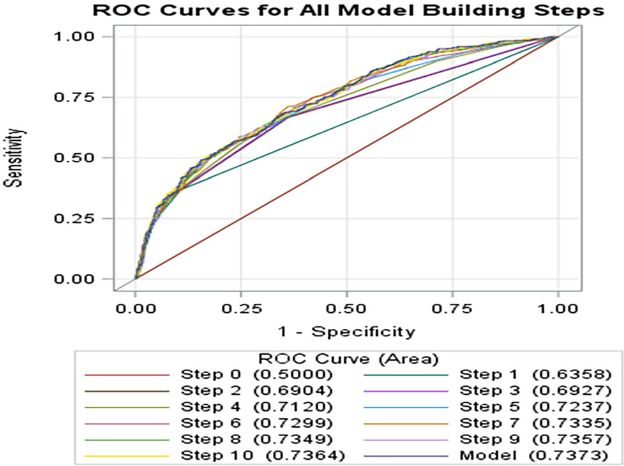

Figure 1: Receiver operating characteristic (ROC) curve.

study found that the stepwise selection method tends to select different variables for each unique set of observations as it was observed across the five random samples used in this study. Since it was identified as the most efficient, stepwise selection method was then applied to the actual dataset (including all the observations) to identify the determinants of the victim's likelihood of the reporting a crime of housebreaking in South Africa.

Bursac et al. (2008) performed a similar comparative study testing the H-L algorithm against the backward, forward and stepwise selection methods. The current study is similar to that by Bursac et al. (2008) since both studies used the stepwise selection procedure versus the H-L algorithm. However, the current study extended the scope of the study by Bursac et al. (2008) by considering the use of bivariate logistic regression methods and the chi-square test of association as the other competing models. In their study, Bursac et al. (2008) used two simulation experiments whereas the current study extended the approach of Bursac et al. (2008) by drawing five equal samples from a real life dataset. Unlike in the study by Bursac et al. (2008) who ran two simulations, both having equal variables (based on an assumption that there are six important variables) under different sample sizes, the current study was not interested in the sample size variation, but on whether the results can be replicated.

Bursac et al. (2008) found that the H-L-algorithm correctly identified and retained variables at a larger rate compared to stepwise selection methods in their study. Unlike Bursac et al. (2008) the current study found stepwise selection to be the better selection method in comparison to the competing methods and the H-L algorithm did not perform as well as in the study by Bursac et al. (2008). The results of the current study may be differing from those obtained by Bursac et al. (2008) due to the difference in methodology followed such as using five samples, and the different cut-off points used.

Authors such as Radwan et al. (2013), Folkerson et al. (2015), Chen et al. (2009), Randall et al. (2013), Griffin et al. (2013), and Bohlke et al. (2009), to name a few, used the H-L algorithm to fit multiple logistic regression models. However, these authors only used this purposeful selection method to fit a multivariate model in order to test associations in their data and they were not interested in confirming whether this novel variable selection method fitted the data well when compared to the other popularly used variable selection methods. Some studies such as the ones conducted by Austin and Tu (2004) and by Bursac et al. (2008) compared some of the approaches considered in the current study but none of these studies considered all the popular approaches simultaneously. As such, the main contribution of this paper to the literature around purposeful variable selection approaches for logistic regression is through a comparison of all four popular variable selection methods simultaneously.

Another contribution of this study is the use of five randomly selected equal samples which were selected from the actual dataset. Although some previous studies such as those conducted by Austin and Tu (2004) and Larkin et al. (2010) have used bootstrap samples, these studies did not compare the competing variable selection methods simultaneously under each of the bootstrap samples. Instead, such studies selected the optimum model from each of the samples, and then they compared all the selected models; therefore, the approach of a simultaneous comparison of models under each sample in this study has never been used previously.

The use of five randomly selected and equal samples was done to ensure the validity of results, whether the results are replicative, and statistical rigour. Through this, the current study ensured that the recommended variable selection method remains to be stepwise selection regardless of using different observations of the same dataset, and was able to identify that although the method is the most efficient, its limitation is that it selected different variables for the different observations of the same dataset, which is something that was never discovered before. To the best of the researchers' knowledge, there has never been a study conducted to statistically determine the determinants of a victim's likelihood of reporting a crime of housebreaking. As such, the current study contributed a new idea to the literature of crime studies.

Based on the methodology implemented in this study, it is recommended that future researchers should not only divide the data into two (training data and application data) or compare models using only one dataset but should ensure validity and replicability of the results by comparing models using several datasets (or samples) with the same characteristics. It is also recommended that researchers who do not specialise in Statistics and who may not be more interested in comparing the models may use the stepwise selection method to attain a good fit for the data with good classification ability as this method was found to outperform the other variable selection methods. However, such researchers are advised to use Bootstrapping to determine the optimum model since stepwise selection was found to select different variables for different samples of the same dataset.

Emanating from the findings of the current study, it is also recommended that policy-makers, crime-fighting agencies, crime researchers and other interested parties should take into consideration the results of this study relative to the determinants of a victim's likelihood to report a crime of housebreaking. By focusing on these variables, these parties may be able to encourage the victims to report such crimes and this will enable law enforcers to address such cases effectively. It is also recommended that future studies may consider comparing the purposeful variable selection methods considered in this study under different conditions such as a mixture of categorical and continuous variables, different sample sizes and in the presence of missing values, to name a few. Similar studies may be conducted for other regression methods other than the binary logistic regression method or for other predictive models such as discriminant analysis.

Funding: This paper forms part of research dissertation which was financially supported by the North West University (NWU), South Africa.
Availability of data and material: The data may be made available on request.
Code availability: The SAS code may be made available on request.
Conflicts of interest/Competing interests (include appropriate disclosures): There is neither conflict of interest nor competing interests.

Appendices

Appendix 1: Descriptive Statistics for Age

Descriptive statistics					
	n	Minimum	Maximum	Mean	Std. deviation
Age of persons in the household	1061	14	103	46.80	15.037

Appendix 2: Frequency Table for Categorical Variables

		Frequency	Percent
House breaking/burglary – reporting crime to the police	Yes	583	54.9
	No	478	45.1
	Total	1061	100.0
Official contact with police	Yes	455	42.9
	No	602	56.7
	Unspecified	4	0.4
	Total	1061	100.0
Police response in emergency	Less than 30 min	195	18.4
	Less than 1 h (but more than 30 min)	239	22.5
	Less than 2 h (but more than 1 h)	186	17.5
	More than 2 h	363	34.2
	Never arrive	77	7.3
	Unspecified	1	0.1
	Total	1061	100.0
Satisfaction with police	Yes	442	41.7
	No	619	58.3
	Total	1061	100.0
Police officers on duty	At least once a day	358	33.7
	At least once a week	268	25.3
	At least once a month	133	12.5
	More than once a month	70	6.6
	Never	220	20.7
	Unspecified	12	1.1
	Total	1061	100.0
Specialised police operations	Yes	223	21.0
	No	836	78.8
	Unspecified	2	0.2
	Total	1061	100.0
Children approach police officer	Yes	943	88.9
	No	107	10.1
	Unspecified	11	1.0
	Total	1061	100.0
Trust in the SAPS	Yes	696	65.6
	No	361	34.0
	Unspecified	4	0.4
	Total	1061	100.0
Trust in metro/traffic police	Yes	725	68.3
	No	332	31.3
	Unspecified	4	0.4
	Total	1061	100.0

(continued)

		Frequency	Percent
Gender of persons in the household	Male	654	61.6
	Female	407	38.4
	Total	1061	100.0
Population group of the persons in the household	Black African	795	74.9
	Coloured	106	10.0
	Indian/Asian	27	2.5
	White	133	12.5
	Total	1061	100.0
Marital status of the persons in the household	Married	386	36.4
	Living together like husband and wife	104	9.8
	Divorced	57	5.4
	Separated, but still legally married	14	1.3
	Widowed	142	13.4
	Single, but have been living together with someone as husband	17	1.6
	Single and have never been married/never lived together as h	334	31.5
	Unspecified	7	0.7
	Total	1061	100.0
Educational attainment of the persons in the household	Grade R/0	4	0.4
	Grade 1/Sub A/Class 1	7	0.7
	Grade 2/Sub B/Class 2	17	1.6
	Grade 3/Standard 1/ABET 1 (Kha Ri Gude, Sanli)	13	1.2
	Grade 4/Standard 2	24	2.3
	Grade 5/Standard 3/ABET 2	21	2.0
	Grade 6/Standard 4	45	4.2
	Grade 7/Standard 5/ABET 3	50	4.7
	Grade 8/Standard 6/Form 1	68	6.4
	Grade 9/Standard 7/Form 2/ABET 4	51	4.8
	Grade 10/Standard 8/Form 3	124	11.7
	Grade 11/Standard 9/Form 4	94	8.9
	Grade 12/Standard 10/Form 5/Matric (No Exemption)	222	20.9
	Grade 12/Standard 10/Form 5/Matric (Exemption *)	21	2.0
	NTC 1/N1/NC (V) level 2	2	0.2
	NTC 2/N2/NC (V) level 3	4	0.4
	NTC 3/N3/NC (V) level 4	7	0.7
	N4/NTC 4	6	0.6
	N5/NTC 5	3	0.3
	N6/NTC 6	5	0.5

(continued)

		Frequency	Percent
	Certificate with less than Grade 12/Std 10	4	0.4
	Diploma with less than Grade 12/Std 10	7	0.7
	Certificate with Grade 12/Std 10	17	1.6
	Diploma with Grade 12/Std 10	72	6.8
	Higher Diploma (Technikon/University of Technology)	15	1.4
	Post Higher Diploma (Technikon/University of Technology, Mas)	3	0.3
	Bachelor?s degree	41	3.9
	Bachelor?s degree and post-graduate diploma	13	1.2
	Honours degree	14	1.3
	Higher degree (Masters, Doctorate)	14	1.3
	Other (specify in the box below)	2	0.2
	Do not know	10	0.9
	No schooling	59	5.6
	Unspecified	2	0.2
	Total	1061	100.0
Nearest police station	Yes	1055	99.4
	No	5	0.5
	Unspecified	1	0.1
	Total	1061	100.0
Average time to police station	Less than 30 min	744	70.1
	Less than 1 h (but more than 30 min)	249	23.5
	Less than 2 h (but more than 1 h)	47	4.4
	More than 2 h	6	0.6
	Not applicable	5	0.5
	Unspecified	10	0.9
	Total	1061	100.0
Visited the police station in three years	Yes	693	65.3
	No	350	33.0
	Not applicable	5	0.5
	Unspecified	13	1.2
	Total	1061	100.0

References

Austin, P. C., and J. V. Tu. 2004. “Automated Variable Selection Methods for Logistic Regression Produced Unstable Models for Predicting Acute Myocardial Infarction Mortality.” *Journal of Clinical Epidemiology* 57 (11): 1138–46.

Bohlke, M., S. S. Marini, M. Rocha, L. Terhorst, R. H. Gomes, F. C. Barcellos, M. C. C. Irigoyen, and R. Sesso. 2009. “Factors Associated with Health-Related Quality of Life after Successful Kidney Transplantation: A Population-Based Study.” *Quality of Life Research* 18 (9): 1185–93.

Bursac, Z., C. H. Gauss, D. K. Williams, and D. W. Hosmer. 2008. “Purposeful Selection of Variables in Logistic Regression.” *Source Code for Biology and Medicine* 3 (1): 17.

Chen, S.-C., Y.-T. Lee, C.-H. Yen, K.-C. Lai, L.-B. Jeng, D.-B. Lin, P.-H. Wang, C.-C. Chen, M.-C. Lee, and W.R. Bell. 2009. “Pyogenic Liver Abscess in the Elderly: Clinical Features, Outcomes and Prognostic Factors.” *Age and Ageing* 38 (3): 271–6.

Fabozzi, F. J., S. M. Focardi, S. T. Rachev, B. G. Arshanapalli, and M. Hoechstoetter. 2014. *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*. Hoboken, New Jersey: Wiley. https://doi.org/10.1002/9781118856406.

Fanelli, M., E. Kupperman, E. Lautenbach, P. H. Edelstein, and D. J. Margolis. 2011. “Antibiotics, Acne, and Staphylococcus Aureus Colonization.” *Archives of Dermatology* 147 (8): 917–21.

Folkerson, L. E., D. Sloan, B. A. Cotton, J. B. Holcomb, J. S. Tomasek, and C. E. Wade. 2015. “Predicting Progressive Hemorrhagic Injury from Isolated Traumatic Brain Injury and Coagulation.” *Surgery* 158 (3): 655–61.

Griffin, A. T., T. L. Wiemken, and F. W. Arnold. 2013. “Risk Factors for Cardiovascular Events in Hospitalized Patients with Community-Acquired Pneumonia.” *International Journal of Infectious Diseases* 17 (12): e1125–9.

Hacke, W., G. Donnan, C. Fieschi, M. Kaste, R. von Kummer, J. P. Broderick, The ATLANTIS, ECASS, and NINDS rt-PA Study Group Investigators. 2004. “Association of Outcome with Early Stroke Treatment: Pooled Analysis of ATLANTIS, ECASS, and NINDS Rt-PA Stroke Trials.” *Lancet (London, England)* 363 (9411): 768–74.

Hajian-Tilaki, K. 2013. “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.” *Caspian Journal of Internal Medicine* 4 (2): 627. PMID: 24009950 and PMCID: PMC3755824.

Heeringa, S. G., B. T. West, and P. A. Berglund. 2010. *Applied Survey Data Analysis*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781420080674.

Hilbe, J. M. 2011. “Logistic Regression.” In *International Encyclopedia of Statistical Science*, edited by M. Lovric, 755–8. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hosmer, D., and S. J. N. Y. Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: John Wiley & Sons. https://doi.org/10.1002/0471722146.

Hosmer, D., and S. Lemeshow. 2004. *Applied Logistic Regression*, 2nd ed. Hoboken, NJ: John Wiley and Sons.

Hosmer David, W., and L. Stanley. 2000. *Applied Logistic Regression*. New York: Wiley. 0-471-61553-6.

Larkin, G. L., W. S. Copes, B. H. Nathanson, and W. Kaye. 2010. “Pre-Resuscitation Factors Associated with Mortality in 49,130 Cases of In-Hospital Cardiac Arrest: A Report from the National Registry for Cardiopulmonary Resuscitation.” *Resuscitation* 81 (3): 302–11.

Menard, S. 2002. *Applied Logistic Regression Analysis*. Thousand Oaks, CA: SAGE Publications. https://dx.doi.org/10.4135/9781412983433.

Moisey, L. L., M. Mourtzakis, B. A. Cotton, T. Premji, D. K. Heyland, C. E. Wade, E. Bulger, R. A. Kozar, and Nutrition and Rehabilitation Investigators Consortium (NUTRIC). 2013. "Skeletal Muscle Predicts Ventilator-Free Days, ICU-Free Days, and Mortality in Elderly ICU Patients." *Critical Care* 17 (5): R206.

Murtaugh, P. A. 2009. "Performance of Several Variable-Selection Methods Applied to Real Ecological Data." *Ecology Letters* 12 (10): 1061–8.

North, R. A., L. M. McCowan, G. A. Dekker, L. Poston, E. H. Chan, A. W. Stewart, R. S. Taylor, P. N. J. B. Baker, and L. C. Kenny. 2011. "Clinical Risk Prediction for Pre-Eclampsia in Nulliparous Women: Development of Model in International Prospective Cohort."*BMJ* 342: d1875.

Olusegun, A. M., H. G. Dikko, and S. U. Gulumbe. 2015. "Identifying the Limitation of Stepwise Selection for Variable Selection in Regression Analysis." *American Journal of Theoretical and Applied Statistics* 4 (5): 414–9.

Park, H. 2013. "An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain." *Journal of Korean Academy of Nursing* 43 (2): 154–64.

Peng, C.-Y. J., T.-S. H. So, F. K. Stage, and E. P. S. John. 2002. The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999. *Research in Higher Education* 43(3): 259–93.

Radwan, Z. A., Y. Bai, N. Matijevic, D. J. del Junco, J. J. McCarthy, C. E. Wade, J. B. Holcomb, and B. A. Cotton. 2013. "An Emergency Department Thawed Plasma Protocol for Severely Injured Patients." *JAMA Surgery* 148 (2): 170–5.

Randall, J. R., B. H. Rowe, K. A. Dong, M. K. Nock, and I. Colman. 2013. "Assessment of Self-Harm Risk Using Implicit Thoughts." *Psychological Assessment* 25 (3): 714–21.

Sarkar, S., and H. Midi. 2010. "Importance of Assessing the Model Adequacy of Binary Logistic Regression." *Journal of Applied Sciences* 10 (6): 479–86.

Schlotzhauer, D. C. 1993. "Some Issues in Using PROC LOGISTIC for Binary Logistic Regression." *Observations: The Technical Journal for SAS Software Users* 2 (4): 12 p.

Shervin, A., D. del Junco, K. Sutter, T. A. McNearney, J. D. Reveille, A. Karnavas, P. Gourh, R. M. Estrada-Y-Martin, M. Fischbach, F. C. Arnett, and M. D. Mayes. 2009. "Clinical and Genetic Factors Predictive of Mortality in Early Systemic Sclerosis." *Arthritis Care & Research* 61 (10): 1403–11.