A REVIEW OF SOME GOODNESS-OF-FIT TEST FOR LOGISTIC REGRESSION MODEL

ABSTRACT

Some goodness-of-fit statistic (Likelihood ratio test, Pseudo (R^2) test, Hosmo-lemeshow test and Chi-square test) were used to determine model fit in binary logistic regression, their estimates were compared to each other, to determine their performances. From our analysis the likelihood ratio test, the pseudo (R^2) test and the chi-square test, produced results that are similar and shows the presences of poor fit, with model1 being the most fit of the three models. Even though hosmo-lemeshow test produced a P-value that is very low, indicating the presence of poor fit in all three models, the differences between these models were not revealed, because the test produced the same estimates for the three models. The likelihood ratio test produced results that are similar to the chi-square test. We also observed that the goodness-of-fit parameters (Null Deviances, Residual Deviance and the Alkaike Information Criteria (AIC)), that accompany model formation in R-statistical software produced result that align with the goodness-of-fit statistic used, the set back is that the summary provided by this parameters, does not reveal model similarities.

Keyword: likelihood ratio, Multivariable, Model, Logistic Regression, Goodness-of-fit.

1.0 INTRODUCTION

The four types of multivariate methods are linear regression, logistic regression, discriminant analysis, and proportional hazard regression which are primarily utilized in the health sciences. Although the expression and output of the outcome variables are different in these methods, they share mathematical similarities. The result variable of linear regression is a continuous quantity, such as an individual's height and weight. An outcome produced by logistic regression is typically a binary event, such as "Yes" or "No," Case vs. Control. In discriminant analysis, which yields outcomes akin to logistic regression, the outcome variable is category or class to which subjects belong for two categories. The result of the proportional hazards regression indicate how long it will take for a binary "failure" event to occur. i.e. (death) throughout an observational follow-up period. Of them, the logistic regression is the most often utilized in the healthcare industry, [16]. A technique called logistic regression (LR) models several independent or predictor factors in order to calculate the likelihood of a binary occurrence. Because the exponentiation of the parameters obtained from logistic regression forms an odds ratio, which measures the strength of association between the independent and dependent variables, epidemiologists primarily examine the impact of multiple independent variables on a binary outcome, such as the presence or absence of disease. The logistic regression model is the most

often used regression technique in applied research, and it has gained widespread acceptance as a useful tool for analyzing binary outcome variables. Finding a source journal without at least one article using the model is frequently difficult. The accessibility of user-friendly software in all major statistical packages and the simplicity of interpreting the fitted model's results are the main causes of its popularity.[9],[2]. A continual stream of new statistical research is being conducted on model evaluation and fit techniques as well as model expansions to novel environments. Even yet, there have been shortcomings in the model fit techniques that are now accessible, which is what led us to conduct this study. The predicted value of the outcome variable is expressed by the logistic model as the total of products of the predicted variable, where each product is created by multiplying the independent variable's value and coefficient. The best mathematical fit for the given model is found in these coefficients. After controlling for all other independent variables, a coefficient shows the effect of each independent variable on the outcome variable. In order to forecast the value of the dependent variable for each new value of the independent variables added to the model, the model may be used for two purposes: first predicting the value of the dependent variable for each new value of the independent variables enter into the model, and secondly for illuminating the relative contributions of each independent variable to the dependent variable while accounting for the effects of other independent variables. This study was conducted to evaluate the results of several goodness-of-fit test in determining model correctness and to illustrate the idea of logistic regression.

2.0 LITERATURE REVIEW

A multivariable technique designed for dichotomous variables is called logistic regression [7]. This technique, sometimes known as the logistic model or logit model is used to simulate the possibility of a given event occurring, such as pass/fail, win/lose, alive/dead. It can be applied to represent how several independent factors are related to a categorical dependent variable. This correlation can be displayed on an S-shaped logistic curve. There are three types of logistic models: ordinal, multinomial, and binary logistic regression. When the dependent variable is dichotomous, the binary model is utilized; when there are more than two categories and the independent variables are either continuous or categorical, the multinomial model is employed; and when there are more than two categories that exhibit a particular pattern, the ordinal model is employed. The reason the logistic model is well-known is that it presents results in a probabilistic manner, with values ranging from 0 to 1. Additionally, an S-shaped curve depicted in the plots of these values illustrates the combined influence of all explanatory variables on the dependent variable[9], [4]. The actual application of the approach is the same even though the kind of data utilized for the dependent variable in logistic regression differs from that of linear regression [1].

The hypothesis that the distribution of the observed outcome variable matches the distribution of the conjectured observation produced by model upon fitting additional data is tested using goodness-of-fit (GOF) statistics. This thesis focuses on GOF tests. When the explanatory variable is discrete in character, the Pearson chi-square test $\left(X^2\right)$ might be a useful tool [11], [9]. However, if one or more of the model's explanatory variables are continuous, each observation will provide a different trend, potentially leading to multiple groupings. The asymptotic theory that underpins the distribution has becomes obsolete if the number of covariate patterns rises at a pace that is almost equal to the number of observations [11],[9] and [5]. Many goodness-of-fit tests for logistic regression have been developed over the last three decades, grouping

observations to address the challenge posed by $\left(X^2\right)$ test i.e., ([5], [6], [12],[15], [13],[14]). Numerous grouping techniques have been developed; some group the outcome variable's predicted probabilities, while others base their groupings strategy on the model's covariates. The number of groups selected and partitioning technique may have an impact on the test results. One of the most widely used goodness-of-fit test created to address the problems that arise when continuous covariates are included in a binary logistic regression model is the [6]. It has received widespread review in the literature and it effectiveness has been compared with numerous other goodness-of-fit statistics for logistic regression (e.g. [10],[8],[12],[13],[14]). Similar to that of the X^2, HL, process groups the anticipated probability by sorting them and assigning them to groups called the "deciles-of-risk" (DOR). The (DOR) method is widely used to create the groups where observations are graded based on their anticipated probability and placed into G = 10 roughly equal sized groups, with HL having an approximate $X^2(G-2)$ distribution. The expected number of observations in each group is determined using the mean of predicted probabilities within that group. According to[3], different grouping arrangement might result in different values of LR when there are ties between the expected probability and the ties fall on the boundary between deciles. This happens because each time the software is run; observation may be assigned to different groups because the ranks of the connected observations are not stable. [11], advise grouping connected observations together to prevent this issue.

2.3 Evaluation Metrics Used for Logistic Regression in R

AIC is a useful metric for model fit and is comparable to the adjusted R square in multiple linear regressions, and aids in preventing overfitting. The smaller it values the better, in the model summary. It aids in penalizing the model's growing number of coefficients. It would be more logical to compare these figures for several models rather than looking at this value for just one. It can therefore be applied to model selection. The model with the lowest AIC, for instance, will be the best of all of them, when having two or many alternative models producing the same result. Another assessment measure that is produced in your R output is called Deviance. It is separated into two parts: Null and Residual Deviance. When calculating the null deviation, the intercept is the only variable taken into account; all other variables are eliminated. We compute the residual deviation for each of the model's covariates. The residual deviation in a linear regression model can be compared to the residual sum of squares, whereas the null deviance can be compare to total sum of squares. A model is considered better if the discrepancies between the null and residual deviation are larger. The model that explains deviation the best is the one with a lower null deviance, which may also be used to compare different models. Furthermore, a model is better when it have lower residual deviation. When assessing model fit, AIC is typically prioritized over deviation.

3.0 MATERIALS AND METHODS

A sigmoid or S-shaped curve of the logistic model, is frequently used to simulate population expansion [17]. The ratio of the likelihood that an event won't occur is (1-K) if the probability of it happening is K. Then the resulting odds can be given by

$$odds = \frac{K}{1 = K}$$

Since logistic regression compares the likelihood of an event occurring to its likelihood of not occurring, the impact of the independent variables is typically described in terms of odds. With logistic regression the mean of the response variable p in terms of an explanatory variable x is modeled relating $\phi(x)$ and x through the equation $\phi(x) = \gamma + \eta x$. Note that this is a flawed model since extreme values of x will provide a result of $\phi(x) = \gamma + \eta x$ that does not fall into the range 0 and 1. Therefore, the natural logarithm should be used to transform the odds in order to solve this problem [13]. Through the use of linear function of the explanatory variable to represent the log odds:

$$\log it(y) = \ln(odds) = \ln\left(\frac{\eta(x)}{1-\eta(x)}\right) = \gamma + \eta x$$

Where $\phi(x)$ is the interest outcome probability and x is the explanatory variable. Keep in mind that this is a basic model, with ϕ and η, serving as the parameters of the logistic regression. By taking the antilog, an equation for estimating the likelihood that an outcome of interest will occur can be derived as

$$\eta(x_i) = \Pr(y_i = 1/x_i) = \frac{e^{\phi+\eta x}}{1+e^{\phi+\eta x}} = \frac{1}{1+e^{-(\phi+\eta x)}}$$

By using the understanding of multiple logistic regression to simple logistic regression, one can create the following equation:

$$\log it(y) = \ln(odds) = \ln\left(\frac{\eta(x)}{1-\eta(x)}\right) = \phi + \eta_1 x_1 + \cdots + \eta_k x_k$$

Therefore, $\eta(x_i) = \Pr(y_i = 1/x_i) = \frac{e^{\phi+\eta_1 x_1+\cdots+\eta_k x_k}}{1+e^{\phi+\eta_1 x_1+\cdots+\eta_k x_k}} = \frac{1}{1+e^{-(\phi+\eta_1 x_1+\cdots+\eta_k x_k)}}$

Keep in mind that the odds ratio (OR) compares two odds in relation to distinct events. That is to say, given tow event y and z, the equivalent chances of y happening in relation to z occurring is

$$odds(yvsz) = \frac{odds(y)}{odds(z)} = \left(\frac{P(y)}{1-P(y)}\right) \Big/ \frac{P(z)}{1-P(z)}$$

An odd ratio is typically used to calculate the correlation between an exposure and a result. The odds ratio (OR) displays the likelihood of an event (disease or illness) occurring in relation to a

certain exposure (health behavior, medical history), as opposed to the likelihood of the outcome occurring in the absence of that exposure. The predicted rise in the log odds of the result for each unit increase in the value of the independent variable is known as the regression coefficient η_i, in logistic regression calculations. Stated otherwise, the OR corresponding to an increase of one unit in the independent variable is the exponential function of the regression coefficient. The OR can also be used to assess the relative importance of different risk factors for a given outcome and to ascertain whether a given exposure is a risk factor for that outcome. If OR=1, it indicates that exposure has no effect on the odds of the result; if OR>1, it indicates that exposure increases the odds of the outcome; and if OR<1, it indicate that exposure reduces the odds of the outcome. For instance, if the odds ratio for the variable "diabetics" is 1.2, the variable is coded as 0=not diabetic and 1=diabetic. Hence, in cases with diabetes, the likelihood of favorable outcome is 1.2 times greater than in those without diabetes. One method for expanding the OR beyond two binary variables is to use logistic regression [13].

3.1 Deviance

The fits of two or more models to the observed data are compared using the deviation. The null and residual deviances are the two components of this likelihood ratio test. The null model that is being examined is the first, and the saturated model that the null is nested in is the second. For every j covariate pattern, there is a parameter in the saturated model. The theory under investigation is that all saturated model parameter equal 0 that are not included in the working model. One way to express the residual deviation is as

$$d\left(\mathbf{z}_i, \hat{\boldsymbol{\lambda}}_i\right) = \pm\left\{2\left[z_i \ln\left(\frac{z_i}{n_i\lambda_i}\right) + (n_i - z_i)\ln\left\{\frac{(n_i - z_i)}{n_i(1-\lambda_i)}\right\}\right]\right\}^{1/2},$$

Where n_i is the number of observation with i^{th} covariate pattern and $\hat{\lambda}_i$ is the probability for the i^{th} covariate pattern. z_i Is the number of observations from the n_i subjects with response $z = 1$. When testing the fit of a binary logistic regression model with k
fitted covariates and i^{th} covariate patterns, the deviance statistic is expressed as

$D = \sum_{i=1}^{k} d(z_i, \lambda_i)_i^2$, When $n_i\hat{\lambda}_i$,is not small,the deviance has an asymptotic distribution that is $X^2(i-k-1)$, ([6],[7],[10]).

3.2 Pearson's Chi-Squared

The chi-square test of independences and the chi-square goodness-of-fit test are the two versions of Pearson's chi-squared test. It is denoted as X^2, it checks whether a variable distribution deviates from its actual distribution. If the test's value is little, it indicates that the observed data closely matched the expected data; if the value is large, it indicates that the observations did not

match the expectations. Showing the link between two category variables is mostly how it is used. With the above-described notation, Pearson's residual is stated as

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\eta}_i)^2}{n_i\hat{\eta}_i(1-\hat{\eta}_i)}$$

$$X^2 \sim X^2_{n-1}, \alpha \quad \alpha = significancelevel$$

3.3 Hosmer-Lemeshow statistic (HL)

Similar to the chi-square test, the Hosmer-Lemeshow (HL) test measures the goodness-of-fit for logistic regression by splitting the data into smaller groups. This test is limited to response variable that are binary. To address the inefficiencies with the chi-square test, the HL test was created. By sorting the estimated probabilities from a model and organizing them into g groups ideally, each group having roughly the same number of members and tied values grouped into the same group. The HL statistics is express as

$$HL = \sum_{g=1}^{G} \frac{\left\{\sum_{i=1_g} (y_i - \hat{\eta}_i)\right\}^2}{n_g\overline{\eta}_g(1-\overline{\eta}_g)}$$

Where n_g is the number of observations in the g^{th} group

$$\overline{\eta}_g = \frac{\sum_{i\varepsilon\varphi_g} \hat{\eta}_i}{n_g}$$

[8], showed that *HL* has an approximate χ 2 *(g* - 2*)* distribution.

3.4 Likelihood Ratio Test

The premise that a model performs best when there are two or more nested model is tested using the likelihood ratio test, also known as the likelihood-ratio chi-square test. The likelihood ratio test is used to determine which model is better when there are two or more built for same purpose using the same data, each with unique quality. The model with the lowest likelihood function will be most suitable. Because these functions can be difficult to compute, statistical software is required for used. This test basically examines how well two models fit each other over time. According to the null hypothesis, the optimal model has fewer variables. If the test statistic is high, the hypothesis is rejected, and the model with the most parameter is therefore deemed to be the best. It is easier to determine the likelihood ratio test when one knows the likelihood of any two models; the procedure is as follows;

$$Z = -2\ln\left(\frac{L(k_1)}{L(k_2)}\right) = 2\left(\log L(k_2) - \log L(k_1)\right)$$

Where $L(k_1)$ = Likelihood Function of mode1

$L(k_1)$ = Likelihood function of model2

$\log L(k_i)$ = natural log of the likelihood functions

The test statistics obtained is chi-square distributed, with degrees of freedom equivalent to the number of bounded parameters.

3.5 McFadden's Pseudo-R squared

Many researchers have proposed various quantities for logistic regression that can inherit the properties of the R-square of linear regression in an attempt to obtain a copy of that R-square. One of such quantity is the McFadden R-square statistics and its pseudo R^2 values, which are reported by stata. Given that regressions are fitted using the maximum likelihood approach, the McFadden Statistics are determined as follows:

$$R^2_{MC.Fadden} = 1 - \frac{Log(k_c)}{Log(k_{Null})}$$

Where

(k_c) =Maximum Likelihood value for the current model

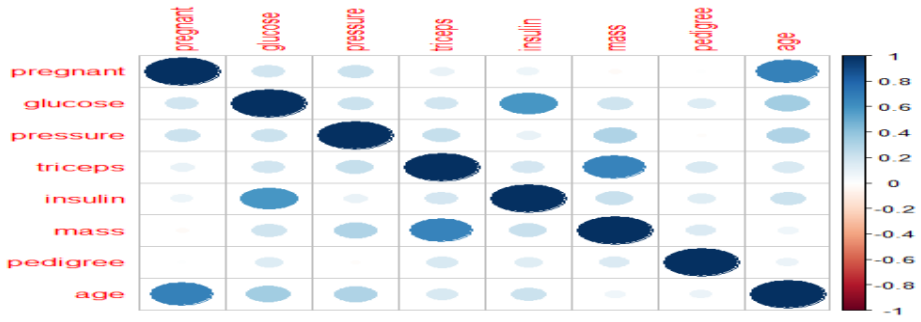
(k_{Null}) = Maximum Likelihood value for the Null model (The model with only an intercept and no covariates)

The metric goes from 0 to slightly about 1, with values closer to zero indicating that the model has little predictive capability

4.0 RESULTS

We used the built-in R data in the PimaIndiansDiabetes2 library caret, which contain 8continuous random variable and 1 categorical variable, to study the response variable Diabetes on the 8 covariates. We develop three distinct models and evaluate each one's goodness-of-fit in order to determine which model best captures the connection between the variables and the response variable. The missing variable was removed from 768 data in PimaIndiansDiabetes2, leaving 392 observations overall that were used in the analysis. Two sets of data were created from this data: the first group included 80% (314) of the observations used to develop the model, and the second group included 20% (78) of the observations used to predict and determine the

accuracy of the model. All of the variables in our data are included in the first model; only the variables that were found to significantly contribute to the response variable in the first model are included in the second model; and the variables that have been found to independently contribute to or promote the response variable are included in the third model.



Data Visualization

Figure 1: correlation graph of variables

Using R, a correlation plot was created using a dot representation: a larger dot indicates a higher correlation. Blue indicates a positive correlation and red a negative correlation. Every variable in the matrix is associated to itself, resulting in symmetry. The data reveals that there is a correlation between age and being pregnant, between insulin and glucose, and between mass and triceps.

Table 1: An Overview of Maximum Likelihood Approximations

Parameter	Estimates	Std. Error	Z-Score	P(Z-Score)	Odds Ratio
Intercept	-1.004	1.218	-8.246	$2.\times10^{-16}$	0.00004
Pregnant	0.0822	0.0554	1.482	0.1383	1.0856
Glucose	0.0383	0.0058	6.635	3.24×10^{-11}	1.039
Pressure	-0.0014	0.0118	-0.12	0.9045	0.9986
Triceps	0.0112	0.0171	0.657	0.5113	1.0113
Insulin	-0.0008	0.0031	-0.632	0.5276	0.9992
Mass	0.0705	0.0273	2.58	0.0099	1.0731
Pedigree	1.141	0.4274	2.669	0.0076	3.1296
Age	0.034	0.0184	1.847	0.0647	1.0345

The plot of the response variable (diabetes) on each of the eight variables in the data is show in table 1. The greatest estimations of the beta coefficients and their significance level are displayed

in the summary above. The estimation of the relationship between each predictor variable and the response variable is provided by the intercept and beta coefficient in column 2. The coefficient of estimates' standard error indicates the accuracy of the estimations; the larger the standard error, the less certain we are of the estimate. The Z-score, or Wald statistic of the estimated coefficients, is calculated by dividing the standard error of (column3) by the coefficient estimate of (column2). The variables' level of importance is shown by the P(z-score); the smaller the values, the more important the variable is to the model. As can be seen from the table, the response variables were significantly influenced by only three of the eight predictor factors. These consist of pedigree, mass and glucose. The positive coefficient of glucose (b=0.0383) indicates that there is a positive correlation between glucose levels and the likelihood of having diabetes. Additionally, the negative coefficient of insulin, b=(-0.0008), shows that a rise in insulin is linked to a fall in the likelihood of having diabetes. The correlation between a predictor (xi) and the response (y) is measured by the odd ratio. The ratio between an event that happens when a predictor is present and an event that happens when the predictor variables are absent is compared. For instance, the odds of becoming pregnant are (1.0685), meaning that the probability of having diabetes will rise by (1.065) for every unit increase in pregnancy concentration.

Table 2: An Overview of Maximum Likelihood Approximations

Parameter	Estimates	Std. Error	Z-Score	P(Z-Score)	Odds Ratio
Intercept	-8.4603	0.6677	-12.670	2.0×10^{-16}	0.0002
Glucose	0.0379	0.0035	10.916	2.0×10^{-16}	1.0386
Mass	0.0809	0.0142	5.690	1.27×10^{-08}	1.0844
Pedigree	0.8675	0.2962	2.929	0.0034	2.3809

Some factors, such as Pregnancy, insulin, triceps and pressure, are not statistically relevant to the model, as can be seen from the result in table 1. As a result, maintaining them in the model may lead to overfitting, thus, the need to remove them. Table 2 is created, which includes a summary of the model with fewer predictors and variables that are statistically significant for being associated with diabetes in the model 1.

Table 3: An Overview of Maximum Likelihood Approximations

Parameter	Estimates	Std. Error	Z-Score	P(Z-Score)	Odds Ratio
Intercept	-9.0149	0.8129	-11.090	2.0×10^{-16}	0.0001
Glucose	0.0346	0.0036	9.697	2.0×10^{-16}	1.0352
Mass	0.0886	0.0155	5.732	9.9×10^{-09}	1.0927
Pedigree	0.9233	0.3040	3.037	0.0024	2.5176
Age	0.0345	0.0085	4.066	4.78×10^{-5}	1.0351
Pressure	-0.0074	0.0085	-0.874	0.3821	0.9926

One of the factors from our correlation chart in figure 1 that was shown to be significantly associated was removed to create table 3. Our chart shows that there are strong correlations between age and pregnancy, glucose and insulin, and mass and triceps. Based on these correlations, we choose to exclude one and select the other, leading to creation of model 3. The model summary is shown in table 3.

Table 4: A Summary of Model Fit.

Parameter	**Model 1**		**Model 2**		**Model 3**	
	Value	**DF**	**Value**	**DF**	**Value**	**DF**
Null Deviance	498.10	391	974.75	751	931.94	723
Residual Deviance	344.02	383	729.76	748	685.66	718
AIC	362.02	-	737.76	-	697.66	-
Mean	0.6097	-	0.7832	-	0.6050	-

The null deviance, residual deviance, and Akaike Information Criteria (AIC) are three factors that R provides in order to examine model fit in the model summary. The residual deviance indicates how well the response variable is predicted when the predictor variables are included, whereas the null deviance indicates how well the response variable is predicted by a model that only includes the intercept. The model is better the larger the difference between the null and residual deviances. For instance, in model 1, the null deviance is 498.10. The residual deviance, or deviation after the addition of the eight predictor variable is 344.02. keeping in mind that the deviation qualifies the poor fit; the lower the value, the more accurate the model, while helpful in model comparisons, the Akaike Information Criterion (AIC) cannot be interpreted in isolation. The model with the lowest (AIC) is thought to be the most fit when there are multiple similar models. Therefore, model 1 will be regarded as the most fit model, among the aforementioned models. The percentage of observations that have been accurately classified serve as a proxy for model accuracy, while the percentage of observations that have been incorrectly classified is known as the classification error. The mean in table 4 provides this forecast. It is evident from the table 4, that model 2 has the greatest mean, having a misclassification error rate of 22%, and classification accuracy is a respectably 78%.

Table 5: Some Statistic of Goodness-of-fit.

	Model 1			**Model 2**			**Model 3**		
Statistic	**Value**	**DF**	**P-value**	**Value**	**DF**	**P-value**	**Value**	**DF**	**P-value**
Likelihood Ratio Test (LR Test)	-172.0	-	-	-181.9	-5	0.0014	-173.6	2	0.3601
Pseudo (R^2)	0.3093	-	-	0.2698	-	-	0.3029	-	-
Hosmer-Lemeshow	392	8	2.2×10^{-16}	392	8	2.2×10^{-16}	392	8	2.2×10^{-16}

Test ($\hat{C}$)									
Chi-square Test (X^2)	344.02	383	-	363.70	388	0.0014	347.23	386	0.3601

To decide which of the three models provides the best match, the likelihood ratio test is used to compare them. It compares the likelihood of the data under a model with full predictors and the likelihood of the data under a model with fewer predictors. A P-value for the total model fit statistic less than 0.05 would force us to reject the null hypothesis. The model with lower likelihood is referred to as less fit. Table 5 indicates that model 2 has a lower fit Value of (-181.9). This difference is statistically significant when compared to model 1 with a P-value of (0.0014), which is less than 0.05. Utilizing the likelihood ratio test, it is determined that there is no statistically significant difference between model 1 and 3 (P-value=0.3601). The Pseudo (R2) test has a value between 0 and 1, where a value closer to zero indicates a less fit model and a value closer to one indicates a very fit model. The values for all three of the models are skewed towards zero, suggesting poor fit, as can be seen from row 2 in table 5. Model 1 has the highest value, (0.3093), indicating the best match. Given the values for the three models are identical, the Hosmer-Lemeshow test yields a P-value (2.2×10^{-16}) of less than 0.05, which indicates that the models are ill-fitting and that there are no significant difference between them. When model 1 is compared to other models, their difference is statistically significant with model 2 and not significant with model (3). The chi-square test looks to be similar to the likelihood ratio test. Model (1) has the lowest value, indicating the most fit. The data suggest that model 1 is superior to the other two models, and that there is no significant difference between model 1 and 3.

5.0 Conclusion

Fitting a logistic regression model and identifying which model best fits our data is our main goal. Using the PimaIndiansDiabetes2, we were able to create three distinct models based on highlighted features. Of the two reduced models, model1, which contain every variable in the data set, seemed to suit the data the best. This was the case both when the goodness of fit test statistic and the model summary statistic were used (the alkaike, residual deviance, and null deviance). When comparing model fit, the information criterion from the model summary of the fit proved to be correct; model 1 hand the lowest values of these parameter in contrast to the other two models. These parameters drawback was that they concealed the models shared characteristics from one another. For example, model1 and model 3 are similar, yet the goodness-of-fit summary statistics for the models does not show this. The goodness of fit statistic allowed us to quantify the model's accuracy as well as the differences between the models. We could then assess whether the discrepancies between the models are significant enough to refute our hypothesis. The findings of our analysis using the likelihood ratio test, the pseudo (R2) test, and the chi-square test are similar and support the model summary that indicate model 1 fits the data the best out of the three models. We also review that model 1 and model 3 are nearly identical; the difference between them was tested and was found not to be statistical significance. Upon evaluation, the difference between models 1 and 2 was shown to be statistically significant. The test yielded the same results for all three models, where the difference between them cannot be determined. However, the hosmo-lemeshow test provided a

very low P-value indicating the presence of poor fit in all three models. The results of the likelihood ratio test and pseudo (R2) test all point to the model's poor fit and complied with the finding of the chi-square test.

6.0 Recommendations

Based on our investigation, we will advise against utilizing the summary of fit statistic generated by the R statistical package during model construction and instead use the goodness of fit statistic, such as the likelihood ratio test, Pseudo (R2) test, chi-square test, etc., when evaluating the fit of the model. When assessing the accuracy of a model, we will advise using the Hosmo-Lemeshow test because, in comparison to other three test statistic employed in this study (likelihood ratio test, pseudo (R2) test, chi-square test), it yield a Pvalue that was most significant. When comparing models, we will advise using the chi-square test or likelihood ratio test.

References

[1] Alexander Henzi, Marius Puke, Timo Dimitriadis, Johanna Ziegel, (2023). A Safe Hosmer-Lemeshow Test, The New England Journal of Statistics in Data Science, 10.51387/23-NEJSDS56, (1-15).

[2] Bagley SC, White H, Golomb BA, (2001). Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 54(10):979-85. doi: 10.1016/s0895-4356(01)00372-9. PMID: 11576808

[3] Bertolini G., D'Amico R., Nardi D., Tinazzi A., Apolone G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. Epidemiol Biostat.5(4):251-3. PMID: 11055275

[4] Daniel Fernández, Louise McMillan, Richard Arnold, Martin Spiess, Ivy Liu, (2022) Goodness-of-Fit and Generalized Estimating Equation Methods for Ordinal Responses Based on the Stereotype Model, Stats, 10.3390/stats5020030, **5**, 2, (507-520).

[5] Horton NJ, Bebchuk JD, Jones CL, Lipsitz SR, Catalano PJ, Zahner GE, Fitzmaurice GM. (1999). Goodness-of-fit for GEE: an example with mental health service utilization. Stat Med. 18(2):213-22.

[6] Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model, Communications in Statistics, A10, 1043—1069

[7] Hosmer, D. W. and Lemeshow, S. (1989). Applied Logistic Regression, Wiley, New York,

[8] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 16(9):965-80. doi: 10.1002/(sici)1097-0258(19970515)16:9.

[9] Kleinbaum, D. G., & Klein, M. (2010). Logistic regression; statistics for biology and health,(3rd ed.). New York, NY: Springer-Verlag New York Inc

[10] Le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary data based on smoothing residuals', Biometrics, 47, 1267—1282

[11] Lemeshow, S., Hosmer, D. W. (1982). A review of goodness-of-fit statistics for use in the development of logistic regression models. American Journal of Epidemiology 115(1):92–106.

[12] Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013). Applied Logistic Regression. 3rd Edition, John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9781118548387

[13] Peng, C. J., Lee, K.L., & Ingersoll, G.M. (2002). An introduction to logistic regression analysis and reporting. The Journal of Educational Research, 96(1), 3-14.

[14] Pulkstenis, E., Robinson, T. J. (2002). Two goodness-of-fit test for logistic regression models with continuous covariates. Statistics in Medicine 21(1):79–93

[15] Shen-Ming Lee, Phuoc-Loc Tran, Chin-Shang Li, (2022). Goodness-of-fit tests for a logistic regression model with missing covariates, Statistical Methods in Medical Research, 10.1177/09622802221079350, **31**, 6, (1031-1050).

[16] Tetrault JM, Sauler M, Wells CK, Concato J, (2008). Reporting of multivariable methods in the medical literature. J Investig Med.;56(7):954-7. doi: 10.2310/JIM.0b013e31818914ff. PMID: 18797413.

[17] Eberhardt, L.L. and Breiwick, J.M. (2012) Models for Population Growth Curves. ISRN Ecology, 2012, Article ID: 815016. http://dx.doi.org/10.5402/2012/815016