American Journal of Applied Mathematics and Statistics, 2018, Vol. 6, No. 4, 126-134
Available online at http://pubs.sciepub.com/ajams/6/4/2
©Science and Education Publishing
DOI:10.12691/ajams-6-4-2

Evaluating Methods of Assessing “Optimism” in Regression Models

Daniel Thoya[1,*], Antony Waititu[1], Thomas Magheto[1], Antony Ngunyi[2]

[1]Department of Statistics and Actuarial Science, Jommo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]Department of Statistics and Actuarial Science, Dedan Kimathi University of Science and Technology, Nyeri, Kenya
*Corresponding author: daniel.thoya@yahoo.com

Abstract The purpose of this study was to evaluate the methods used to assess “optimism” in regression models. Particularly, focus was on the use of pseudo R^2 values of cox &snail and the Nagelkerke to identify the best statistic for measuring “optimism” in regression models, measure model performance and determine the relationship between “optimism” and over fitting. Different underlying data sets assume different models that fit their data accurately. However, the fitted regression models usually fit the data they are based on better than new data. This is what we call ‘optimism’. Specific focus will be on determining the best statistic for measuring optimism in regression models, assess model performance using ‘optimism’ through cross-validation and also determining the relationship between optimism and over fitting of regression models. The study focus on three models (Cox-regression, Logistic regression and Linear Regression) and bootstrap procedure was used.

Keywords: *optimism, pseudo-r-square, cox & snell, Nagelkerke*

Cite This Article: Antony Waititu, Thomas Magheto, and Antony Ngunyi, “Evaluating Methods of Assessing “Optimism” in Regression Models.” *American Journal of Applied Mathematics and Statistics*, vol. 6, no. 4 (2018): 126-134. doi: 10.12691/ajams-6-4-2.

1. Introduction

Regression models are powerful tools that are frequently used variously by both researchers and scholars in studies [1] in his work on regression model and forecasting he found out that regression models provided the analysis and estimation of parameters for forecasts. Regression models can also handle partially observed (censored) responses [2] in his study on survival analysis censoring marked the difference between other statistical analysis and survival analysis. A fitted regression model will fit the data it was based on better than any other new data [3] in studying the procedure when adjusting for optimism and over fitting in measures of predictive ability using bootstrapping prognostic models performed differently with test data from the training data. It is a requirement for analysts to create prognostic models that have the ability to reflect accurately the patterns that exist in different underlying data sets.

1.1. Measuring Prediction Error

Usually it is paramount for a researcher to assess the quality of every model before using it in any data set. By virtue of natural grounds, common practice, most models are highly optimized for the data in which they were trained [4] when investigating the role of noise variables in model building the characteristics of the training data plays a major role in the complexity and overall performance of any model. Expected errors exhibited on new data will always be higher than the expected errors on training data [5] used model validation in studying optimism for training error where it was discovered that test data had higher values of optimism whenever used to test model performance.

$$\begin{aligned}&\textit{Trueprediction error}\\&= \textit{training error} + \textit{training optimism.}\end{aligned}$$

When the modeler is more optimistic, then the training error will be better compared to the value of the true error.

1.1.1. The Danger of Over Fitting

From this perspective, a model that minimizes training error will automatically reduce the predication error for a new data set. It is therefore recommendable to ignore the distinction between training error and the prediction error to allow for model selection [6] when lying down the criteria for model selection the role of training and prediction errors were observed in the overall method assumed by the modeler and hence their assumption. The reason being that optimism is a function of model complexity, as complexity increases so does optimism. The relationship for the true prediction error takes the following form;

$$\begin{aligned}&\textit{Trueprediction error}\\&= \textit{Training Error} + f(\textit{Model complexity}).\end{aligned}$$

Model complexity increases when the number of parameters is increased and this will ensure the model does a better job for training data, which is a fundamental property of statistical models [7] when studying mathematical and computer modeling it was seen that lack of methodology in choosing the best model fit stems from a poor understanding on the ways in which the study rely on the particular model.

1.1.2. Data Reduction

Excellent tools of prediction have always been models well fit [8] when studying level of crime in the city of Salinas, the absence of statistical tools in predicting was a major blow until when regression models were applied. The understanding is vested on the total predictor degrees of freedom (d.f) p, as the number of parameters examined during analysis.

Use of informal analytical methods like graphical work makes one unable to determine the value of p. Instead it is paramount to estimate the effective number of parameters considered according to the flexibility of fits considered at the initial stages of analysis. The predictor degrees of freedom p, is the number of parameters allowed for consideration, in other words, it is the number of regression coefficients estimated without algebraic restrictions. It is suggested that as a rough rule of thumb, in order for one to validate a new sample using predictive discrimination, the predictor degree of freedom should not be more than $m/10$, where m is the number of uncensored event times in the training sample. When we consider binary outcome, m is the number of outcome in the less frequent event. If the quantity $p \geq m/10$ then the analyst has to choose a data reduction technique that takes care of this and shrinkage is the best method.

1.2. Problem Statement

Assessing “Optimism” in regression models has been approached differently using different methods. There is need to evaluate some of these methods so that a better one is identified. However, these fitted regression models usually fits the data they are based on better than any new data. This is what we call ‘optimism’ of the model.

1.3. Justification of the Study

The main goal of this study is to assess the methods for evaluating “optimism” in regression models and the first step is to identify a statistic for measuring “optimism” focusing on pseudo R^2 values of Cox&Snell and the Nagelkerke. Using these statistics, the model performance would also be evaluated and again used to determine the relationship between “optimism” and over fitting. These methods have varied degree of measure and assessment. The tragedy is identifying the best out of these. The most important of all these techniques is the ability of a technique to ascertain the estimated model performance and the model’s variance and stability.

1.4. Objectives of the Study

1.4.1. General Objective

The main objective of this study is to evaluate the methods of assessing ‘optimism’ in regression models.

1.4.2. Specific Objectives

i. To determine the best statistic for assessing “optimism” in regression models
ii. To assess model performance using ‘optimism’ through cross-validation.
iii. To determine the relationship between “optimism” and over fitting of regression models.

1.5. Hypotheses of the Study

This study seeks to evaluate the methods of assessing “optimism” in regression models.

The study seeks to test the following hypotheses;

i. Null hypothesis: there exists no statistic for assessing “optimism” in regression models
ii. Null hypothesis: there is no significant difference in performance among the three models (cox, logistic and linear regression models)
iii. Null hypothesis: there is no relationship between optimism and over fitting of regression models.

1.6. Significance of the Study

The common goal for every model builder, researchers, scholars and academicians is the zeal to come up with prognostic models reliable and accurate for training and unforeseen data [9] in studying prognostic models in chronic liver diseases, the prognostic structure in data can be studied in many different ways however the most recent and accurate method was the use of regression models, the cox regression models.

1.7. Scope of the Study

The study purports to rivet on assessment of optimism exhibited by cox regression, logistic and linear regression models. Bootstrapping resampling technique was used in evaluating the two pseudo-R-square measures of Cox&Snell and Nagelkerke for assessing optimism with regard to these three categories of regression models. This is due to the nature and wide use of these models by both scholars and researchers in different fields and applications.

2. Literature Review

2.1. Introduction

In this chapter the study gave a detailed review of past studies on regression models, different types of regression models, more specifically cox regression models logistic and linear regression models.

2.2. Types of Regression Models

In predictive modeling, most people have considered linear and logistic regression models as the first algorithms [10] listed logistic and linear models as the most used models in data modeling and research. However it is paramount to understand the existence of several types and forms of regression models that have use and applications in different types of research data [7] tried to use linear regression to model binary data but he resorted to logistic regression when he could not infer.

2.2.1. Linear Regression

Among the modeling techniques, linear regression occupies the first position. The metric R-Square can easily be used for model performance evaluation. The variance of coefficient estimates can increase due to the effect of multicollinearity, this makes the estimates very sensitive to minor changes in the model, prediction errors pronounced optimism [11] in his study on multiple regression, he found out that multicollinearity occurs when independent variables in a regression model are correlated. For the simple linear regression model we assume a model of the form;

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

Where β_0 and β_1 are two unknown constants that represent the intercept and the slope. They are also known as the coefficients or parameters and ε are the error term.

If we consider some estimates $\overset{\wedge}{\beta_0}$ and $\overset{\wedge}{\beta_1}$ of the model coefficients parameters, we predict the future using;

$$\hat{y} = \overset{\wedge}{\beta_0} + \overset{\wedge}{\beta_1} x,$$

Where $\hat{y}$ indicate a prediction for Y on the basis of $X = x$.

Multiple linear regressions

The expression for the multiple linear regressions assumes the following form;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

$\beta_{j's}$ give the average effect on Y of unit increase of the independent variables $X_{j's}$, holding all other factors constant.

Interpretation of the regression coefficients

Each of the coefficients can be estimated and tested separately. Correlations amongst predictors cause problems. When the predictors, $X_{j's}$ change, then interpretation become hazardous since everything else changes. For observational data, claims of casualty should be avoided.

Obtaining the likelihood function for linear regression model

The simple linear regression model states that the errors are independent and normally distributed with mean 0 and variance σ^2.

The linearity condition; $Y_i = \alpha + \beta(x_i - \overline{\chi}) + \varepsilon_i$

Therefore implies

$$Y_i \sim \mathrm{N}(\alpha + \beta(\mathrm{x_i} - \overline{\chi}), \sigma^2)$$

Therefore the likelihood function is;

$$L_{Y_i}(\alpha, \beta, \sigma^2) = \coprod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(Y_i - \alpha - \beta(\chi_i - \overline{\chi}))^2}{2\sigma^2} \right]$$

This can be written as;

$$L = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha - \beta(\chi_i - \overline{\chi}))^2 \right].$$

Taking the log of both sides; we obtain;

$$LogL = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha - \beta(\chi_i - \overline{\chi}))^2.$$

This is the log-likelihood function of the linear regression model.

2.2.2. Logistic Regression

The application of logistic regression is strictly on binary data (0/1, True/False, Yes/No). It is of great importance to note that the values of the response variable range from 0 to 1. The ideal equation for logistic regression is as shown below;

$$odds = p/(1-p)$$
$$= \frac{probability\ of\ eventoccurence}{probabilitry\ of\ eventnotoccuring}$$
$$\ln(odds) = \ln(p/(1-p))$$
$$\log it(p) = \ln(p/(1-p))$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 .. + \beta_k x_k.$$

Because logistic regression predicts probabilities rather than classes, we can fit it using the likelihood function. For each training data points, we have a vector of features χ_i and the observed class, y_i. The probability of that class is either p if, $y_i = 1$ or $1 - p$ if $y_i = 0$.

The likelihood is then;

$$L(\beta_0, \beta) = \coprod P(\chi_i)^{y_I} (1 - P(\chi_i))^{1-y_i}.$$

Taking the log of both sides, the resulting equation becomes;

$$L(\beta_0, \beta) = \sum y_i \log p(\chi_i) + (1 - y_i)\log 1 - p(\chi_i)$$
$$= \sum_{i=1}^{n} \log 1 - p(\chi_i) + \sum_{i=1}^{n} y_i \log \frac{p(\chi_i)}{1 - p(\chi_i)}$$
$$= \sum_{i=1}^{n} -\log 1 + e^{\beta_0 + \chi_i \beta} + \sum_{i=1}^{n} y_i (\beta_0 + \chi_i . \beta).$$

This is the log likelihood function for a logistic regression model.

2.2.3. Cox Regression Model

Cox regression model provides an estimate of the treatment effects on survival after adjustment of the explanatory variables. The regression employed by cox is proportional hazards regression analysis. The cox PH model takes the following form;

$$h(t, X) = h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i X_i \right),$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_P)$ are the predictor/explanatory variables. Explanation of the formula; product of two quantities; $h_o(t)$ Is called the baseline hazard, exponential sum of β_i and X_i

$X^{'}s$ Zero reduces to baseline hazard. The baseline hazard is an unspecified function.

Important properties of the cox PH formula;

The baseline hazard $h_o(t)$ does not depend on X but on t

The exponential involves the $X^{'}s$ but not t, the $X^{'}s$ are time dependent. The proportional hazard function follows. There exist a number of reasons that make the cox PH model popular;

1. Robustness; the cox model is a "safe" choice of model in most modeling situations that researchers can go for.
2. The model form; $h(t,X) = \underbrace{h_o(t)}_{\geq 0} \times \underbrace{\exp\left(\sum_{i=1}^{p} \beta_i X_i\right)}_{\geq 0}$ makes the estimated hazards to be always non-negative.
3. $h(t,X)$ and $s(t,X)$ can be estimated for a cox model using a minimum of assumptions.
4. The cox model is preferred to logistic model in survival data modeling because logistic model ignores survival times and censoring information.

Computing the hazard ratio;

The hazard ratio is defined as;

$$HR = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)},$$

where $X^* = (X_1^*, X_2^*, X_3^*, \ldots., X_P^*)$ and

$$X = (X_1, X_2, X_3, \ldots, X_P).$$

Obtaining the likelihood function under censored data

Assuming we have n units whose lifetimes are governed by a survivor function $S(t)$ with associated density $f(t)$ and hazard $\lambda(t)$. Suppose unit i is observed for a time t_i, if the unit died at t_i its contribution to the likelihood function is the density at that duration which is the product of the survivor and the hazard function;

$$L_i = f(t_i) = S(t_i)\lambda(t_i).$$

If the event is still alive at time t_i then under non-informative censoring, the life time will exceed t_i with probability given as; $L_i = S(t_i)$ which becomes the contribution of the censored observation to the likelihood. Let d_i be a death indicator taking the value i if one unit died and zero otherwise, then the likelihood will be written as;

$$L = \prod_{i=1}^{n} L_i = \prod_{i} \lambda(t_i)^{d_i} S(t_i).$$

Taking the log and recalling the initial expression that links the survival function $S(t_i)$ and the cumulative hazard function $\Lambda(t)$, we obtain the log likelihood function for the survival model given as;

$$\log L = \sum_{i=1}^{n} \{d_i \log \lambda(t_i) - \Lambda(t_i)\}.$$

Interpretation of the Hazard ratio

The hazard for one individual is divided by the hazard for a different individual. During interpretation, one usually wants $HR \geq 1$. That is; $\hat{h}(t, X^*) \geq \hat{h}(t, X)$. This therefore means

X^*: the group with larger hazard and X the group with the smaller hazard.

2.3. Model Validation and Assessment

Scrutiny of the manifest accuracy of a multivariable model is not useful when using training dataset [12] when applying parametric spectral analysis to multichannel event related potentials during cognitive experiments found out that model assessment was crucial for proper data processing and prediction.

3. Research Methodology

3.1. Introduction

This chapter gives stringent interrogatory features that were palpable in the study.

3.2. Design of the Study

The study design was simulation. The simulated data formed our population and original data from where the bootstrap samples were obtained. The discrimination statistics which are the two pseudo R-square values, the cox&snail and the Nagelkerke (Cragg & Uhler's) r-squared were be compared.

The cox&snell;

$$R^2{}_{cox\&snell} = 1 - \left\{\frac{L(M_{INTERCEPT})}{L(M_{FULL})}\right\}^{\frac{2}{N}}$$

value together with the Nagelkerke:

$$R^2_{Nagel\ker ke} = \frac{1 - \left\{\frac{L(M_{INTERCEPT})}{L(M_{FULL})}\right\}^{2/N}}{1 - L(M_{INTERCEPT})^{2/N}}$$

were obtained from the simulated data

Where;

$L(M_{INTERCEPT})$ is the likelihood of the intercept model (model without predictors)

$L(M_{FULL})$ is the likelihood of the model with parameters

N is the number of observed data sets

3.2.1. General Bootstrap Procedure for Measuring Optimism

1. The first step is to develop the model using all the n subjects and carry out any testing that may be necessary. Let $R^2{}_{cox\&snellapp}$ denote the apparent

$R^2{}_{cox\&snell}$ from the model formed. This is the scaled chi-square computed on the same sample from which the fit has been derived from.

2. It follows that we generate a sample of size n with replacement from the original sample considering both predictors and response.
3. From the bootstrap sample compute the apparent $R^2{}_{cox\&snellapp}$ from this model and denote it as $R^2{}_{cox\&snellboot}$
4. Let $R^2{}_{cox\&snellorig}$ denote the apparent $R^2{}_{cox\&snell}$ from the original dataset. Then 'freeze' the developed model and evaluate its performance on the original dataset.
5. Compute the optimism by $R^2{}_{cox\&snellboot}$ - $R^2{}_{cox\&snellorig}$
6. Steps 2 to 6 are then repeated 100-B times

The corrected bootstrap performance of the original stepwise model is $R^2{}_{cox\&snellorig}$ -0 this difference is closer to the unbiased estimate of the expected value of the external predictive discrimination that generated $R^2{}_{cox\&snellorig}$ which is an estimate of internal validity penalizing over fitting.

Using data from a hypothetical population, simulations were conducted at individual setting of variables. Averages of performance measure were taken over *B* repetitions for a chosen *m* number events on the predictor variables.

From the model for obtaining optimism;

$$Optimism \\ = sample\ measure\ statistic - apparent\ measure\ statistic$$

we obtain optimism for the two statistics for the three models as follows;

$$Optimism1 = AverageR^2{}_{cox\&snellboot} - R^2{}_{cox\&snellorig}$$

$AverageR^2{}_{cox\&snellboot-cx} - R^2{}_{cox\&snellorig-cx}$ For the cox-regression model

$AverageR^2{}_{cox\&snellboot-\lg t} - R^2{}_{cox\&snellorig-\lg t}$ - For the logistic regression model

$AverageR^2{}_{cox\&snellboot-\ln} - R^2{}_{cox\&snellorig-\ln}$ For the linear regression model

Optimism 2

$AverageR^2_{Nagel\ker keboot}$ - $R^2_{Nagel\ker ke-orig}$

$AverageR^2_{Nagel\ker keboot-cx}$ - $R^2_{Nagel\ker keorig-cx}$ For the cox-regression model

$AverageR^2_{Nagel\ker keboot-\lg t}$ - $R^2_{Nagel\ker keorig-\lg t}$ For the logistic regression model

$AverageR^2_{Nagel\ker keboot-\ln}$ - $R^2_{Nagel\ker keorig-\ln}$ For the linear regression model

It is from these measures that the least optimism was be obtained and hence the one that was closer to zero was the best statistic.

3.2.3. To Assess Model Performance Using 'optimism' through Cross-validation

Using the optimal statistic lets denote this as C_{optim} obtained in procedure above between $R^2{}_{cox\&snell}$ and $R^2_{Nagel\ker ke}$ the values were compared across the three models. Build the original model from the training data and obtain the values of C_{optim} B Bootstrap samples serve as the testing sets for B -cox, logistic and linear models. For each of the model obtain the difference for the values of C_{optim} average them to get the value of "optimism". The model with its value of C_{optim} closer to zero will be regarded as the best performing under optimism.

3.2.4. Using the Number of Parameters to Determine the Relationship between "optimism" and over Fitting

Build models with at least three parameters and obtain the value of C_{optim} for each of the three categories of interest (cox, logistic and linear) regression models. From the models initially built, increase the number of parameters from three to four, five, six, seven and if possible eight parameters. Obtain the values of C_{optim}. Obtain the values of "optimism" and compare across the models with reference to the number of parameters modeled.

4. Results and Discussions

4.1. Data

The data for this study was obtained through simulations of hypothetical populations. For the cox regression and logistic models, at least two categorical variables formed part of the predictor variables. The package "simstudy" alongside "survival" and "BaylorEdPsych" in R were used.

4.1.1. Diagnostic Checks for Simulated Data

The study was mainly based on simulated data and therefore to ensure it was fit for use in the achieving the objectives of the study, it was exposed to a number of checks.

4.1.2. Trend Test for Survival Data

To perform a trend test, the cox model was fit with a factor predictor variable scored as 1, 2, 3… and later a post hoc trend test was conducted.

Table 1. Trend test for survival data

	coef	exp(coef)	se(coef)	z	p
factor (v4)1	0.369	1.446	0.199	1.86	0.063
factor (v4)2	0.916	2.500	0.225	4.08	4.5e-05
factor (v4)3	2.208	9.097	1.026	2.15	0.031
Likelihood ratio test=18.4 on 3 df, p=0.000356					

It can be noted that for the status variable status=0 which is implicitly part of the contrast. The full coefficient vector is (0, 0.369, 0.916, and 2.208) and the linear contrast zz is 0,1,2,3. Thus the data is fit for continued analysis.

4.1.3. Random Forests Model

These give the average probability of an event surviving within a given period of time as shown in the graph below;

The qqplots below provide a visual diagnistics for binary data and the simulated data for linear modelling.

4.2. Identifying the Best Statistic for Assessing “optimism” in Regression Models

The two inferential Pseudo r-square discrimination statistics were obtained forming our apparent values of interest. The values from the three models were as shown below;

From Table 2, the original values were 0.07347012 and 0.078304671 for the cox regression model for the cox&snell and Nagelkerke pseudo R-square values respectively. The average pseudo r-square values after running 800 bootstrap samples were 0.07998 and 0.85951625.This gave an optimism value of 0.00651 and 0.00765 for the cox&snell and Nagelkerke pseudo R-square values respectively. For the linear regression model the original values were 0.516475 and 0.711327 again for the cox&snell and Nagelkerke pseudo R-square values in that order while the average Pseudo R-square values after 800 bootstrap samples were 0.519114 and 0.672839 giving an optimism value of 0.00264 and 0.03849. Thirdly for the logistic regression model the original values were 0.018877 and 0.012092 while the average Pseudo R-square values were 0.121288 and 0.148447 yielding an “optimism” value of 0.10241 and 0.13636. The different values of “optimism” exhibited by the two statistics for the three models nicely confirms the fact that “optimism” in measure of predictive ability of a model is a function of the size of data set holding other things constant. In reference to [3] this does not rule out the fact that “optimism” is also a function of the complexity of the fitted model.

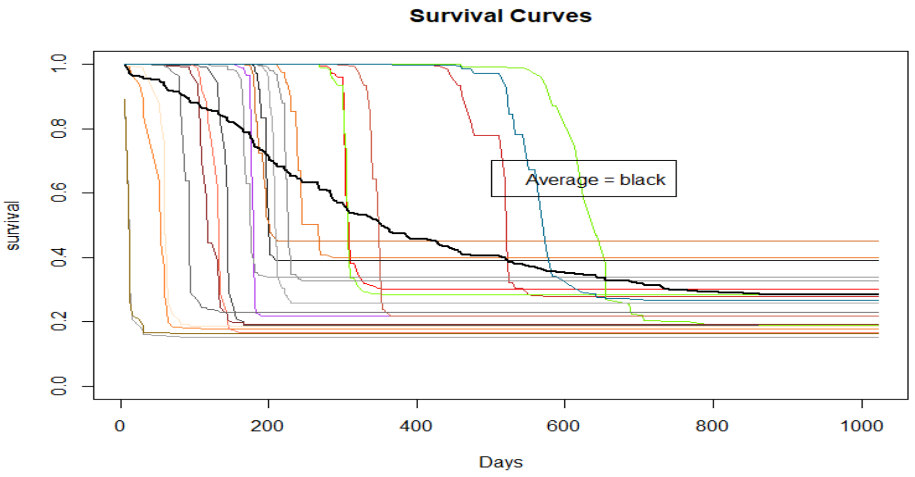

Figure 1. Diagnostic plot-survival data

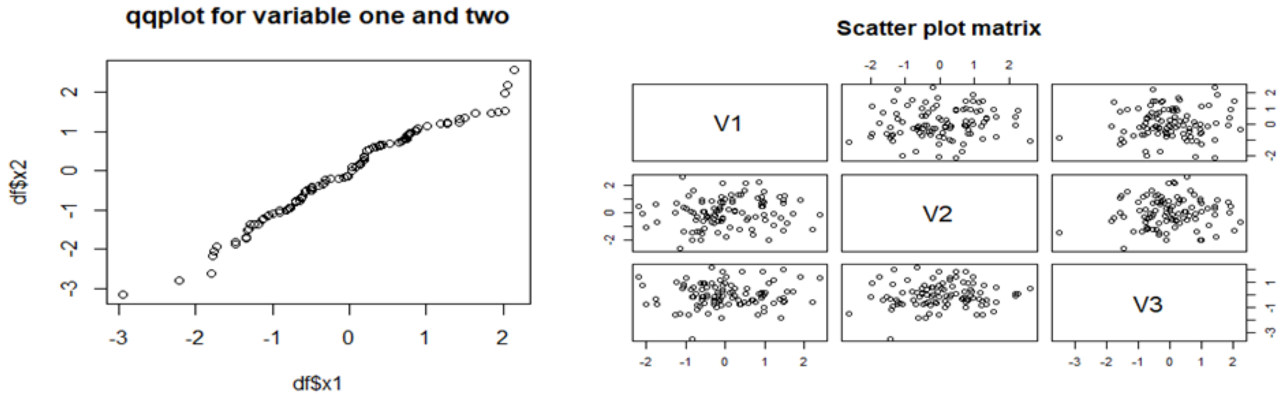

Figure 2. Diagnostic plots-binary and linear data

Table 2. Best statistic for assessing “optimism” in regression models

Model	Cox-regression model		Linear regression model		Logistic regression model	
Sample	$R^2_{cox\&snell}$	$R^2_{Nagel\ker ke}$	$R^2_{cox\&snell}$	$R^2_{Nagel\ker ke}$	$R^2_{cox\&snell}$	$R^2_{Nagel\ker ke}$
Original	0.07347012	0.078304671	0.516475332	0.711327013	0.018877944	0.012092061
sampleaverage	0.07998105	0.085951625	0.519114265	0.672838729	0.121288194	0.148447378
optimism	**0.006513**	**0.00765334**	**0.0026453**	**0.03849556**	**0.1024150**	**0.1363632**

The Cox & Snell pseudo R-square is given as;

$$R^2_{cox\&snell} = 1 - \left\{ \frac{L(M_{INTERCEPT})}{L(M_{FULL})} \right\}^{\frac{2}{N}}$$

While the Nagelkerke pseudo R-square takes the form;

$$R^2_{Nagel\ker ke} = \frac{1 - \left\{ \frac{L(M_{INTERCEPT})}{L(M_{FULL})} \right\}^{2/N}}{1 - L(M_{INTERCEPT})^{2/N}}$$

Where $L(M_{INTERCEPT})$ is the likelihood function for the intercept model i.e. the model with only the intercept variable. While $L(M_{FULL})$ is the likelihood function of the full model.

The ratio of the likelihoods shows the improvement of the full model over the intercept model (the smaller the ratio the better the improvement). If there are N observations in the data set, then $L(M)$ is the product of N such probabilities. Thus obtaining the n^{th} root of the product $L(M)$ provides an estimate of the likelihood of each Y value.

It is clear from the results above that indeed Cox &Snell pseudo r-square statistic has a good measure of "optimism".

4.3. Assessing Model Performance Using 'optimism' through Cross-validation

For the second objective on assessing model performance using "optimism" through cross validation, the focus was on the optimal statistic from the first objective and this was the Cox &Snell pseudo r-square statistic. The simulated data was partitioned into two; the training and the testing data set. Seventy five (75%) percent was used for developing (training) the model while twenty five percent (25%) was used to test and validate the model. The optimism values for the three models were obtained as shown in the table below; linear regression model had the lowest "optimism" value of 0.04438 followed by cox-regression model with an "optimism" value of 0.06473 and coming third was logistic regression model with an "optimism" value of 0.15682.

Table 3. Model performance "optimism" using cross validation

Model	Training	Testing	"optimism"
Cox-regression model	0.05417	0.11890	**0.06473**
Logistic regression model	0.44214	0.59895	**0.15682**
Linear regression model	0.02032	0.06470	**0.04438**

This means therefore linear regression models perform better in prediction compared to cox and logistic regression models. According to Oredein et al, (2011) model validity is the reasonableness and stability in performance on prognostic measures of interest. These results show that indeed the value of "optimism" can be used to measure model performance under cox&snell pseudo r-square statistic.

4.4. Determining the Relationship between "optimism" and Over Fitting of Regression Models

To achieve this objective, the study employed two strategies that influence prediction and performance of prognostic models. These were sample size and the number of predictor variables.

4.4.1. Determining the Relation between "optimism" and Sample Size

Boot strap samples of different sizes were drawn and the size of "optimism" using the Cox & Snell statistic determined. These were compared across the three models; cox regression model, logistic model and linear regression model. The results were as shown in the table below; small sample sizes have low "optimism" while large samples experience increasing "optimism" as can be seen for n=400, the value of optimism is 0.00805740 while for n=3000, 'optimism' is 0.0520398 for the cox regression model, similarly for the logistic regression models and linear models, "optimism" increases with increase in sample size. This confirms the Peduzzi and Concato (1995) 5-10 events per variable rule that indeed it results to small sample sizes leading to over fitting and optimism. The correlation between sample size and optimism for the three models is a positive one increasing with an increase in sample size. Correlation between sample size and optimism for the cox regression model is 0.3696021, for logistic regression we have 0.4388737 and 0.6382342 for the linear regression model.

Table 4. Relationship between "optimism" and sample size

Model	Cox-regression model	Linear regression model	Logistic regression model
Sample size(n)	Optimism1	Optimism2	Optimism3
400	0.00805740	0.0143361	0.02325167
500	0.0206130	0.0147324	0.05234892
600	0.02095420	0.0218329	0.03351217
700	0.02514201	0.013616	0.07224059
800	0.06027282	0.051221	0.05497586
900	0.03695850	0.0187048	0.07786700
1000	0.03825738	0.0162528	0.07790918
1500	0.04084817	0.0162528	0.08108266
2000	0.05010051	0.016803	0.08188801
3000	0.0520398	0.0215589	0.08040247

This is a clear indication that there exist a positive relationship between sample size and "optimism" for prognostic models.

4.4.2. Determining the Relationship between "optimism" and Over Fitting of Regression Models Using the Number of Predictor Variables

To achieve this study obtained the 'optimism' values of model fit with a minimum of four predictors. Using three predictors as the reference measure of "optimism", the values of "optimism" for the other models were obtained are as shown in the table below; "optimism" was increasing with increase in number of predictor variables for the three models. When the predictors were four in number, the value of optimism was 0.083384358 for the cox regression model, 0.00711203 for linear model and 0.081523 for logistic model. When the predictors were increased to eight, the optimism values obtained were 0.088940566 for cox regression model, 0.03582399 for linear model and 0.0853161 for logistic model.

According to [13] over fitting results to "optimism" about a model's performance on new data. In over fitting, a model describes the random error or the noise instead of the underlying relationship. This agrees with the theory of model complexity that an attempt to over fit a prognostic model will automatically result to the model becoming optimistic. Hence it can be drawn from the result that over fitting has a direct positive relation with optimism.

The multiple line graphs below give a pictorial presentation of the relationship between 'optimism' and over fitting for the three prognostic models.

It can be seen logistic model has the highest 'optimism' values when the predictor values are increased. Linear regression models have the least tendency of being 'optimistic'.

Table 5. Relationship between "optimism" and over fitting-number of predictor variables

Model	Cox-regression model	Linear regression model	Logistic regression model
Number predictor variables	Optimism	Optimism1	Optimism2
4	0.083384358	0.00711203	0.081523
5	0.084218083	0.01359329	0.0749533
6	0.088714951	0.02405008	0.0694775
7	0.088730511	0.0332091	0.0658123
8	0.088940566	0.03582399	0.0853161

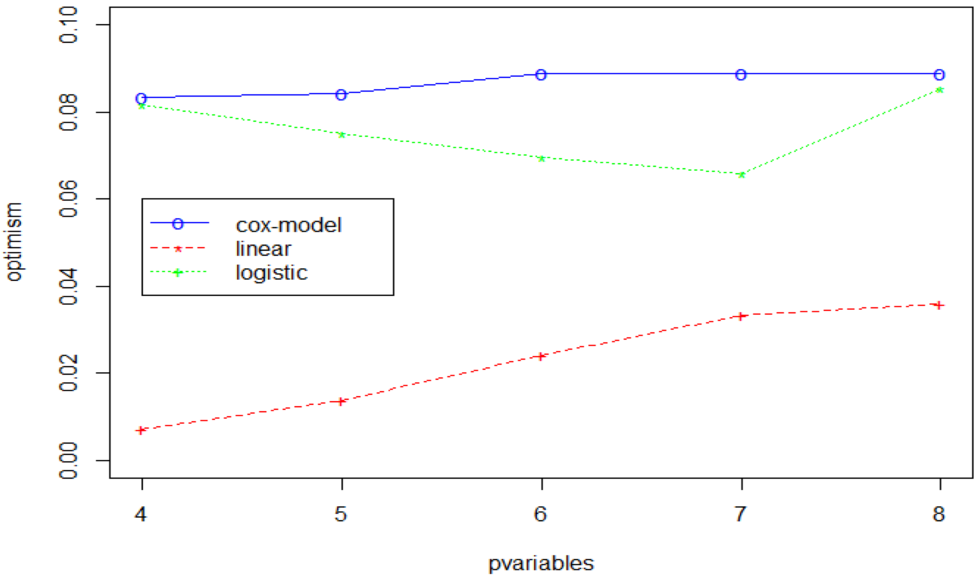

Figure 3. Multiple line graph showing the relationship between 'optimism' and over fitting

5. Conclusions and Recommendation

The main objective of the study was to evaluate methods used to assess "optimism" in regression models; the use of inferential pseudo r-square statistics through bootstrapping is indeed very informative and reliable. The use of cox&snell pseudo r-square statistic provided a platform to measure optimism in models that cannot be determined using the ordinary r-square statistic; a special case is the logistic regression model. Note that choosing cox&snell pseudo r-square as the best statistic to measure optimism does not leave out Nagelkerke statistic since they all determine model performance which is an important element to every model builder. Under large samples, they both give similar results. Levels of Optimism have a direct influence to model performance. Optimistic models will give unreliable results since they will only predict well the data that was used to develop them.

Larger samples minimize prediction errors however, when the noise variables are modeled as opposed to the underlying variables of interest then the model fails to stand the test of good fit and prediction. When the samples are large we have a wide window of modeling noise variables as opposed to smaller samples however care should be taken when deciding the sample size to avoid under fitting, where underlying model fails to capture the trend of the data at hand. The more the predictor variables the more optimistic the model becomes rendering the model less reliable in prediction. From the results of the study it would be plausible to recommend the use of pseudo r-square statistics in determining "optimism" of regression models.

Further studies need to be conducted on assessing the discrimination ability of models using the pseudo r-square statistics, Possibility of using the pseudo statistics when inferring on model fit as opposed to the ordinary r-square since they can be computed for all prognostic models.

References

[1] Curtis, K. (2012). Book Review: Spatial Regression Models Ward M.D.GleditschK.S.2008. Spatial Regression Models. Thousand Oaks, CA: Sage. ISBN 978-1-4129-5415-0. Sociological Methods & Research, 41(4), 671-674.

[2] Fahrmeir, L. (2013).Regression. Berlin; Springer.

[3] Bartlett, J. (2014). Adjusting For Optimism/Overfitting in Measures of Predictive Ability Using Bootstrapping.

[4] Kasza, J., & Wolfe, R. (2013). Interpretation of commonly used statistical regression models. Respirology, 19(1), 14-21.

[5] J. Rispoli, F., & Shah, V. (2015). Using Simulation to Test the Reliability of Regression Models. Energy and Environment Research, 5(1).

[6] Sugiyama, M. (2016). Model Selection for Maximum Likelihood Estimation. Introduction to Statistical Machine Learning, 147-156.

[7] Ziegel, E. R., & Staff, S. I. (1996). Logistic Regression Examples Using the SAS System. Technometrics, 38(1), 86.

[8] Shingleton, J. (2003). Crime Trend Prediction Using Re Gression Models For Salinas, California.

[9] Christensen, E. (1997). Prognostic models in chronic liver disease: validity, usefulness and future role.

[10] Smith, H. (2014). Regression Models, Types of. Wiley StatsRef: Statistics Reference Online.

[11] Mannan, H. R., & McNeil, J. J. (2012). Computer programs to estimate overoptimism in measures of discrimination for predicting the risk of cardiovascular diseases. Journal of Evaluation in Clinical Practice, 19(2), 358-362.

[12] Leon, L., & Cai, T. (2012). Model checking techniques for assessing functional form specifications in censored linear regression models. Statistica Sinica, 22(2).

[13] Steyerberg, E. (1999). Stepwise Selection in Small Data Sets A Simulation Study of Bias in Logistic Regression Analysis. Journal of Clinical Epidemiology, 52(10), 935-942.

[14] Kazak, A., & Kazak, R. (2003). Does cross validation provide additional information in the evaluation of regression models? Canadian Journal of Forest Research, 33(6), 976-987.