ON THE PREDICTIVE PERFORMANCE OF SOME PREDICTORS OF THE LINEAR PROBABILITY MODEL

ISSN: 2141 – 3290
www.wojast.com

MOFFAT, I. U. AND GEORGE E. U.
Department of Mathematics & Statistics,
University of Uyo, Uyo, 520231, Nigeria
moffitto2011@gmail.com and kemmyg.01@gmail.com)

ABSTRACT

In making inferences, researchers often go for a single quantity for summarizing model-fit. In regression analysis, the coefficient of determination serves this purpose when the dependent variable is normally distributed. When the dependent variable is categorical, the case is different. This paper examines three methods for assessing the goodness-of-fit of the Ordinary Least Square (OLS) model when the dependent variable is binary or categorical. The methods considered are the Percentage Correct Predictions (PCP), Percentage Reduction in Error (PRE), and Expected Percentage Correct Prediction (ePCP). After examining the three methods, the Expected Percentage Correct Prediction is seen to be the most reliable and statistically tractable measure of prediction even in the presence of discordant observations.

INTRODUCTION

Fitting models to data is a common phenomenon in mathematical and social sciences. When such models are fit, it is always important to ascertain the degree to which the model is valid, as this determines the correctness of the predicted values, using the model. A well-fit regression model results in predicted values close to the observed data values. The mean model, which uses the mean for every predicted value, generally would be used if there were no informative predictor variables. The fit of a proposed regression model should therefore be better than the fit of the mean model (Karen, 2012).

Statistically, a model fits the data well if the difference between the observed values and the model's predicted values are insignificant and unbiased. For instance, the coefficient of determination or the coefficient of multiple determinations for multiple regressions (R-squared) serves this purpose in Regression Analyses when the dependent variable is normally distributed. It is a statistical measure for determining how close the data are to the fit regression line as it describes the percentage of variation of the response variable that is explained by a linear model (Moffat, 2014).

However, assessing the goodness of fit of a regression line involves considering several things as no single characteristic of data is sufficient for a good assessment. After fitting a linear model using regression analysis, ANOVA or Design of Experiments (DOE), we need to determine how well the model fits the data.

In a binary series where the dependent variable has only two values, the data could be modeled using the Linear Probability Model (LPM), Logistic regression, Probit regression, etc. In such cases, R-squared cannot be obtained since the response variable is not normally distributed.

According to Park (2010), when a dependent variable is categorical, the Ordinary Least Squares (OLS) method can no longer produce the Best Linear Unbiased Estimator (BLUE). However, there is need for assessing model fit for such datasets. It is against this background that this

paper looks at some of the ways of achieving this result by the PCP, PRE, and the *e*PCP under the Linear Probability model.

MATERIALS AND METHOD

The Cumulative Grade-Point Averages (CGPAs) of 32, 400 level students in 2012/2013 class were considered. For the dependent binary variable, a CGPA of 2.49 and below was regarded as a failure and graded "0" while a CGPA of 2.50 and above was regarded as a pass and graded "1". The Predictor variables were real numbers in the open interval (0, 5).

Binary Response

Whenever the variable that we want to model is binary, it is natural to think in terms of probabilities. When the dependent variable y is binary, it is typically equal to one for all observations in the data for which the event of interest has happened (Success) and zero for the remaining observations (Failure). If a random sample is considered, the sample mean of this binary variable is an unbiased estimate of the unconditional probability that the event happens.

In the binary system, the expected value of *Y* is:

$$E[y] = 1 \text{ x} Pr(y = 1) + 0 \text{ x} Pr(y = 0) = Pr(y = 1) = x_i\beta \quad (1)$$

Thus the RHS of (1) must be interpreted as a *probability*, i.e. restricted to between 0 and 1. For this reason, a linear regression model with a dependent variable that is either 0 or 1 is called the Linear Probability Model (LPM). The LPM predicts the probability of an event occurring, and, like other linear models, says that the effects of *X*'s on the probabilities are linear.

According to Ash (2012), this model is associated with the following problems:

i. Unbounded predicted values: Pr (y = 1) can take on values greater than 1 and less than 0.
ii. Conditional heteroscedasticity: Actually, the classical regression model or OLS assumes that the residuals are identically distributed with mean zero and a constant variance (Homoscedasticity), i.e. Var (ε) = σ^2. But the variance of the residual is related to the value of x.
    That is, Var(y) = E[y](1 – E[y]) = $x_i\beta(1 - x_i\beta)$.
    Thus, the variance of y depends on the values of X and β and is, therefore, heteroskedastic by construction. That is why it always uses robust standard errors.
iii. Non – Normal errors: The errors can only take on two values,
    1- x_i or $-x_i\beta$. As a result, the errors can never be normally distributed, therefore causing problems for hypothesis testing.
iv. Functional Form: Given the nature of probabilities, we expect that the marginal impact of an independent variable would exhibit diminishing returns; that is, as the value of the independent variable increases, its impact on *y* should decrease. The LPM does not allow for this possibility.

However, none of these challenges actually causes a problem with the point estimates; that is, they do not cause any bias. Thus, the OLS point estimates remain unbiased estimates of the true parameter values of the slope.

According to Melton (2012), under the same assumptions as OLS, the LPM is unbiased. Although the LPM has its problems as some of its predictions are unbounded outside the range of 0 and 1 (unlike a 'true' probability model), the deficiency is accommodated in this study. Here, a predicted probability value that is negative is taken as '0' while the value that is greater than one is taken as '1'. Thus, we have established the best measure for the linear evaluation of model-fit in binary response setting.

Percentage of Correct Predictions (PCP)

The inherent issue is that the fit model has generated some predicted probabilities $(\hat{p}_i)$ and we want to go from this predicted probability to a predicted outcome - either 0 or $1_{(\hat{y}_i)}$. Once we

have the predicted outcome $(\hat{y}_i)$, we can then compare this to the actual outcome (y_i). In binary time series predictions, it is always important to determine the percentage of the observations that is correctly predicted. The most common way to do this is to say that any observation with a predicted probability greater than or equal to 0.5 should be classified as '1' and any observation with a predicted probability less than 0.5 should be classified as '0'. Thus, we have the following steps to calculate PCP:

1. Estimate $(\hat{p}_i)$ for each observation, i, using the fit model.

2. For those observations with $\hat{p}_i \geq 0.5$, we set $\hat{y} = 1$; otherwise we set $\hat{y} = 0$.

3. Call each observation C_i with $y_i = \hat{y}_i$ a correct prediction.

In reality, one of the problems associated with PCP is that of overstating its precision.

Percentage Reduction in Error (PRE)

An alternative to PCP is the Percentage Reduction in Error. This is achieved by comparing PCP with the Percentage of observations in the Model Category of the observed data (PMC).

PMC = Percentage in the Model Category.

$$PRE = \frac{PCP - PMC}{1 - PMC}$$

Expected Percentage of Correct Predictions (*ePCP*)

Herron (1999) proposed an expected Percent Correctly Predicted (ePCP). This statistics essentially provides the expected percentage of correct predictions given as:

$$ePCP = \frac{1}{n}\left[\sum_{y_i=i} \hat{p}_i + \sum_{y_i=0}\left(1 - \hat{p}_i\right)\right].$$

Assessing Model Fit

Applying these steps to the data at hand, we, first estimate the parameters of the model for the data. The probit model is then fit, and the predicted values are obtained for the calculation. From our analysis, the Linear Probability Model (LPM) for the data is given as:

The regression equation is y = - 0.498 + 0.406 x

Predictor	Coef	StDev	T	P
Constant	-0.4977	0.1221	-4.08	0.000
x	0.40553	0.04994	8.12	0.000

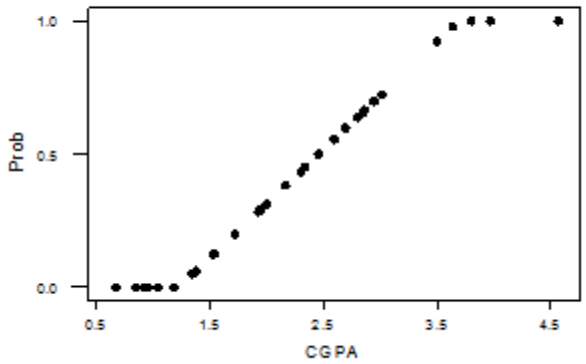
S = 0.2836 R-Sq = 68.7% R-Sq(adj) = 67.7%

ANOVA Table

Source	DF	SS	MS	F	P
Regression	1	5.3052	5.3052	65.94	0.000
Error	30	2.4136	0.0805		
Total	31	7.7188			

From the ANOVA table, it is obvious that the estimates of the parameters for the model are significant. (Fig. 1)

Fig. 1. The Graph of Predicted Probabilities against CGPA



Estimation of PCP

This model is used to obtain the probability values, hence the predicted outcomes (Table 1).

Table1: Predicted outcome.

S/N	CGPA (X_i)	RESPONSE (Y_i)	$\hat{P}_i$	$\hat{Y}_i$	C_i
1	2.84	1	0.655	1	1
2	3.97	1	1.000	1	1
3	1.35	0	0.050	1	1
4	2.46	0	0.501	1	_
5	4.57	1	1.000	1	1
6	2.84	1	0.655	1	1
7	2.59	1	0.554	1	1
8	2.70	1	0.598	1	1
9	3.64	1	0.980	1	1
10	2.34	0	0.452	0	1
11	1.72	0	0.200	1	1
12	3.80	1	1.000	1	1
13	0.96	0	0.000	0	1
14	3.50	1	0.923	1	1
15	1.18	0	0.000	0	1
16	3.01	1	0.724	1	1
17	2.00	0	0.314	0	1
18	1.93	0	0.286	0	1
19	1.38	0	0.062	0	1
20	1.53	0	0.123	0	1
21	0.85	0	0.000	0	1
22	2.94	1	0.696	1	1
23	1.05	0	0.000	0	1
24	1.95	0	0.294	0	1
25	0.96	0	0.000	0	1
26	2.17	0	0.383	0	1
27	0.92	0	0.000	0	1
28	0.68	0	0.000	0	1
29	2.30	0	0.436	0	1
30	2.86	1	0.663	1	1
31	1.54	0	0.127	0	1
32	2.80	1	0.639	1	1
		13			31

From Table 1, we have that;

$$PCP = \frac{100}{n}\sum_{i=1}^{n} C_i \qquad \Rightarrow PCP = \frac{100*31}{32} = 96.9\%$$

Estimation of the Percentage Reduction in Error (PRE)

To calculate the PRE, the percentage of the observations in the Model Category is needed. Here, PMC is the percentage of observations with $Y_i = 1$. For instance, Table1 shows that $\sum_{i=1}^{32} Y_i = 13$.

$$PMC = \frac{100}{32}\sum_{i=1}^{32} Y_i = \frac{100 * 13}{32} = 40.63\% \cdot$$

PRE is calculated as;

$$PRE = \frac{PCP - PMC}{1 - PMC} = \frac{0.969 - 0.406}{1 - 0.406} = 0.948 = 94.80\%$$

ePCP

Table 2 gives the values of Y_i and their corresponding $\hat{p}_i$. It should be noted for the last column that, $\hat{p}_i = \hat{p}_i$ for $Y_i = 1$; and $\hat{p}_i = 1 - \hat{p}_i$ for $Y_i = 0$.

Table 2

S/N	Y_i	$\hat{p}_i$	
1	1	0.655	0.655
2	1	1.000	1.000
3	0	0.050	0.950
4	0	0.501	0.499
5	1	1.000	1.000
6	1	0.655	0.655
7	1	0.554	0.554
8	1	0.598	0.598
9	1	0.980	0.980
10	0	0.452	0.548
11	0	0.200	0.800
12	1	1.000	1.000
13	0	0.000	1.000
14	1	0.923	0.923
15	0	0.000	1.000
16	1	0.724	0.724
17	0	0.314	0.686
18	0	0.286	0.714
19	0	0.062	0.938
20	0	0.123	0.877
21	0	0.000	1.000
22	1	0.696	0.696
23	0	0.000	1.000
24	0	0.294	0.706
25	0	0.000	1.000
26	0	0.383	0.617
27	0	0.000	1.000
28	0	0.000	1.000
29	0	0.436	0.564
30	1	0.663	0.663
31	0	0.127	0.873
32	1	0.639	0.639
			25.859

Thus we have from Table 2 that, $ePCP = \frac{25.859}{32} \times 100 = 80.8\%$

DISCUSSION

One problem that few have noticed is that *PCPs* incorporate a notion that is opposed to the meaning of probabilities (Train, 2007). For instance, *PCP* treats an observation with $\hat{p}_i = 0.50$ the same as an observation with $\hat{p}_i = 0.99$ despite the fact that the former value of $\hat{p}_i$ says much less than the latter. This leaves the precision of the method in doubt. Of course, the *PRE* offers a little or no remedy to the problem encountered with *PCP* since it is calculated as a function of correct and incorrect predictions (and the probabilities of correct and incorrect predictions for *ePCP*). Thus, it still has the precision problems associated with *PCP*.

The expected Percentage Correct Predictions is seen to salvage the problem of precision associated with the *PCP* as it avoids the problem of treating an observation with $\hat{p}_i = 0.50$ the same way it treats an observation with $\hat{p}_i = 0.99$. Thus, this method is comparatively the most reliable amongst the three methods considered here. However, the computational tractability of these methods under nonlinear models remains our challenge in subsequent researches.

REFERENCES

Ash, M. A.(2012). Regression with a Binary Dependent Variable. Lecture 22.

Herron, M. (1999). Postestimation Uncertainty in Limited Dependent Variable Models. *Political Analysis* 8: 83 - 98.

Karen, G. M. (2012). Assessing the Fit of Regression Models. Cornell Statistical Consulting unit, StatNews #68, 1 - 2.

Melton, J. (2012). Models for Binary Dependent Variables. Institute for Advanced Studies.

Moffat, I. U. (2014). A Probabilistic Model for Predicting Examination Performance: A Binary Time Series Regression Approach. International Journal of Sciences: Basic and Applied Research, 16 (2), 375 – 394.

Park, H. M. (2010). Regression Models for Binary Dependent Variables Using STATA, SAS, R, LIMDEP, and SPSS. University Information Technology Services, Indiana University.

Train, K. E. (2007). Discrete Choice Models with Simulation. New York: Cambridge University Press.