Using non-parametric count model for credit scoring

Sami Mestiri [1] and Abdeljelil Farhat [2]

Research Unit EAS-Mahdia
Faculty of Science Mahdia Management and Economics,
University of Monastir, Tunisia.

Summary : The purpose of this paper is to apply count data models to predict the number of times a credit applicant will not pay the amount awarded to repay the credit. Poisson models and negative binomial distribution models, taking into account the observed heterogeneity, are generally used in situations where the dependent variable is discrete. Alternatively, we propose to use non parametric model where the relationship form between conditional mean and the explanatory variables is unknown. The empirical results found suggest that the nonparametric poisson model regression has the best prediction of the number of default payment.

Keywords. : Count data ; non-parametric model ; credit scoring ; prediction.

1 Introduction

Credit scoring systems were created for the evaluation of new credit applications. This system is generally based on available statistical information, related to the behaviour of former clients with credits. Financial institutions usually apply parametric econometric models such as the logistic regression model or classifications techniques such as discriminant analysis or neural networks to create these systems.

In this work, we propose to predict future repayment behaviour using the expected number of default payment as an alternative technique. The use of this last variable suggests that appropriate models might be interesting, in which some covariant exogenous variables are included in order to specify the expected level of indebtedness. These models can be used as explanatory tools when assessing the level of risk associated with personal credit transactions.

Thus, instead of using techniques to classify individuals into groups, we suggest in our work that a judicious approach is to model the variable count "the number of default payments ", which is a way to get a model to predict the expected level of debt for new

1. Email :mestirisami2007@gmail.com
2. Email :farhat.abdeljelil@gmail.com

applicants. Poisson models and negative binomial distribution models, taking into account the observed heterogeneity, are generally used in situations where the dependent variable is discrete (Cameron and Trivedi, 1986).

Alternatively, we propose to use two non-parametric poisson models where the relationship form between conditional mean and the explanatory variables is unknown. The first model, denoted NP, estimates a totally non-parametric regression using local linear regression. A Gaussian kernel of second order is used for the explanatory variables. The second model, noted as INDEX, is a single-index model estimated using the semi-parametric least squares method of Ichimura (1993), which jointly estimates the bandwidth and coefficients using the method. non-linear least squares leaving-one.

The contribution of this paper is to develop a credit scoring system based on the nonparametric poisson model. This means that a financial institution wanted to find a method of ranking new clients requesting credit into three different classes : the good, the medium, and the bad in a more efficient way. Good customers would return the money completely, while bad customers would be the default.

2 Econometric models

2.1 The Poisson regression model

The basic model of the econometric literature for the representation and analysis of count data is the Poisson model. The endogenous variable, for example, the number of default payment noted y_i, is assumed to follow a Poisson distribution. The probability for a customer to have unpaid instalments is therefore :

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad (y_i = 0, 1, 2, ..)$$

where μ_i is the parameter of the Poisson distribution, such that : $E(y_i) = Var(y_i) = \mu_i$ This parameter is related to p exogenous variables by the log-linear form :

$$ln(\mu_i) = x_i\beta \quad \forall \quad i = 1, ..., n$$

where x_i is a vector (1, p) associated with the parameter vector $\beta_{(p,1)}$. The choice of the log-linear specification is mainly due to the need to have positive μ_i parameters. For a sample of size n, the Poisson counting model can be estimated a priori by the maximum likelihood method. The log-likelihood of this specification is :

$$lnL = \sum_{i=1}^{n}[-\mu_i + y_i x_i^{'}\beta - ln(y_i!)]$$

The likelihood equations are :

$$\frac{\partial lnL}{\partial \beta} = \sum_{i=1}^{n}(y_i - \mu_i)x_i = 0$$

The Hessian is given by :

$$\frac{\partial^2 lnL}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \mu_i x_i x_i^{'}$$

Hessian is negative definite for all x and β. Newton's method is a simple algorithm for estimation this model and will converges quickly. The estimated asymptotic variance-covariance matrix of the maximum likelihood estimator is deduced : $-[\sum_{i=1}^{n} -\hat{\mu}_i x_i x_i^{'}]^{-1}$. Given the estimation of $\hat{\beta}$, the prediction for observation i is $\hat{\mu}_i = exp(x_i^{'}\hat{\beta})$

2.2 The Negative binomial regression model

The equidispersion hypothesis in the poisson model is very restrictive. In practice due to an abundance of null values and or the presence of some extreme values, the variance is often above average. In this case, we speak of an over-dispersion of the variable Y (see Cox (1983), Hinde and Demétrio (1998)). This situation may call into question the use of this model, by an underestimation of the variances of the parameters of the model. Hence the idea of using an alternative counting model, based on the negative binomial law, which takes into account this over-dispersion by introducing an additional parameter α which makes it possible to capture the heterogeneity unobserved from the endogenous variable (which may imply unobserved over dispersion).

In a negative binomial regression model, we define the probability that Y takes the value y_i

$$P(Y_i = y_i / X_i = x_i) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i 1)\Gamma(\frac{1}{\alpha})}(\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}(\frac{\mu_i}{1+\alpha\mu_i})^{yi}$$

or α is an auxiliary parameter that measures the degree of over-dispersion. This law has a conditional mean μ_i and a conditional variance $\mu_i(1+\alpha\mu_i)$. The Negative Binomial Law tends to Poisson's Law when α goes to zero. If $\alpha > 0$, the poisson model is rejected to the negative binomial model profile.

2.3 The nonparametric poisson model

When random variables Y are univariate continuous and variable X are continuous multivariate random , the kernel estimate of the conditional mean of Y given $X = x$ is

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} y_i K(x_i - x, h)}{\sum_{i=1}^{n} K(x_i - x, h)}$$

where K(.) is a product kernel. The bandwidth h can be chosen by leave-one-out cross-validation such as generalized cross-validation and expected Kullback-Liebler cross validation (based on AIC for the nonparametric regression model).
When random variables Y are discrete we can use frequency methods, replacing the kernel weighting function $K((x_i - x), h)$ by the indicator function $1[x_i = x]$. But in practice this requires a large sample size and discrete random variables that take only a few distinct values.

Hall, Racine, and Li (2004) and Li and Racine (2007), propose use of alternative weighting functions that lead to smoother estimation, thereby reducing estimator variance at the expense of introducing some bias as in the continuous case, and that enable use of cross-validation for bandwidth selection.
For scalar probability mass function estimation with discrete random variable Y that takes c distinct values, the kernel function $K((y_i - y), h)$, for example, is replaced by the weight function

$$\begin{aligned} Kd(y_i, y, \lambda) &= 1 - \lambda \quad if \quad y_i = y \\ &= \frac{\lambda}{(1-c)} \quad if \quad y_i \neq y, \end{aligned} \tag{1}$$

where $\lambda = 0$ yields the frequency estimate and $\lambda = 1$ corresponds to a uniform weight. For nonparametric regression with discrete regressor X, one can more simply replace the kernel $K((y_i - y), h)$ with $Kd(y_i, y, \lambda) = 1$ if $x_i = x$ and $Kd(y_i, y, \lambda) = \lambda$ if $x_i = x$. When discrete data are ordered, nearby observations can be exploited in estimation, as in the continuous case. Then the kernel $K((y_i - y), h)$ is replaced with

$$Kord(y_i, y, \lambda) = \frac{c!}{j!(c-j)!}\lambda^j(1-\lambda)^{c-j} \quad if \quad |y_i - y| = j,$$

where y takes the ordered values $0, 1, ..., c-1$. If the discrete data take a large number of values, as can be the case for count data, then this will yield similar results to the continuous case and it can be simpler to use the usual kernel methods.

2.4 Poisson semiparametric models

As alternative approach, we can consider single-index poison models where the conditional mean a scalar is a function of a linear combination of the regressors, with $E[y|x] = g(x'\beta)$, where the scalar function g(.) is unspecified.

For an unknown function g(.) the single-index model β is only identified up to location and scale. To see this, note that for scalar v the function $g^*(a+bv)$ can always be expressed as $g(v)$, so the function $g^*(a + bx\beta)$ is equivalent to $g(x'\beta)$. Common normalizations are to drop the intercept and restrict. Additionally g(.) must be differentiable. In the simplest

case all regressors are continuous. If instead some regressors are discrete, then at least one regressor must be continuous ; see Ichimura (1993).

Several different estimators have been proposed that lead to a root-n consistent and asymptotically normal estimator of β and an estimator of the function g(.) that is consistent, though with a convergence rate less than root-n. These estimators include semiparametric least squares (Ichimura, 1993) and average derivative estimation (Hardle and Stoker, 1989). See, for example, Pagan and Ullah (1999) and Li and Racine (2007). These estimators ignore the intrinsic heteroskedasticity of count data, so will be inefficient.

For generalized linear models with a specified variance function, Weisberg and Welsh (1994) propose a more efficient version of Ichimura's semiparametric least squares. We suppose

$$E[y_i|x_i] = g(x_i^{'}\beta)$$
$$V[y_i|x_i] = \phi v(g(x_i^{'}\beta)),$$

where the functional form for the mean function g(.) is not specified, but that for the variance function v(.) is specified. For counts usually $v(\mu) = \phi\mu$ or $v(\mu) = \phi\mu + \alpha\mu^2$. If g() were known, then β solves

$$\sum_{i=1}^{n} \frac{(y - g(x_i\beta))}{v(g(x_i\beta))} g_(x_i\beta)x_i = 0.$$

With g(.) unknown estimation follows an alternating procedure. Given an initial estimate $\hat{\beta}$, for example from standard Poisson regression, estimate $\hat{g}(.)$ by kernel regression of y_i on $x_i\beta$ and then, given $\hat{g}(.)$ and $\hat{\beta}$, estimate the first derivative $g^{'}(.)$ by kernel methods. Then re-estimate β based on the equations with the unknown functions $g(.)$ and $g^{'}(.)$ replaced by estimates $\hat{g}(.)$ and $\hat{g}^{'}(.)$, and so on. Weisberg and Welsh (1994) show that the resulting estimator of β has the same asymptotic distribution as in the usual GLM case where g() is known, and that if a second-order kernel is used the estimate $\hat{g}(.)$ converges to g(.) at the optimal convergence rate of $n^{2/5}$

3 Prediction in credit scoring models

Usually, studies in this area take a part of the sample for estimation purposes and another part is used to check the preditive performance of the estimated models. Definition of good and bad clients was based on the number of monthly payment that were defaulted. When using poisson model, a score was associated to each individual. The score is a transformation of the probability of having been drawn from each of the two populations under study. If the estimated probability of being a good client is greater that the estimated probability of being bad, the prediction for the individual is that it belongs to the good group (and conversely, for a smaller probability). This prediction is compared

reports	Number of non-payments.
age	in years plus twelfths of a year.
income	Yearly income (in USD 10,000).
expenditure	Average monthly credit card expenditure.
owner	Factor. Does the individual own their home ?
selfemp	Factor. Is the individual self-employed ?
dependents	Number of dependent.

TABLE 1 – Description of the regressors used in the study

to the actual client behaviour. When this is done for all individuals in the sample, an estimation of classification rates is obtained.

Eventually,the performance of credit scoring models is evaluated through the percentage of correct classification for the individuals who already applied for credit, according to their subsequent behaviour. Nevertheless, the percentage of bad clients that would be classified as good by the scoring is a very important issue. It is this measure that is to be minimized since the smaller it is, the smaller the risk of granting credit to potential defaulters.

For count data models, prediction has to be performed in two steps. Firstly, the number of expected defaulted monthly instalments is found. Afterwards, the definition of predicted good or predicted bad is assigned to the individual following the same criterium that is used to define good and bad clients in the sample. At the end, predicted and real behaviour are compared to obtain estimated classification rates that may be used to evaluate the performance of this methodology to traditional approaches.

4 Estimation results

4.1 Data description

The data used in this study refer to the number of defaulted payments as those which have been analysed by Green (2003). There is a random subsample of 1002 clients for all the bank clients at a given date. They contain information about clients who had obtained loans for consumption. The interest credit is a personal loan which characterized by the fact that the amount of money granted is moderate. Usually, the loan is repaid over a short period of time and is often repaid monthly with constant payments throughout the repayment period and small in relation to individual income

The dependent variable is the number of monthly non-payments. The largest value in

Variable	Poisson	NB2
Intercept	−2.80 (0.218)	−2.844 (0.25)
income	0.128 (0.053)	0.135 (0.063)
owneryes	−0.104 (0.218)	−0.089 (0.243)
selfempyes	0.595 (0.316)	0.49 (0.39)
expenditure	0.0007 (0.0002)	0.0008 (0.0003)
α		0.572 (0.239)
log-likelihood	−306.247	−299.379
AIC	622.49	610.76
BIC	4617.494	5404.758

TABLE 2 – Poisson and NB2 fitted models

the sample is 4. The number of zero counts is 813. The proportion of clients with zero non-payments is 81.32 %. A description of the variables used in this paper can be found in Table 1.

4.2 Model comparison

For estimation purposes, some individuals were eliminated from the original sample. Individuals with repayment lasting more than four months at sample collection were excluded from the estimation process on the grounds that there was not enough information about their repayment behaviour and that posterior classification could be misleading.

Table 2 shows estimates and standard errors for two parametric models. The negative binomial model (NB2) fits the distribution of the data much better than the Poison, with log-likelihood increasing from -306.247 to -299.379. It is interesting to see that parameter estimates are the same, except for variable selfempyes, but note how estimation of a Poison model leads to distorted standard errors due to the fact that heterogeneity is not taken into account.

We also performed nonparametric kernel estimation of the conditional density of dependant variable **reports** given exploratory variables **income, owneryes, selfempyes** , and **expenditure**. Then $\hat{f}(y|x)$ is obtained as the ratio of a four-dimensional kernel density estimate to a four dimensional kernel density estimate, where **reports** is treated as ordered discrete data with the weighting function (), **income** and **expenditure** are treated as continuous with a second-order Gaussian kernel of fixed bandwidth, and **owneryes, selfempyes** , is an unordered binary discrete variable with the weighting function

	Prévues by NB2					Predicted by semi .param				
Actual	0	1	2	3	4	0	1	2	3	4
0	154	20	1	0	1	62	62	37	8	7
1	16	2	0	0	0	4	9	4	1	0
2	6	0	0	0	0	2	1	2	1	0
3	1	0	0	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0

TABLE 3 – Reports : Actual versus prediction

Predicted	Mean	S.D	Min	Max	$Cor[y, \hat{y}]^2$
Reports (y)	0.164	0.48	0	3	.
Poisson	0.134	0.42	0	4	0.1
NB2	0.139	0.44	0	4	- 0.04
NP	1.129	1.01	0	5	0.16
INDEX	1.075	1.03	0	4	-0.1

TABLE 4 – Reports :Summary of various fitted means

(). The bandwidth is chosen using expected Kullback-Liebler cross-validation.

We additionally estimate semiparametric models of the conditional mean of **reports** given the three regressors **income, owneryes,selfempyes** , and **expenditure**. The fourth model, denoted INDEX, is a single-index model estimated using the semiparametric least squares method of Ichimura (1993) that jointly estimates the bandwidth and coefficients using leave-one-out nonlinear least squares.

4.3 Prediction

Table 3 is a classification table that compares the actual count y_i to the predicted count $\hat{y}_i$, where $\hat{y}_i = k$ if the conditional density estimate $f(y|x_i)$ is maximized when $y = k$. The semiparametric estimates predict zeros well and underpredict intermediate and larger counts. By contrast the NB2 model, does similarly well in predicting zeros and ones, but underpredicts intermediate and larger counts much more. For example, for the 18 observations with **reports** equal 1 the NB2 model predicts that only 2 count , whereas the semiparametric model predicts that 9 counts. For the 7 observations with **reports** excess of 2 the NB2 model predicts that 0 count , whereas The semiparametric predicts two counts in excess of 2 .

Table 4 presents descriptive statistics for the predicted values the number of default payments of studies models. The NB2 model does particularly poorly, with the lowest squared correlation (of -0.004) between the actual and fitted values. Fitting the entire

	Reports	yhat.poiss	yhat.nb	yhat.npreg	yhat.npindex
Reports	1.00	0.06	-0.06	0.11	0.01
yhat.poiss	0.06	1.00	0.42	0.04	-0.10
yhat.nb	0.06	0.42	1.00	-0.04	-0.09
yhat.npreg	0.11	0.04	-0.04	1.00	-0.10
yhat.npindex	0.01	-0.10	-0.09	-0.10	1.00

TABLE 5 – Reports : Correlations of various fitted means

distribution using an NB2 model in this data example leads to poorer fit of the mean, as is also evident from the average fitted mean of 0.139 being substantially higher than the sample mean of 0.164. The empirical results found suggest that the best fit models are semi-parametric single-index regression. The nonparametric model leads to fitted values that are fairly similar to those for the index models. The nonparametric model is preferred because $Cor[y, \hat{y}]^2$ is 0.16 compared to -0.1 for the single-index model.

Table 5 presents correlations for the fitted values. The nonparametric poisson model fitted values are highly correlated with the actual number of default paiement, suggesting that the nonparametric poisson model may be a good model for these data.

FIGURE 1 – Fitted values from four models plotted against actual value.

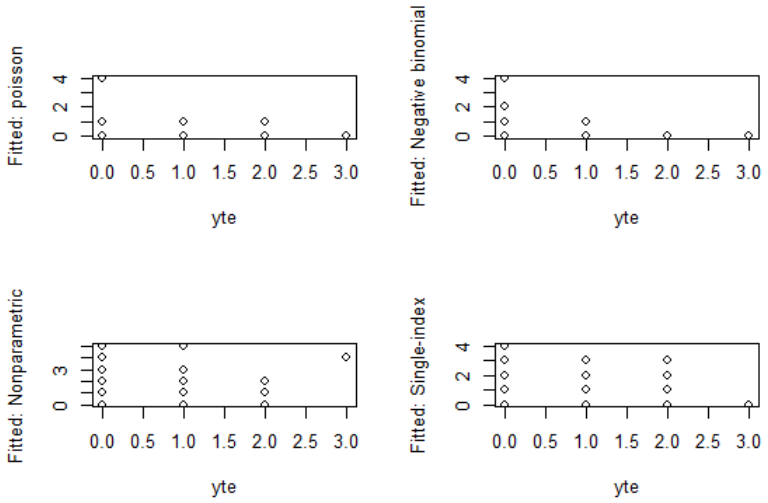

Figure 1 presents plots of the fitted values from all these models except for the Poisson against the actual number of default paiement. These plots also suggest that the best fitting model is the nonparametric poisson model.

5 Conclusion

In this paper, we used nonparametric poison models to account for both heterogeneity and zero inflation present in a data set for credit-scoring purposes aiming at analysing the credit-scoring behaviour for individual loans and identifying the number of classes of clients without making assumptions about the parametric form of the heterogeneity term. We fitted four competing models in order to capture the present heterogeneity and to better describe the data.

The main contribution of the paper is that we tried to model with more sophisticated models the number of defaulted payments, allowing for a different kind of credit scoring rather than the traditional good versus bad categorization. Our results verify in a statistically concrete basis what is well-known in credit-scoring literature, namely that the two-class categorization is not sufficient and that the population consists of more groups. Classification problems in the context of credit granting decisions may use count data models due to the characteristics of the dependent variable. In fact, the number of defaulted payments is the variable used to define whether a client is good (repaying) or bad (defaulter). Adequate use of count data models with non parametric form is useful to find which are the most influential variables in the studied process. It has to be noted that estimation required assymptotic approximations for standard errors.

Further research is needed to desetangle some obscure points such as model selection or missspecification in non parametric poisson models. In this situation although prepayment has not been considered, one should see the way to include duration of repayment at sample collection and its influence in final estimation and classification results.

References

[1] Dionne, G., M. Artis and M. Guillen (1996). Count data models for a credit scoring system. *J. Empirical Finance.*

[2] Cox, D.R. (1983), "Some Remarks on Overdispersion," *Biometrika*, 70, 269–274.

[3] Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

[4] Cameron, A.C., Trivedi, P.K., (1986). Regression based tests for overdispersion in the Poisson model, *Journal of Econometrics*, 46, 347-364

[5] Cameron, A.C., Trivedi, P.K., (1997). Regression Analysis of Count Data, *Econometric Society Monograph n 30, Cambridge University Press.*

[6] Hall, B.H., Z. Griliches, and J.A. Hausman (1986), "Patents and R and D : Is There a Lag ?," *International Economic Review*, 27, 265–283.

[7] Hall, P., J. Racine, Q. Li (2004), "Cross-Validation and the Estimation of Conditional Density Functions," *Journal of the American Statistical Association*, 99, 1015–1026.

[8] Greene, W.H., (1997). FIML estimation of sample selection models for count data,

Working paper EC-97-02, Stern School ef Business, New York University.
[9] Greene,W.H. (2011), Econometric Analysis, *edition 7, Upper Saddle River, NJ, Prentice Hall.*
[10] Mullahy, J., (1986). Specification and testing of some modified count data models, *Journal of Econometrics*, 33, 341-365
[11] Ichimura, H. (1993), "Semiparametric Least Squares (SLS) andWeighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120.
[12] Ichimura, H., and P. Todd (2007), "Implementing Nonparametric and Semiparametric Estimators," *Handbook of Econometrics volume* 6, Part B, 5364–5468. Amsterdam, North-Holland.
[13] Li, Q., and J. Racine (2007), Nonparametric Econometrics, *Princeton, NJ, Princeton University Press.*
[14] Li, Q., and J. Racine (2008), "Nonparametric Estimation of Conditional CDF and Quantile
Functions With Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics*, 26, 423–434.
[15] Pagan, A.R., and A. Ullah (1999), Nonparametric Econometrics, *Cambridge, UK, Cambridge University Press.*
[16] Weisberg, S., and A.H. Welsh (1994), "Adapting for the Missing Link," *Annals of Statistics*, 22, 1674–1700.

6 Appendix

Estimation of Poisson model

```
formula.model<- reports ~ age + income + expenditure
ccpois <- glm(formula.model, data = CreditCard, family = poisson)
summary(ccpois)
logLik(ccpois)
pre<-round(predict(ccpois,datastq),3)
lambda <-round(exp(pre),3)
yhat.poiss <- rpois(nrow(datastq),lambda )
table(yhat.poiss)
predict1 <- cbind(yte,yhat.poiss)
table(yte,yhat.poiss)
```

Estimation of Negatif binomial model

```
library(MASS)
model.nb <- glm.nb(formula.model, data = CreditCard)
summary(model.nb)
pre1<-round(predict(model.nb,datastq),3)
lambda1 <-round(exp(pre1),3)
yhat.nb = rpois(nrow(datastq),lambda1 )
table(yhat.nb)
predict2 <- cbind(yte,yhat.nb)
table(yte,yhat.nb)
```

Nonparametric conditional mean estimation (local linear kernel)

```
library(np)
bw.npreg <- npregbw(formula.model,,regtype="ll",bwmethod="cv.aic",data =
CreditCard)
summary(bw.npreg)
model.npreg <- npreg(bws=bw.npreg, gradients=TRUE)
summary(model.npreg)
pre3<-round(fitted(model.npreg,datastq),3)
lambda2 <-round(exp(pre3),3)
yhat.npreg = rpois(nrow(datastq),lambda2)
table(yhat.npreg)
predict3 <- cbind(yte,yhat.npreg)
table(yte,yhat.npreg)
```

Semiparametric single index conditional mean estimation

```
bw.npindex <- npindexbw(formula.model,data = CreditCard)
summary(bw.npindex)
model.npindex <- npindex(bws=bw.npindex, gradients=TRUE)
summary(model.npindex)
pre4<-round(fitted(model.npindex ,datastq),3)
lambda3 <-round(exp(pre4),3)
yhat.npindex = rpois(nrow(datastq),lambda3)
table(yhat.npindex)
predict4 <- cbind(yte,yhat.npindex)
table(yte,yhat.npindex)
```

Compare the various predicted conditional means

```
predictedmeans <- cbind(reports,yhat.poiss,yhat.nb,yhat.npreg,yhat.npindex)
apply(predictedmeans,2,mean)
apply(predictedmeans,2,sd)
summary(predictedmeans)
round(cor(predictedmeans),2)
```

Plots of the fitted values

```
par(mar=c(2, 2))
plot(yhat.poiss~ reports ,ylab="Fitted : poisson")
plot(yhat.nb~reports ,ylab="Fitted : Negative binomial")
plot(yhat.npreg~reports ,ylab="Fitted : Nonparametric")
plot(yhat.npindex~reports,ylab="Fitted : Single-index")
```