Journal of Statistical Computation and Simulation
Vol. 00, No. 0, Month 2004, pp. 1–15

Taylor & Francis
Taylor & Francis Group

CATEGORICAL DATA REGRESSION DIAGNOSTICS FOR REMOTE ACCESS SERVERS

JEROME P. REITER* and CHRISTINE N. KOHNEN

Practice of Statistics and Decision Sciences, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708, USA

Q1

(Revised 24 November 2003; In final form 5 August 2004)

Owing to the growing concerns over data confidentiality, many national statistical agencies are considering remote access servers to disseminate data to the public. With remote servers, users submit requests for output from statistical models fit using the collected data, but they are not allowed access to the data. Remote servers also should enable users to check the fit of their models; however, standard diagnostics like residuals or influence statistics can disclose individual data values. In this article, we present diagnostics for categorical data regressions that can be safely and usefully employed in remote servers. We illustrate the diagnostics with simulation studies.

Keywords: Categorical data; Confidentiality; Diagnostic; Disclosure; Regression; Remote access

1 INTRODUCTION

As concerns over data confidentiality grow, in the near future national statistical agencies may not be willing or legally allowed to release genuine data on individual units. In such a world, one data dissemination strategy is remote access servers, to which users submit requests for output from statistical models fit using the collected data, but they are not allowed access to the data (Keller-McNulty and Unger, 1998; Duncan and Mukherjee, 2000; Schouten and Cigrang, 2003). In fact, several statistical agencies are developing or already use remote servers as part of their data dissemination strategies, including the Australian Bureau of Statistics, Statistics Canada, Statistics Denmark, Statistics Netherlands, Statistics Sweden, the US Bureau of the Census, the US National Agricultural Statistics Service, the US National Center for Education Statistics, and the US National Center for Health Statistics (Rowland, 2003).

Remote access servers have advantages over disclosure limitation strategies that ultimately release altered versions of the original data, such as collapsing categories or receding continuous variables into intervals (Willenborg and de Waal, 2001), swapping data values of randomly selected units (Dalenius and Reiss, 1982), and adding noise to the original data (Fuller, 1993). First, remote server analyses are based on the original data and so are free from biases injected by data perturbation methods. Second, users of remote servers can fit standard statistical models; there is no need to make corrections for measurement errors, as is needed when analyzing perturbed data. Third, remote servers can protect confidentiality more

* Corresponding author. E-mail: jerry@stat.duke.edu

ISSN 0094-9655 print; ISSN 1563-5163 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/00949650412331299184

effectively than releasing altered data, since no actual or close-to-actual values for individual units are purposefully released.

In addition to providing basic output such as estimated model parameters and their standard errors, the remote server output also should include some way for users to check the fit of their models. Unfortunately, releasing the usual diagnostic statistics can disclose values. For example, when actual residuals and fitted values are released for a submitted regression model, the user can obtain the values of the dependent variable by simply adding the residual to the fitted value. A way around this problem for linear regression models was proposed by Reiter (2003): remote servers can provide synthetic, *i.e.* simulated, diagnostics that mimic real-data diagnostics. For example, agencies can release synthetic values of dependent and independent variables, residuals, and fitted values. Users then can treat these synthetic values like ordinary diagnostic quantities, such as examining scatter plots of the synthetic residuals versus the synthetic independent variables or versus the synthetic fitted values.

In this article, we propose remote server diagnostics for regressions involving categorical dependent variables, in particular logistic and multinomial regressions. Unlike the diagnostics for linear regressions in Reiter (2003), which are based on unit-specific residuals, the categorical data regression diagnostics are based on grouped residuals (Landwehr *et al.*, 1984; Gelman *et al.*, 2000). The article is organized as follows. Sections 2 and 3 present remote server diagnostics for logistic and multinomial regressions. Section 4 illustrates the diagnostics using genuine data. Section 5 concludes with a discussion of the diagnostic measures.

We focus solely on diagnostics, assuming that the server has provided estimated coefficients and standard errors for the submitted regression. It is possible that the estimated coefficients and standard errors themselves are disclosure risks, regardless of the availability of diagnostic measures. For example, when all units with the same covariate pattern have the same outcome values, the fitted regression that includes an indicator variable for that pattern predicts outcomes exactly for those units. We presume the server has been programmed not to provide output for such unsafe regressions, so that any released estimated coefficients and standard errors carry acceptable risks. We do not consider here strategies for assessing which regressions have releasable output. This is an area of active research.

2 DIAGNOSTICS FOR LOGISTIC REGRESSIONS

Releasing unit-specific, logistic regression diagnostic measures, such as residuals or case influence statistics, can result in disclosures. All units with outcome equal to one must have negative residuals, and all units with outcome equal to zero must have positive residuals. Hence, when the user is able to determine the sign of a unit's residual, he or she knows the value of that unit's outcome. Similar problems exist for popular case influence statistics, such as the changes in the total deviance or change in the Chi-squared test statistic after deleting each observation from the data set (Pregibon, 1981; Hosmer and Lemeshow, 2000). Any unit with a large case influence statistic does not fit the pattern of the model, so that its value of the outcome must be as far as possible from its predicted probability. In other words, units with large case influence statistics and small predicted probabilities have outcomes equal to one, and units with large case influence statistics and large predicted probabilities have outcomes equal to zero.

Adding noise to the residuals or to the case influence statistics generally will not transform them into simultaneously safe and useful diagnostics. For residuals, the noise must be large enough to cause users to be unsure of the signs of the actual residuals; this practically eliminates the diagnostic utility of the released residuals.

For case influence statistics, adding random noise leaves similar problems: when the noise is small, the user still can associate large values with the units' outcomes, and when the noise is large the diagnostics have limited diagnostic utility. Hence, we do not believe it is feasible in general to release case-specific, remote server diagnostic measures for logistic regressions.

We suggest instead to release grouped diagnostics, related to those proposed by Landwehr *et al.* (1984) and Gelman *et al.* (2000). Let $\mathbf{x}_p$, for $p = 1, \ldots, d$, denote the independent variables in the user's submitted logistic regression. The variables in $\mathbf{x}_p$ may be continuous or discrete. The server generates grouped diagnostics separately for each $\mathbf{x}_p$ in four steps: (i) partition $\mathbf{x}_p$ into m_p disjoint categories; (ii) calculate the percentages of ones in the dependent variable for each of the m_p categories, and add a small amount of noise to these percentages; (iii) calculate the averages of the predicted probabilities from the fitted model for each of the m_p categories; and (iv) in each of the m_p categories, release the medians of the values of the $\mathbf{x}_p$, the perturbed percentages of ones, and the averaged predicted probabilities. When the predicted probabilities are substantially different from the perturbed percentages in some regions of $\mathbf{x}_p$, the model needs to be adjusted to fit better in those problematic regions. In step (ii), server administrators can skip adding random noise to the observed percentages when releasing exact percentages in the categories is not considered a disclosure risk.

The partitioned versions of $\mathbf{x}_p$ are used only for diagnostic purposes. Coefficients and standard errors of the logistic regression are estimated using $\mathbf{x}_p$ as submitted by the user, without any partitioning. Partitioning the $\mathbf{x}_p$ for diagnostics ensures that the server does not release individual units' predictor values, which may be sensitive. The larger the partition sizes, the greater the protection of individuals' values of $\mathbf{x}_p$ but the weaker the potential for detecting model inadequacies. The smaller the partition sizes, the weaker the protection but the greater the potential for detecting model inadequacies. The categories of any $\mathbf{x}_p$ should be the same for all submitted models, so that calls to different regressions do not inadvertently reveal additional information about individuals' predictor values.

The $\mathbf{x}_p$ are partitioned marginally, *i.e.* without concern for the joint distributions among the predictors. Partitioning marginally helps protect confidentiality by avoiding the creation of cross-classified categories with small numbers of units. It also simplifies the construction of the server: it need only generate diagnostics for the submitted regression. A downside to drawing marginally is that the utility of the diagnostics is reduced. For example, suppose the submitted logistic regression includes predictors for sex and for a dichotomous version of race (minority and non-minority). Marginal diagnostics provide information on the model fit only for the main effects of sex and race, which may not reveal race–sex interactions. The user can assess the importance of interactions directly by submitting regressions that include interaction terms, provided of course that the server is willing to release the output from the fitted regression. Alternatively, the server could be programmed to provide diagnostics on the basis of categorizations of joint distributions when they are safe to release. These judgments are made before the server goes on-line.

Adding random noise to the observed percentages is advisable when the outcomes are almost all zeros or ones for some categories of $\mathbf{x}_p$. For example, suppose only one unit in some category of $\mathbf{x}_p$ has a zero outcome. If the observed percentage is not perturbed, the unit with that zero knows that all other units in its category have ones, which is a disclosure. We perturb the number of ones in each category of $\mathbf{x}_p$ by randomly adding a uniform draw from $(-2, -1, 1, 2)$. When the resulting, perturbed percentage of ones in any category is less than zero or greater than one, we redraw until the perturbed percentage is in fact between 0% and 100%. Adding or subtracting a maximum of two units ensures that a released percentage could have been generated from at least two values of the real-data percentages, which helps protect confidentiality. Excluding zero from the set of possible random draws ensures that no exact percentages are released. The same random seed should be used to generate the random noise

for all logistic regressions involving the same outcome variable, so that users cannot refine guesses of the observed percentages with repeated calls for the output from the same logistic regression.

For categorical $\mathbf{x}_p$, the m_p categories can be the levels of $\mathbf{x}_p$, provided there are sufficient number of units in each level so as to prevent disclosures. When specific levels of $\mathbf{x}_p$ do not contain a sufficient number of units, they can be merged with others until an agency-specified minimum size is reached. These determinations are made before the server is made available to the public. For continuous $\mathbf{x}_p$, we recommend forming categories of 100 units with similar values of $\mathbf{x}_p$, because adding or subtracting noise of at most 2% should not greatly distort relationships between the real-data percentages and predicted probabilities for most values of these quantities. The released diagnostics should include the number of units in each category of x_p, so that the user can determine how much error is added by any perturbations.

To illustrate the utility of these diagnostics, we simulate two scenarios using different specifications for the independent variables in the logistic regressions. For each simulation, we generate $n = 10{,}000$ observations. The independent variables, x_1 and x_2, are randomly drawn from $N(0, 2^2)$. To generate the dichotomous dependent variable, we draw from Bernoulli distributions with probabilities, $\exp(g(x_1, x_2))/(1 + \exp(g(x_1, x_2)))$, where the $g(x_1, x_2)$ are as shown in Table I. We generate grouped diagnostics using the methods outlined previously. Each independent variable is split into $m = 100$ categories, each comprising $k = 100$ units. Random noise is added to the real-data percentages, drawn uniformly from $(-0.02, -0.01, 0.01, 0.02)$.

For the non-linearity simulation, we fit two models to the data. The first is a logistic regression that uses a linear function of x_1, an incorrect model for the data. The second is the correct logistic regression that fits a quadratic function of x_1. Figure 1 displays the perturbed percentages (labeled with $\circ$), and the predicted probabilities (labeled with $*$) from the incorrect model, plotted against the values of the midpoints of the categories of x_1 The predicted probabilities

TABLE I Scenarios for logistic regression simulations.

Case	*Logistic link*
Non-linearity	$g(x_1, x_2) = -3 + 4.3x_1 + 1.5x_1^2$
Interactions	$g(x_1, x_2) = 1 + 2.2x_1 - 4.6x_2 + x_1x_2$

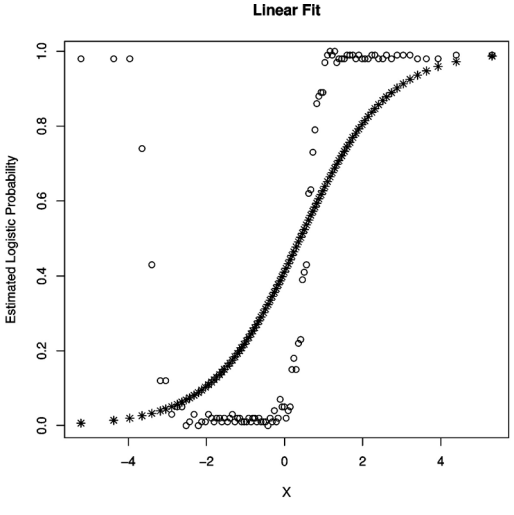

FIGURE 1 Diagnostics for non-linearity simulation when using the incorrect logistic regression model.

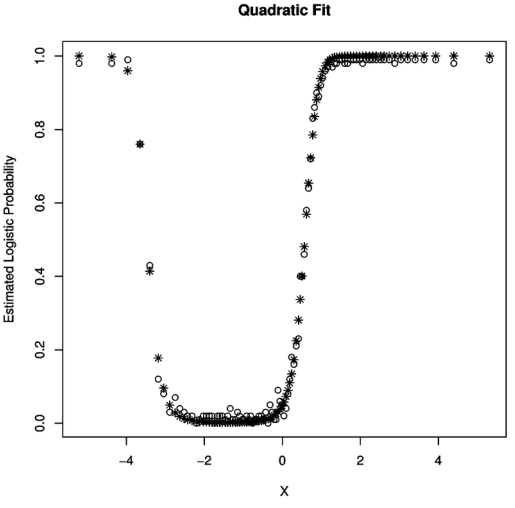

FIGURE 2 Diagnostics for non-linearity simulation when using the correct logistic regression model.

differ greatly from the perturbed percentages, an indication that the user should modify the model. As shown in Figure 2, adding the quadratic term to the model greatly improves the fit, and the predicted probabilities closely resemble the perturbed percentages. This clearly illustrates the utility of these diagnostics in model evaluation over including only the coefficients and standard errors in the logistic regression output.

For the interactions scenario, we also fit an incorrect and the correct model. The incorrect model uses just the main effects of x_1 and x_2, whereas the correct model includes the main effects and the interaction. As shown in Figure 3, the main effects model leads to poor correspondence between the perturbed percentages and predicted probabilities; the diagnostics reveal that this model does not fit the data. Figure 4, on the other hand, shows coherence between the predicted probabilities and the perturbed percentages, correctly suggesting that this model is a reasonable fit to the data.

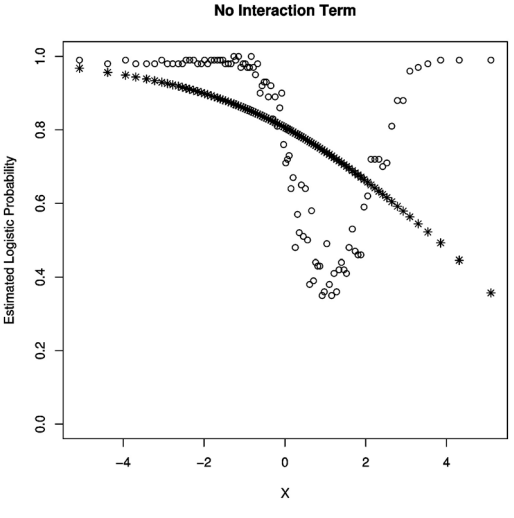

FIGURE 3 Diagnostics for interactions simulation when using the incorrect logistic regression model.

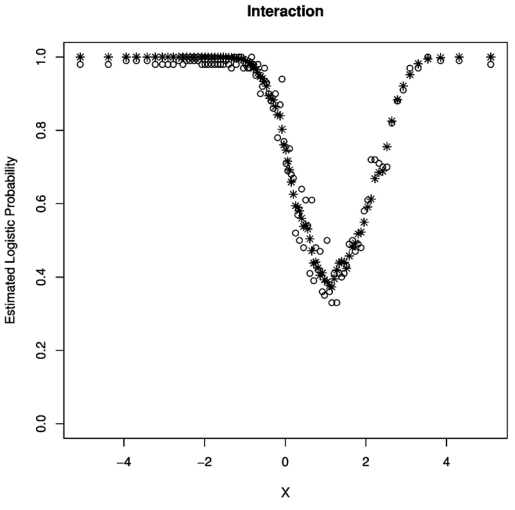

FIGURE 4 Diagnostics for interactions simulation when using the correct logistic regression model.

3 DIAGNOSTICS FOR MULTINOMIAL REGRESSIONS

In multinomial regressions, the outcome variable has $h > 2$ categories. Each unit's residuals are an $h \times 1$ vector, $(y_1 - \hat{\pi}_1, y_2 - \hat{\pi}_2, \ldots, y_h - \hat{\pi}_h)$, where $y_j = 1$ when the unit's outcome is in category j and $y_j = 0$ otherwise. The only positive value in the residual vector is for the category matching the unit's outcome. Hence, releasing even the signs of the unit's residuals from a particular multinomial regression discloses the unit's outcome. Introducing doubt about the direction of the signs would require so much noise so as to render the released residuals useless for diagnostics.

As for logistic regression, we propose to release grouped diagnostics rather than case-specific diagnostics. The $\mathbf{x}_p$ are partitioned into m_p categories. Once again, the partitioning is used only for diagnostics; the multinomial regression coefficients are estimated using $\mathbf{x}_p$ as submitted by the user. For each of the m_p categories, the server computes the number of units in each of the h outcome categories, and, if desired by the server administrator, adds random noise to these counts drawn from $(-2, -1, 1, 2)$. For each level of the categorized $\mathbf{x}_p$, the server releases the medians of the values of $\mathbf{x}_p$, the perturbed percentages of units in each of the outcome categories, and the average predicted probabilities. Users can examine these quantities to determine whether or not the predicted probabilities closely resemble the perturbed percentages. Similar diagnostics can be used on proportional Odds models for ordered categorical data, on the basis of predicted and perturbed cumulative probabilities.

We illustrate these diagnostics using simulations like those in Section 2. We generate $n = 10{,}000$ observations with independent variables, x_1 and x_2, drawn randomly from $N(0, 2^2)$. To generate the dependent variables, we draw from multinomial distributions with $h = 3$ categories and the following probabilities for the three categories:

$$\pi_0 = \frac{1}{1 + \exp(g_1(x_1, x_2)) + \exp(g_2(x_1, x_2))}$$

$$\pi_1 = \frac{\exp(g_1(x_1, x_2))}{1 + \exp(g_1(x_1, x_2)) + \exp(g_2(x_1, x_2))}$$

$$\pi_2 = \frac{\exp(g_2(x_1, x_2))}{1 + \exp(g_1(x_1, x_2)) + \exp(g_2(x_1, x_2))}$$

with the values of $\exp(g_1(x_1, x_2))$ and $\exp(g_2(x_1, x_2))$ shown in Table II.

Using these data sets, we generate grouped diagnostics by splitting each independent variable into $m = 100$ categories, each consisting of $k = 100$ units. The real-data percentages for each category are perturbed by adding random draws from $(-2, -1, 1, 2)$ to the numbers of ones. All multinomial regressions are fit using the 'multinom' function in the software package R.

For the non-linearity simulation, we fit two functions of x_1 in the multinomial regression models: an incorrect linear function and the correct quadratic function. Figure 5 displays the perturbed percentages (labeled with $\circ$) and the predicted probabilities (labeled with $*$) from the incorrect, linear function for the three values of dependent variables ($y = 1, 2$, and 3). The predicted probabilities differ widely from the perturbed percentages, reflecting the lack of fit. As displayed in Figure 6, adding the quadratic term to the model greatly improves the fit.

In the interaction simulation, we once again fit both an incorrect and correct model. The incorrect model only uses the main effects of x_1 and x_2, whereas the correct model includes these main effects along with the interaction. The synthetic diagnostics in Figure 7 properly show the lack of fit for the main effects model, particularly for larger values of x_1. Figure 8 shows that once the interaction is included in the model, there is better correspondence between the predicted probabilities and perturbed percentages, which suggests that the model is a better fit.

TABLE II Scenarios for multinomial regression simulations.

Case	*Logistic link*
Non-linearity	$g_1(x_1, x_2) = 3.1 - 2.5x_1 + 1.8x_1^2$ $g_2(x_1, x_2) = 2 - 1.6x_1 + 1.4x_1^2$
Interactions	$g_1(x_1, x_2) = 2 + 1.5x_1 - 0.8x_2 + x_1x_2$ $g_2(x_1, x_2) = 1.5 + 2.2x_1 - 0.4x_2 + x_1x_2$

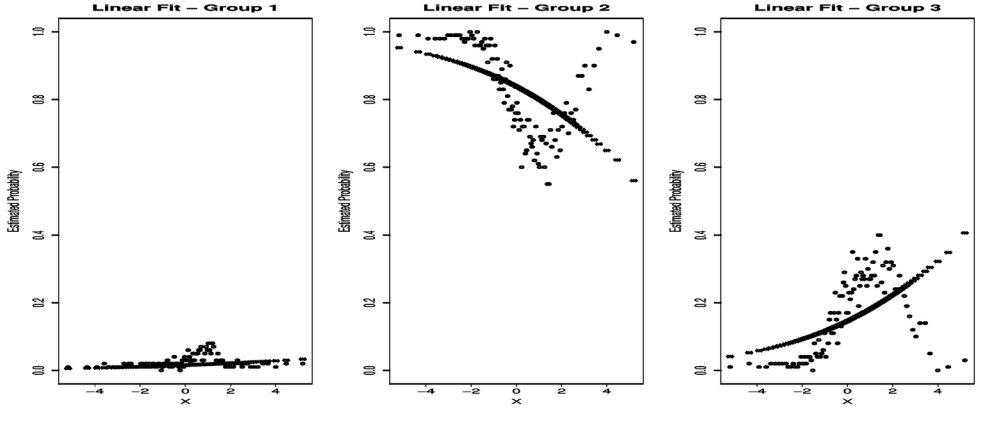

FIGURE 5 Diagnostics for non-linearity simulation when using the incorrect multinomial regression model.

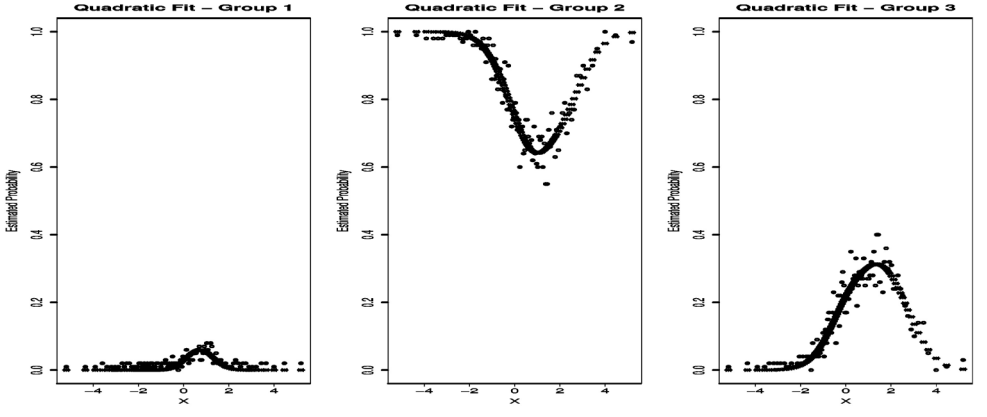

FIGURE 6 Diagnostics for non-linearity simulation when using the correct multinomial regression model.

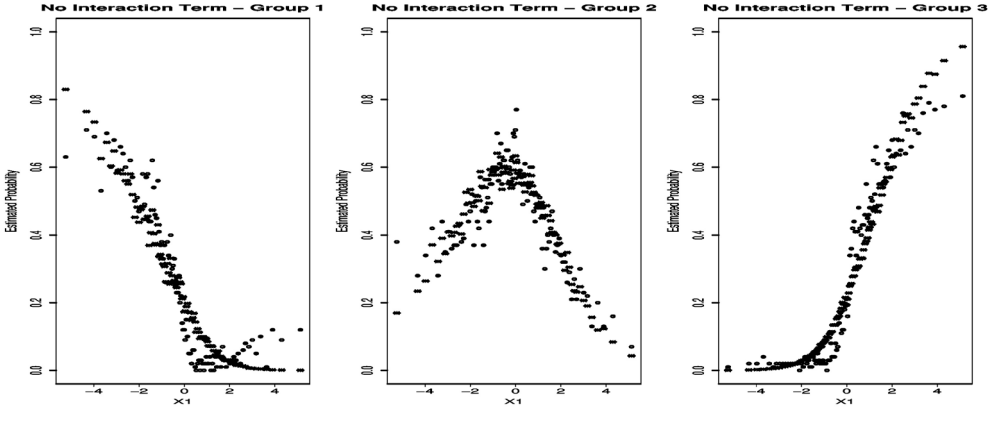

FIGURE 7 Diagnostics for interactions simulation when using the incorrect multinomial regression model.

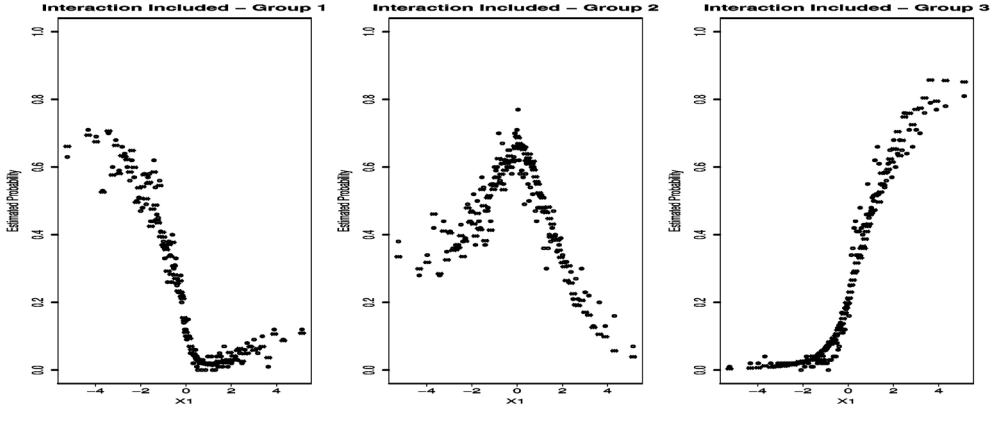

FIGURE 8 Diagnostics for interactions simulation when using the correct multinomial regression model.

4 SIMULATIONS WITH GENUINE DATA

We next investigate the utility of these remote server diagnostics for logistic and multinomial regressions on genuine data. The data comprise 10,000 randomly sampled heads of households from the public release files of the March 2000 US Current Population Survey. Table III describes the variables used in the models.

Marital status, M, has seven types: $M = 1$ for married civilians with both spouses present at the home; $M = 2$ for married people in the armed forces with both spouses present at the home; $M = 3$ for married people with one spouse not present at the home; $M = 4$ for widowers; $M = 5$ for divorced people; $M = 6$ for separated people; and, $M = 7$ for people who never have been married. Highest attained education level, E, increases from 31 to 46 in correspondence with years of schooling. As examples, $E = 31$ represents highest educational attainments of less than first grade; $E = 39$ represents a high school degree; $E = 43$ represents a bachelor's degree; and, $E = 46$ represents a doctoral degree. Out of the 10,000 households, 6612 have positive property taxes, P, and the remainder have zero property tax. All 10,000 households have positive income, I. Both monetary variables have long right tails.

From these data, we seek to fit a logistic regression model to predict whether or not households have positive property taxes, and to fit a multinomial regression model to predict marital status for the household head. For simplicity, complications due to the complex sampling design are ignored, and the 10,000 households are treated as a simple random sample. The models we consider are not the best models for predicting property taxes, but they are useful for illustrating the remote server diagnostics.

4.1 Property Tax Logistic Regression

In the model descriptions to follow, a bold-face letter corresponds to an independent variable fit, as a continuous variable and a plain letter represents variables fit as a series of indicator variables. Using the GLIM notation of McCullagh and Nelder (1989), the initial function of predictors in the fitted logistic regression model is

$$1 + \mathbf{I} + \mathbf{G} + \mathbf{E} + \mathbf{N} + \mathbf{Y} + M + X + R \tag{1}$$

When constructing the remote server diagnostics, we categorize the continuous variables in groups of 100 and add random noise from $(-2, -1, 1, 2)$ to the observed numbers of people with positive property taxes. In the plots involving $\mathbf{N}$, all levels of $\mathbf{N} \geq 7$ are collapsed into one level of 'at least 7' because there are few households of those sizes. This helps protect the confidentiality of these large households.

Figure 9 displays the remote server diagnostics for each of the eight independent variables. For most variables, the predicted probabilities and the perturbed percentages are similar.

TABLE III Variables from CPS data used in simulations.

Variable	*Label*	*Range*
Sex	X	Male, female
Race	R	White, black, Asian, Amer. Indian
Marital status	M	Seven categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0–90
Number of people in house	N	1–16
Number of youths (under 18 years) in house	Y	0–10
Household property taxes ($)	P	0–99,997
Household income ($)	I	1–768,742

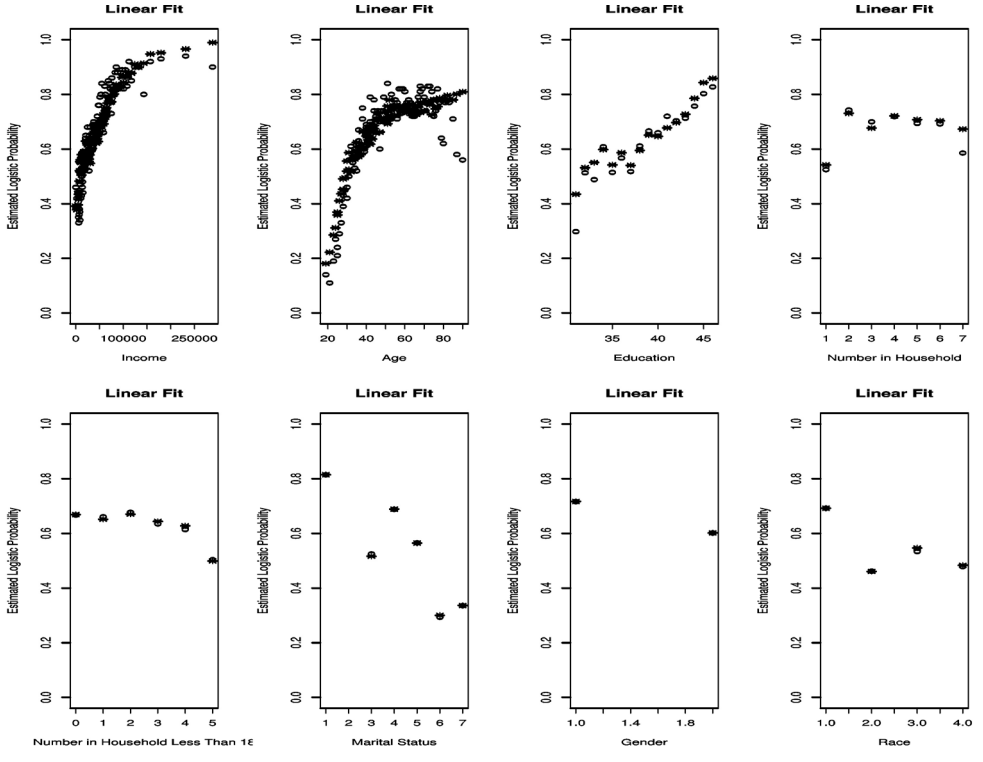

FIGURE 9 Diagnostics for initial logistic regression model for property taxes. For number in household, '7' includes households with at least seven people; for number in household less than age 18, '5' includes households with at least five youths.

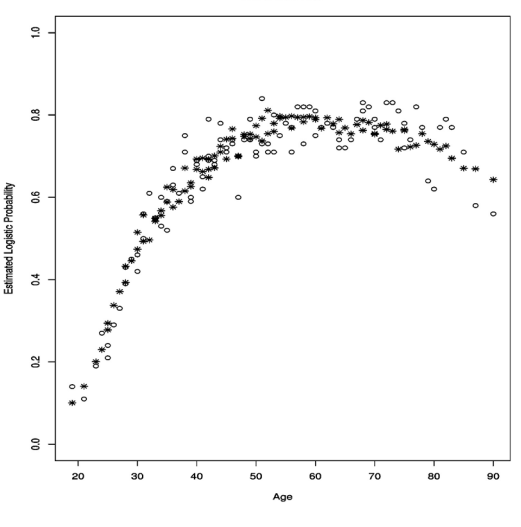

FIGURE 10 Diagnostics plot for age for updated logistic regression model for property tax.

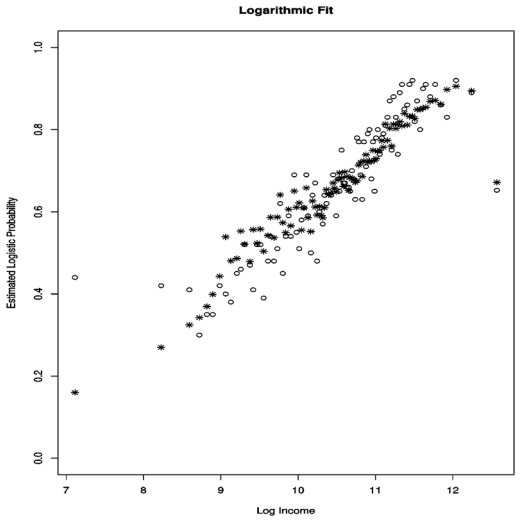

FIGURE 11 Diagnostics plot for log(income) for updated logistic regression model for property tax.

They deviate most noticeably for **I** and **G**, particularly at high values of those variables. To improve model fit, we add $\mathbf{G}^2$ and replace **I** with log(**I**). The quadratic function of age is plausible: younger and older individuals are less likely to own a home than middle-aged individuals. Taking the logarithm of income pulls in extreme values of income. There also appears to be a difference in the predicted and perturbed percentages for households of size at least seven, which suggests adding a quadratic term in household size may improve fit.

On the basis of these diagnostics, we fit the updated logistic regression model with the predictor function:

$$1 + \log(\mathbf{I}) + \mathbf{G} + \mathbf{G}^2 + \mathbf{E} + \mathbf{N} + \mathbf{N}^2 + \mathbf{Y} + M + X + R \tag{2}$$

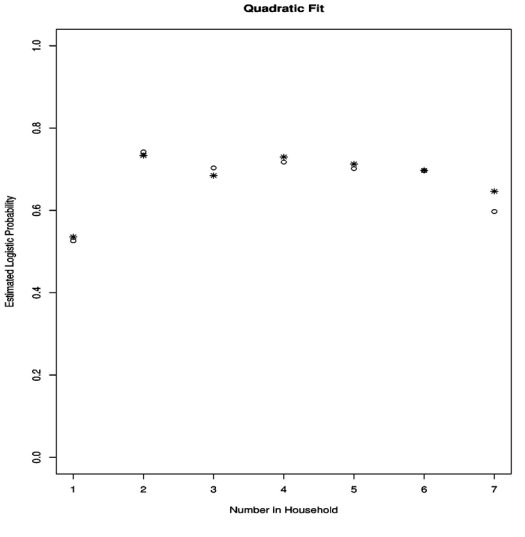

FIGURE 12 Diagnostics plot for number in household for updated logistic regression model for property tax.

TABLE IV Δ of deviance test statistics for updated logistic regression model for property tax. The baseline model is the initial model in Eq. (2).

Added term	*Δ Deviance*	*Δ df*	*Proposed model*	*Comparison model*
$\mathbf{G}^2$	150.3	1	$(1) + \mathbf{G}^2$	(1)
$\log(\mathbf{I})$	47.8	1	$(1) + \mathbf{G}^2 - \mathbf{I} + \log(\mathbf{I})$	$(1) + \mathbf{G}^2$
$\mathbf{N}^2$	9.5	1	$(1) + \mathbf{G}^2 - \mathbf{I} + \log(\mathbf{I}) + \mathbf{N}^2$	$(1) + \mathbf{G}^2 - \mathbf{I} + \log(\mathbf{I})$

Figures 10–12 show the remote server diagnostics for the updated model for age, income, and household size. The predicted probabilities for the updated model are closer to the perturbed percentages than those for the initial model, indicating improvement in the model fit. The additional predictors are statistically significant, as evident in the change of deviance test statistics displayed in Table IV.

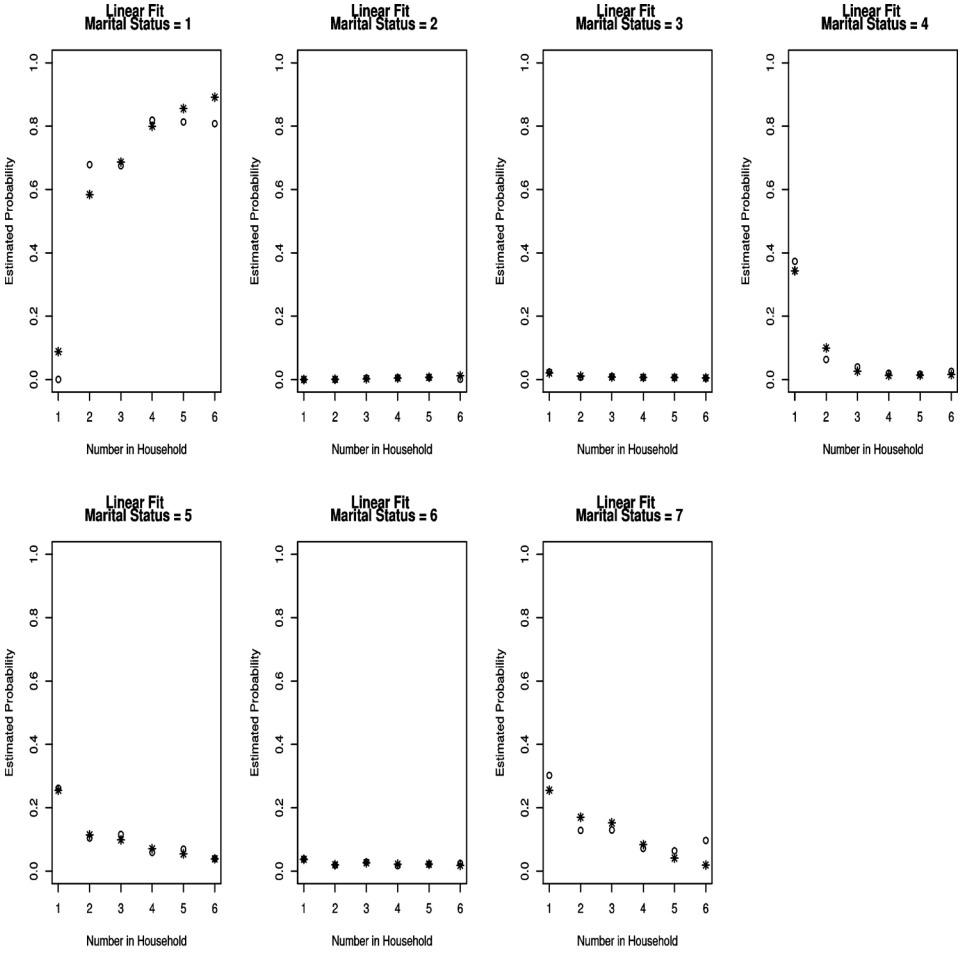

FIGURE 13 Diagnostics using linear function of number in household for multinomial regression for marital status.

4.2 Marital Status Multinomial Regression

For the multinomial regression for the predicting marital status, the initial function of the predictors is:

$$1 + \mathbf{P} + \mathbf{I} + \mathbf{G} + \mathbf{N} + \mathbf{Y} + \mathbf{E} + X + R \tag{3}$$

As before, when constructing the remote server diagnostics, we categorize the continuous variables in groups of 100, except for those with prespecified levels which required some form of level collapsing. We then added random noise from $(-2, -1, 1, 2)$ to the observed number in each category.

The resulting graphical displays are too numerous to present here. We focus on the diagnostic plots involving number of people in the household, displayed in Figure 13. To protect confidentiality, we combine all households with six or more people into one category to attain sufficient numbers; there are few households of those sizes and they pose a potential disclosure risk.

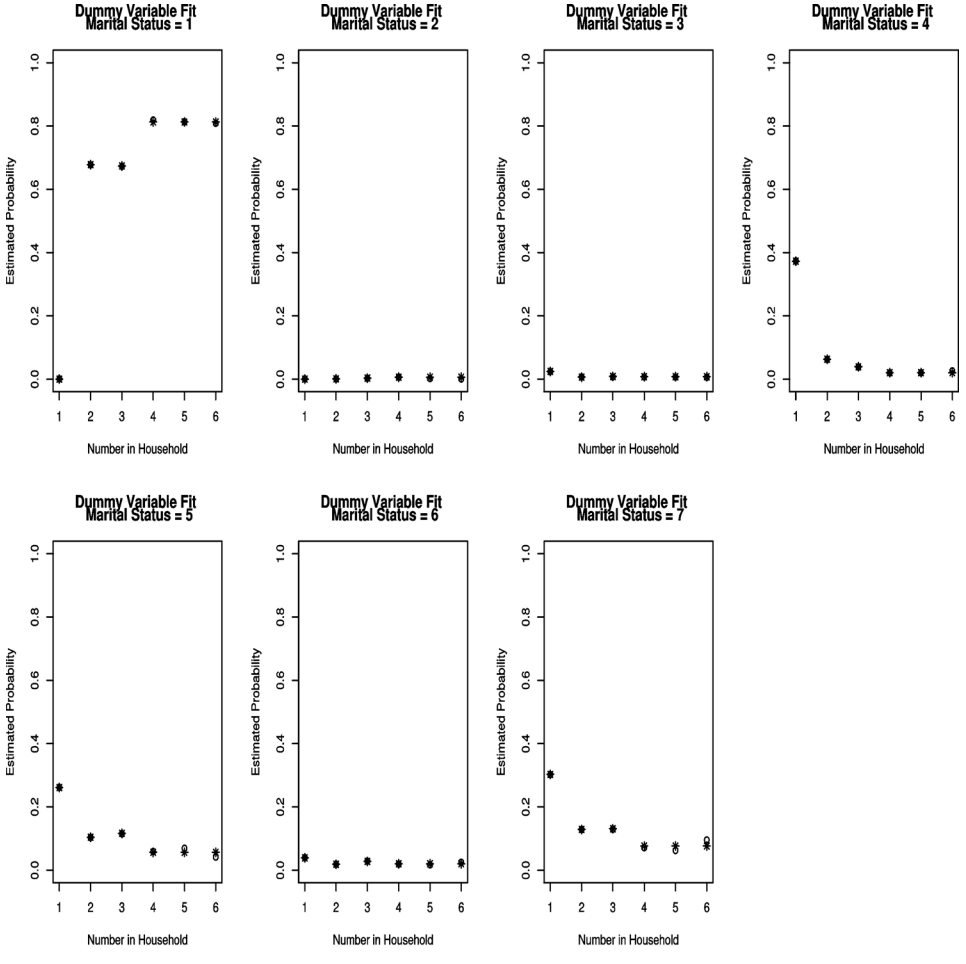

FIGURE 14 Diagnostics using indicator variables for number in household for multinomial regression for marital status.

The linear trend in the predicted probabilities does not match the perturbed percentages. There are several large discrepancies for $M = 1$ and 7. These indicate that the model does not fit the data well and can be improved. Owing to the discrete nature of household size, it is reasonable to model N as a series of indicator variables. We note that once $N > 3$, the perturbed percentages appear to be fairly steady. Hence, we are led to fit N as a series of indicators, one each for $N = 1, 2, 3$, and ≥ 4, using the latter as a baseline category. The resulting model fits the data much better, as indicated by the updated diagnostics in Figure 14.

The diagnostic plots for the other variables, not shown here, suggest modeling number of youths in the household as a series of indicator variables, and adding a quadratic term for age. The other variables appear reasonably modeled with main effects and linear functions.

5 CONCLUDING REMARKS

The simulation studies indicate that the diagnostics proposed here can improve the utility of remote access servers relative to releasing only the basic output from queries, without substantially compromising confidentiality. The general approach of these diagnostics can be applied for any type of link function submitted by the user, including non-linear or spline functions of the predictors. The diagnostics may not work well when the probabilities of 'success' for the dependent variable change radically within small regions of $\mathbf{x}_p$, especially when the partition sizes must be relatively large to ensure data confidentiality. This type of risk/utility tradeoff is typical of disclosure limitation techniques.

The ideas underpinning these diagnostics can aid model construction. For example, for large data sets, one approach is to split the data set into training and evaluation units. The training units are used to estimate coefficients of the models, and the evaluation units are used to check the model fit. To assist users' model specification, the server can generate the diagnostics on the basis of the training units' data, as done in this article. The diagnostics also can be fruitfully applied to the evaluation units' data. That is, for the evaluation data, the server releases the medians of categorized versions of the $\mathbf{x}_p$, the averages of the predicted probabilities, and the averages of the perturbed percentages. Large discrepancies in the average predicted probabilities and the perturbed percentages indicate lack of fit.

Acknowledgement

This research was supported by NSF grant EIA-0131884 to the National Institute of Statistical Sciences.

References

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference*, **6**, 73–85.

Duncan, G. T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, **95**, 720–729.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, **9**, 383–406.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. and Van Mechelen, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics*, **49**, 247–268.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, New York.

Keller-McNulty, S. and Unger, E. A. (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, **14**, 347–360.

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models (with discussion). *Journal of the American Statistical Association*, **79**, 61–71.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models: Second Edition*. Chapman & Hall, London.
Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–724.
Reiter, J. P. (2003). Model diagnostics for remote access servers. *Statistics and Computing*, 371–380.
Rowland, S. (2003). An examination of monitored, remote access microdata access systems. In *National Academy of Sciences Workshop on Data Access*.
Schouten, B. and Cigrang, M. (2003). Remote access systems for statistical analysis of microdata. *Statistics and Computing*, 381–389.
Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer-Verlag, New York.

Taylor & Francis
Taylor & Francis Group

Journal **…J. Statist. Comput. Simul.,** Article ID **…GSCS 041058**

TO: CORRESPONDING AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer the queries.

Q1	Please check the author affiliation.	

Production Editorial Department, Taylor & Francis Ltd.
4 Park Square, Milton Park, Abingdon OX14 4RN

Telephone: +44 (0) 1235 828600
Facsimile: +44 (0) 1235 829000