Model Diagnostics for Remote Access Regression Servers

Jerome P. Reiter*

Key Words: Confidentiality, Diagnostics, Disclosure, Regression, Remote access, Synthetic Data

Abstract

To protect public-use microdata, one approach is not to allow users access to the microdata. Instead, users submit analyses to a remote computer that reports back basic output from the fitted model, such as coefficients and standard errors. To be most useful, this remote server also should provide some way for users to check the fit of their models, without disclosing actual data values. This paper discusses regression diagnostics for remote servers. The proposal is to release synthetic diagnostics–i.e. simulated values of residuals and dependent and independent variables–constructed to mimic the relationships among the real-data residuals and independent variables. Using simulations, it is shown that the proposed synthetic diagnostics can reveal model inadequacies without substantial increase in the risk of disclosures. This approach also can be used to develop remote server diagnostics for generalized linear models.

1 Introduction

To protect public-use microdata, one approach is not to release any microdata at all. Instead, agencies can require users to submit analyses to a remote computer that reports back basic output from the fitted model, such as coefficients and standard errors. This approach has been suggested by several authors (e.g., Keller-McNulty and Unger, 1998; Duncan and Mukherjee, 2000), and versions of remote access servers have

*Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708. This research was completed while the author was a Senior Fellow at the U.S. National Institute of Statistical Sciences. It was supported by NISS as part of the U.S. National Science Foundation digital government project.

been developed by Statistics Canada (Mantel and Nadon, 1999; Bustros, 2000), the U.S. National Center for Health Statistics, and the U.S. Bureau of the Census. Remote regression servers also are being developed by Statistics Netherlands and the U.S. National Institute of Statistical Sciences. See Schouten and Cigrang (2003) for a general discussion of the issues involved in setting up remote servers.

The remote access approach has an advantage over methods of disclosure avoidance that swap, recode, or add noise (e.g., see Willenborg and de Waal, 2001) to the original data: parameter estimates are based on the real, unmasked data. Additionally, confidentiality may be protected more effectively since no real microdata are released. To be most useful, a remote server should provide some way for users to check the fit of their submitted models, without disclosing actual data values. However, such methods are not part of existing remote access servers. Furthermore, diagnostic methods for remote servers do not appear in published statistical or computing literature.

In this paper, I propose remote server diagnostic methods for linear regressions. Specifically, I propose that remote servers provide synthetic, i.e. simulated, values of dependent and independent variables, residuals, and fitted values (values of the dependent variable on the estimated regression line). Users then can treat these synthetic values like ordinary diagnostic quantities, for example by examining scatter plots of the synthetic residuals versus the synthetic independent variables or versus the synthetic fitted values. Synthetic values of dependent and independent variables and fitted values are drawn from their marginal distributions as approximated by normal kernel density estimators (Wegman, 1972). Corresponding synthetic residuals are generated by (i) fitting generalized additive models (Hastie and Tibshirani, 1990) for the real-data residuals on each of the real-data independent variables and real-data fitted values, (ii) using these models to predict residuals at the values of the synthetic independent variables and synthetic fitted values, and (iii) adding noise to these predictions. This approach can be used for remote server diagnostics for other generalized linear models; work on such diagnostics is in progress. The synthetic diagnostics are illustrated with several simulated data sets and a subset of data from the U.S. Current Population Survey.

2 Desiderata for remote server diagnostics

To construct and evaluate diagnostics for remote servers, I consider three main criteria: confidentiality, utility, and feasibility. These are similar in spirit to the criteria developed by Duncan *et al.* (2001) and Schouten and Cigrang (2003).

2.1 Confidentiality

The diagnostics should not substantially compromise confidentiality relative to releasing only coefficients and standard errors. For example, the server should not release real-data residuals, since they can be used to determine the actual values of the dependent variable. Releasing any useful diagnostic variable adds to the risk of disclosures; we seek to limit the additional risk.

2.2 Utility

The diagnostics should enable users to obtain better-fitting models relative to releasing only coefficients and standard errors. Because synthetic diagnostics at best approximate real-data diagnostics, we expect they may miss some model inadequacies that can be revealed using real-data diagnostics. We seek remote server diagnostics that do not miss too many of them.

2.3 Feasibility

The diagnostics should be provided almost immediately to the user after he or she requests them. This is particularly important for users who submit many queries to the remote server. The diagnostics also should be easy for adminstrators to incorporate as part of their remote servers.

3 Method of generating synthetic diagnostics

The remote server diagnostics are generated in a two-step procedure. First, before the server is made accessible to outside users, synthetic values are generated for all variables in the data file. Second, once the

server is on-line and a regression is submitted by some user, synthetic residuals and synthetic fitted values are generated dynamically for that regression. The synthetic variables, synthetic fitted values, and synthetic residuals are provided in spreadsheets or graphical displays when requested by the user.

3.1 Generation of synthetic variables and synthetic fitted values

For some collected data set with d variables, let $\mathbf{x}_p$, where $p = 1, \ldots, d$, be a variable in that data set. Before making the server accessible to outside users, the administrator generates synthetic variables $\mathbf{x}_p^s$, for $p = 1, \ldots, d$, corresponding to the original d variables. Hereafter, a superscript s stands for "synthetic." The generated $\mathbf{x}_p^s$ are stored on the remote server, so that all requests involving particular subsets of variables are provided with the same synthetic values of these variables. When users submit analyses with transformations of variables, the stored values of the $\mathbf{x}_p^s$ are suitably transformed and provided. The range of each $\mathbf{x}_p^s$ is required to lie within the range of the real-data $\mathbf{x}_p$, so that users know approximately how far they can extrapolate regression predictions. To save storage space, synthetic values for interactions among independent variables are generated dynamically, i.e. only when users fit such terms in the models. The same random seed is used for all interactions, so that users always obtain the same synthetic values whenever any particular interaction is included in the submitted model.

For categorical $\mathbf{x}_p$, the $\mathbf{x}_p^s$ are generated by sampling with replacement from the real-data values. The remote server can release the real-data $\mathbf{x}_p$ when it is safe to provide the exact number of units in each category. Special care needs to be taken when some categories of $\mathbf{x}_p$ are sparse, as shall be discussed in Section 3.2.

For continuous $\mathbf{x}_p$, first a normal kernel density estimator is fit to the real-data values, estimating the values of the density curve at 100,000 evenly spaced points bounded by the minimum and maximum values of $\mathbf{x}_p$. Then, to generate $\mathbf{x}_p^s$, random samples are taken from the density curve using an inverse-cdf method. This is computationally fast, and appropriate routines exist in commercially available software. This same procedure can be used to draw synthetic fitted values for submitted regressions. To simplify notation, the real-data and synthetic fitted values for the submitted regression are written as $\mathbf{x}_0$ and $\mathbf{x}_0^s$, respectively, and

labels tieing the fitted values to particular regressions are omitted; i.e., fitted values are indexed by $p = 0$.

For all p, the $\mathbf{x}_p^s$ are drawn marginally. Hence, when synthetic variables or fitted values are provided in a spreadsheet, users must be informed that the rows in the spreadsheet do not correspond to units, so that analyses of relationships among these variables are meaningless. Drawing marginally from nonparametric distributions is computationally faster and conceptually easier than drawing jointly from nonparametric distributions. It also may help protect confidentiality by providing less information for attackers seeking to match synthetic values to external data. A downside to drawing marginally is that the utility of the diagnostics is reduced. For example, in a plot of synthetic residuals versus a synthetic independent variable, the values of a third independent variable cannot be meaningfully superimposed on the plot. This makes it harder to identify interaction effects using these plots. The methods of generating synthetic residuals in this paper can be applied when synthetic variables are drawn jointly.

The number of synthetic values in an $\mathbf{x}_p^s$ need not equal the number of units used in the regression, although generally this is easiest to implement. For example, administrators may choose to reduce the number of synthetic values in an $\mathbf{x}_p^s$ to protect confidentiality further. Admininstrators also should carefully examine the $\mathbf{x}_p^s$ before putting the remote system on-line to make sure that confidentiality is sufficiently protected and the marginal distributions of the $\mathbf{x}_p$ are reasonably reproduced.

I also investigated adding Gaussian noise to the original values to generate synthetic values. This approach can fail to reflect important aspects of the distributions of the variables. For example, it does not capture point masses at zero for monetary variables. Hence, I do not recommend generating values of variables or fitted values by solely adding random noise to the original values.

3.2 Generation of synthetic residuals

In the development below, labels tieing real-data or synthetic values to particular regressions are omitted to simplify notation. Of course, each regression has its own real-data residuals and real-data fitted values, and the remote server would generate different synthetic residuals and synthetic fitted values for each regression.

For $i = 1, \ldots, n$, let e_i be the real-data residual for unit i when a regression is fit using the real data.

The real-data standardized residual for that same unit in that regression equals

$$t_i = e_i / \hat{\sigma} \sqrt{(1 - h_i)} \tag{1}$$

where $\hat{\sigma}$ is the estimated root mean squared error of the regression, and h_i is the ith diagonal element of the hat matrix for the regression. Let $\mathbf{t}$ be the set of the t_i for the regression for all units i. When the regression fits the data, all elements of $\mathbf{t}$ have variance equal to one, which makes $\mathbf{t}$ preferable to the e_i for diagnosing non-constant variance. Standardized residuals also are more convenient to simulate than ordinary residuals, because noise used to generate synthetic standardized residuals can come from distributions with variances on the same scale; this is explained below. Therefore, standardized residuals are used for the remainder of this paper.

Let t^s_{kp} be the synthetic, standardized residual attached to synthetic value k of variable p in the fitted regression. We seek to generate the t^s_{kp} so that the relationship between $\mathbf{t}^s_p$ and $\mathbf{x}^s_p$ looks like the relationship between $\mathbf{t}$ and $\mathbf{x}_p$. To this end, each t^s_{kp} is determined as follows:

$$t^s_{kp} = b_{kp} + v_{kp} + n_{kp}. \tag{2}$$

The b_{kp} places t^s_{kp} on a curve consistent with the general relationship between $\mathbf{t}$ and $\mathbf{x}_p$, and the v_{kp} moves the synthetic residual off that curve in a way that is consistent with the variation in the real-data residuals near $x_{ip} = k$. The n_{kp} is noise added to reduce the risk of disclosing the values of the real-data residuals. These three pieces are described in detail below.

To determine b_{kp} when p indexes the fitted values (i.e., $p = 0$) or a continuous independent variable, a smooth curve is fit to the relationship between $\mathbf{t}$ and $\mathbf{x}_p$ using a generalized additive model (Hastie and Tibshirani, 1990). The b_{kp} equals the value of this curve at k. For the generalized additive model, I use a Gaussian link function and locally weighted linear regression, i.e. loess, smoothers (Cleveland, 1979) with a smoothing parameter equal to 0.5, and fit the model using local scoring and the backfitting algorithm

of Hastie and Tibshirani (1990, Chapter 6). These particular specifications are defaults for the "gam" routine for fitting generalized additive models in the software package *S-Plus* (Venables and Ripley, 1997); administrators can use other specifications when they better describe the relationship between $\mathbf{t}$ and $\mathbf{x}_p$. When p indexes a categorical variable, $b_{kp} = 0$ because a smooth curve between $\mathbf{t}$ and $\mathbf{x}_p$ is not needed; all values in $\mathbf{x}_p^s$ are in $\mathbf{x}_p$.

To determine each v_{kp}, first the unit j is found such that $j = \arg\min_i |k - x_{ip}|$; this is the unit whose value in the real-data $\mathbf{x}_p$ is closest to the synthetic value k. When more than one unit satisfies the arg-min condition, unit j can be obtained by sampling randomly from the qualifying units. When $p = 0$ or when p indexes a continuous independent variable, $v_{kp} = t_j - b_{jp}$, where b_{jp} is the value at x_{jp} on the curve obtained from the generalized additive model. When p indexes a categorical independent variable, $v_{kp} = t_j$. Effectively, this randomly selects a standardized residual from the units with $x_{ip} = k$.

Each n_{kp} is drawn from an independent $N(0, \tau)$, where τ is specified by the administrator of the remote server. Different values of τ can be used for different regressions. However, a single τ should be used for all synthetic residuals from the same regression, so as not to introduce artificially non-constant variance in the synthetic residuals. All queries that use the same dependent variable should use a common random seed to generate the n_{kp}. This prevents users from refining any guesses about a real-data t_i by averaging synthetic residuals from repeated calls to the same or similar regressions.

In many cases, setting $\tau = 1$ should provide adequate disclosure protection when releasing synthetic residuals. To see this, it is helpful to mimic potential behavior of an attacker. Assume the attacker knows the exact values of the independent variables for some unit i and, therefore, is able to obtain a reliable fitted value for that unit. Further, assume the attacker somehow identifies a released t_{kp}^s as being relatively close to t_i. Knowing the t_{kp}^s contains additive noise with unit variance, the attacker can form an approximate 95% prediction interval for t_i as

$$t_{kp}^s \pm 2. \tag{3}$$

Using the fitted value, the attacker can transform (3) to a 95% prediction interval for the value of the dependent variable. The width of this interval is roughly $2\hat{\sigma}$, which equals the width of the usual prediction interval obtained from the regression output. Hence, relative to releasing only basic regression output, attackers gain little from the additional release of synthetic residuals, at least for units whose fitted values fall close to the regression line.

In other cases, remote server administrators may want to set τ at values other than one. For example, the administrator might select $\tau = \frac{w\hat{\sigma}_y^2}{\hat{\sigma}^2} < 1$, where w is some small constant and $\hat{\sigma}_y^2$ is the unconditional variance of the dependent variable in the real data. Alternatively, to mitigate the effects of outliers, the administrator might select τ proportional to a ratio of a robust estimate of the regression variance over $\hat{\sigma}^2$. When $\tau < 1$, the differences between the synthetic and real-data residuals are typically smaller than when $\tau \geq 1$, so that the synthetic diagnostics mimic more closely the relationships between $\mathbf{t}$ and the $\mathbf{x}_p$. However, when $\tau < 1$, an attacker can obtain a 95% interval for some t_i with narrower width than when $\tau \geq 1$. Conversely, setting $\tau > 1$ can improve confidentiality protection but possibly diminish the utility of the diagnostics. Such confidentialty/utility trade-offs are common for disclosure avoidance techniques that involve adding noise to actual values (Fuller, 1993; Duncan and Mukherjee, 2000; Duncan *et al.*, 2001).

Releasing synthetic residuals may not adequately protect units whose fitted values are far from the regression line. For example, suppose for some regression unit i has $t_i = 15$, and the attacker knows there is only one unit in the population with an extreme value of the dependent variable. Releasing a synthetic standardized residual in the range of $13 \leq t_{kp}^s \leq 17$ may be considered a disclosure by some administrators, particularly if $\hat{\sigma}$ is relatively small. To protect units with extreme residuals, administrators can top-code synthetic residuals, e.g., by forcing all $|t_{kp}^s| \leq C$. Setting the top-code limit at some value C corresponds roughly to not revealing synthetic residuals more than C root mean squared errors from the estimated regression line.

Inappropriate disclosures also can occur when p indexes a categorical variable with some sparse levels. For example, suppose the attacker knows there is only one unit i with a certain value k of some $\mathbf{x}_p$. Further, suppose the server generates four values of $\mathbf{x}_p^s$ equal to that k. By fitting a simple regression of the sensitive

variable on $\mathbf{x}_p$, the attacker can obtain a refined estimate of the t_i by averaging the four values of t^s_{kp}. To avoid such problems, administrators can implement the restriction that no level of a categorical variable is released unless there is a minimum number of people in the category. This is similar in spirit to some release rules for tabular data (Willenborg and de Waal, 2001).

A related problem may occur when the dependent variable takes on a small set of discrete values. For example, suppose the dependent variable is ordered from one to seven, and the solitary independent variable is dichotomous. There are at most fourteen distinct values of real-data residuals. It is possible that large, positive residuals exist only when the dependent variable equals seven, or that large negative residuals exist only when the dependent variable equals one. When the synthetic residuals mimic this structure, they may disclose the dependent variable for some units. To lessen the chance of such disclosures, administrators can increase the variance of the additive noise, τ, in the synthetic residuals.

The next two sections illustrate the performance of these diagnostics with two sets of simulation studies. The first set is comprised of simple regressions and a variety of relationships between the dependent and independent variable. These data sets can be considered litmus tests for the procedures. The second set is comprised of multiple regressions using data from the U.S. Current Population Survey.

4 Simulation A: Some litmus tests with simulated data

The simulations include eight data sets with $n = 100$ observations each. Each data set is comprised of an independent variable, x_1, whose values are drawn randomly from a normal distribution with mean equal to five and variance equal to one. Values of the dependent variable, x_2, differ across the data sets. They are described in Table 1.

For each of these eight scenarios, the model is an ordinary least squares regression of the untransformed x_2 on the untransformed x_1. Synthetic diagnostics for the models are generated using the methods of Section 3, setting $\tau = 1$ for all regressions. The "density" routine in the software package *S-Plus* is used to fit the normal kernel density estimators for generating the $\mathbf{x}^s_p$. The "gam" routine in *S-Plus*, with default settings,

Description	Generation model for dependent variable
Good fit, $R^2 = .99$	$x_2 = 10x_1 + \epsilon$
Good fit, $R^2 = .50$	$x_2 = x_1 + \epsilon$
Good fit, $R^2 = .02$	$x_2 = .25x_1 + \epsilon$
Heteroscedastic	$x_2 = 10x_1 + \epsilon, \quad \epsilon \sim N(0, x_1^2)$
Influential point	$x_2 = 10x_1 + \epsilon + 200\delta$, where $\delta = 1$ for the largest value of x_1.
Outliers	$x_2 = 10x_1 + \epsilon + 30\alpha - 25\beta$, where $\alpha = 1$ for $x_1 = 3.8$, and $\beta = 1$ for $x_1 = 5.8$
Curvilinear	$x_2 = 10(x_1 - 5) - 5(x_1 - 5)^2 + \epsilon$
Piecewise	$x_2 = 20 + \epsilon$, for $x_1 < 4.5$
	$x_2 = 10 + 8x_1 + \epsilon$, for $4.5 \leq x_1 < 5.5$
	$x_2 = -4 + 14x_1 + \epsilon$, for $x_1 \geq 5.5$

Table 1: Data generation models. Unless indicated, the ϵ are drawn from independent, standard normals.

is used to fit the generalized additive models. Plots of real-data and synthetic standardized residuals versus real-data and synthetic independent variables are displayed in Figures 1 - 4.

For the three good-fitting regressions (Figure 1), the real-data and synthetic-data plots look similar. None of the plots suggest violations of the regression assumptions. For the curvilinear and piecewise regressions (Figure 2), the synthetic-data plots reveal the violations of the regression assumptions, albeit not as sharply as the real-data plots due to the additive noise. For the piecewise regression, the skilled data analyst can recognize three clouds of points: one less than 4.5, one between 4.5 and 5.5, and one greater than 5.5. For the outlier and influential point regressions (Figure 3), the synthetic-data plots reflect the impact of the points with extreme residuals. However, the synthetic-data plots do not include some of these points. This is because no synthetic values of x_1 are close enough to the real-data values of x_1 associated with these points. For the heteroscedastic regression (Figure 4), the synthetic data plot shows the classic fan-shape associated with non-constant variance.

The failure of the synthetic diagnostics to reveal all the influential and outlying points suggests that the server should report ranges of the values of the $\mathbf{x}_p$ associated with all unusual points as part of the diagnostic output. These ranges should be determined so that they do not result in disclosures. Additionally, the fuzziness in the curvilinear, piecewise, and heteroscedastic plots suggests that weak relationships could be

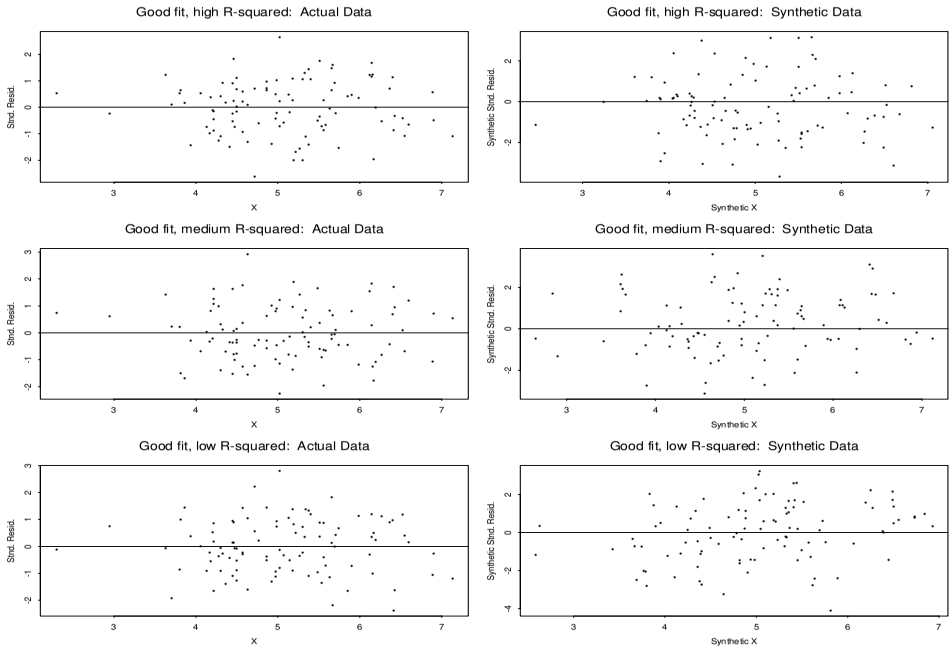

Figure 1: Side-by-side comparisons of real-data and synthetic-data plots of standardized residuals versus the independent variable for the three good-fitting regressions.

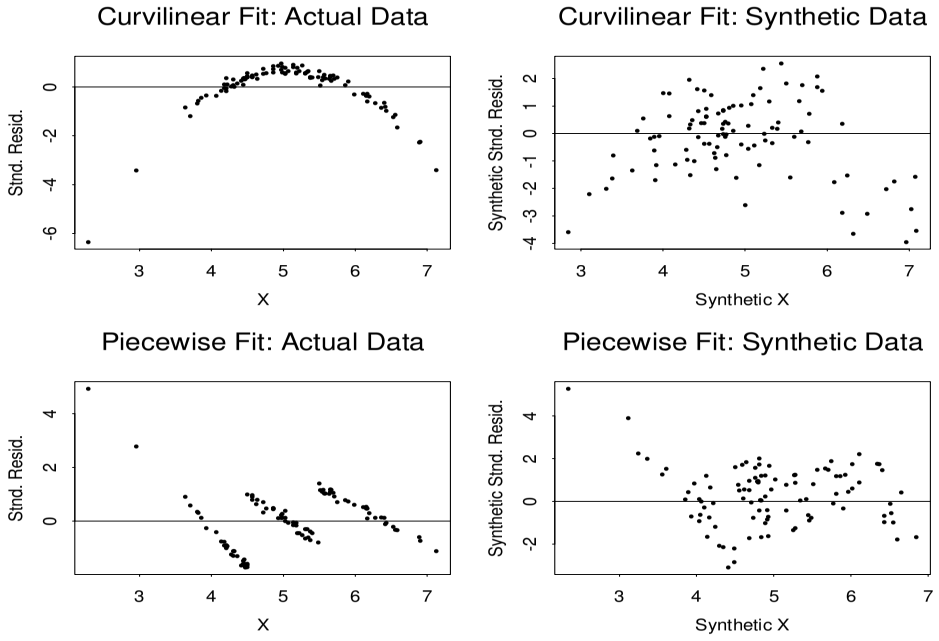

Figure 2: Side-by-side comparisons of real-data and synthetic-data plots for the curvilinear and piecewise scenarios.

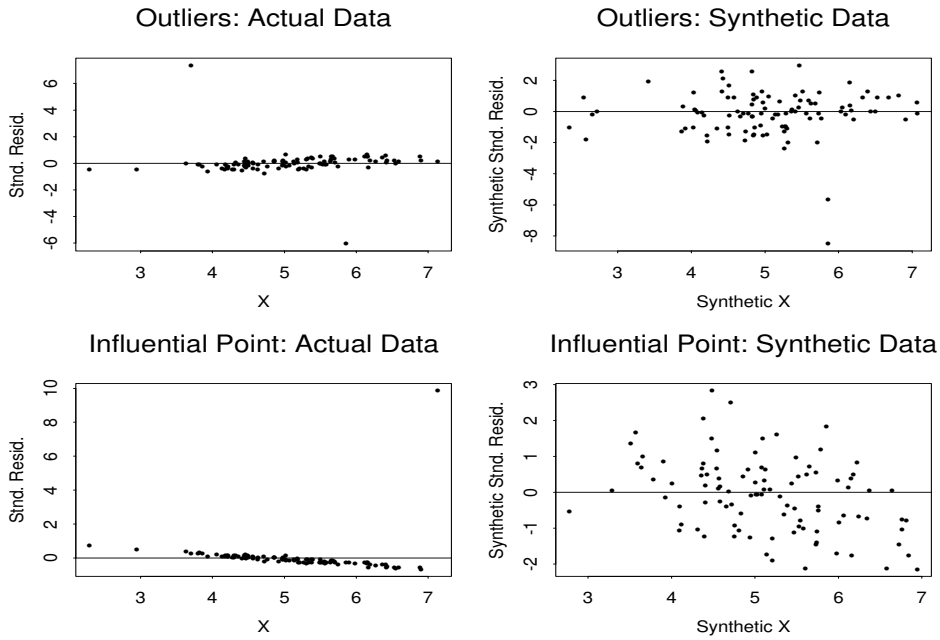

Figure 3: Side-by-side comparisons of real-data and synthetic-data plots for the outlier and influential point scenarios.

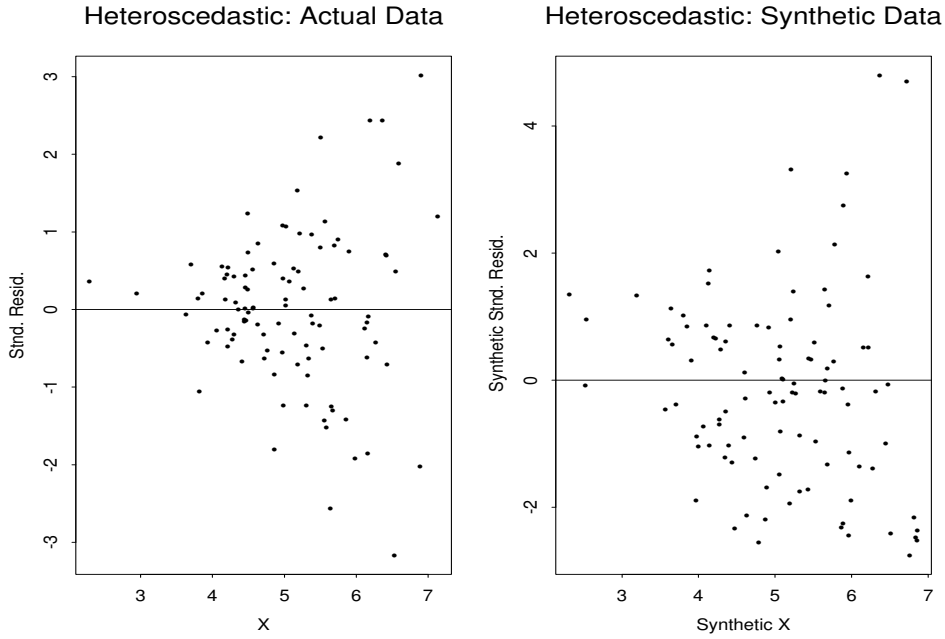

Figure 4: Side-by-side comparisons of real-data and synthetic-data plots for the heteroscedastic scenario.

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

Table 2: Variables used in CPS data simulations.

obscured by the added noise. Nonetheless, these synthetic diagnostics clearly have positive utility: they show model violations that would go unnoticed from solely the basic output of the regression.

5 Simulation B: Performance on genuine data

This set of simulations examines multiple regressions based on a subset of public release data from the March 2000 U.S. Current Population Survey. The data are comprised of 10,000 randomly sampled heads of households from the public-use data file. The variables of interest are shown in Table 2.

Marital status, M, has seven types, ranging from $M = 1$ for married civilians with both spouses present at the home to $M = 7$ 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, 6,612 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 a prediction model for positive property taxes as a function of the other variables. For simplicity, complications due to the complex sampling design are ignored, and the 6,612 households with $P > 0$ are treated as a simple random sample. In the descriptions of the models below, a bold-faced letter indicates that its corresponding independent variable is fit as a continuous variable. A plain letter indicates

that its corresponding variable is fit as a series of dummy variables. The notation ("$Letter'' < c$) represents an indicator that equals one when the variable associated with "Letter" is less than the value c, and it equals zero otherwise. These models are not the best models for predicting property taxes, but they are useful for illustrating the synthetic diagnostics.

The dependent variable is $\log(P)$. Since property taxes are monetary variables, we can expect many analysts to use this transformation *a priori* of looking at the data. Analysts also could examine the distribution of released synthetic property tax values to discover the long right tails. The initial formulation of the conditional expectation of $\log(P)$ is

$$E\left(\log(P)\right) = \mathbf{I} + \mathbf{G} + \mathbf{E} + \mathbf{N} + \mathbf{Y} + M + X + R \tag{4}$$

Figures 5 and 6 display plots of real-data and synthetic residuals versus fitted values and versus four selected independent variables for the model in (4). Synthetic diagnostics are generated using the methods of Section 3, setting $\tau = 1$. When simulating values, income is treated as continuous and all other independent variables are treated as categorical. Generating the synthetic diagnostics requires only a few seconds of computer time on a typical workstation.

Both the real-data and synthetic-data plots in Figure 5 show that the spread of the residuals decreases as the fitted values increase. This suggests some violation of the regression assumptions. Unlike the real-data plot, the synthetic-data plot does not include a fitted value near 10.5. This does not affect the overall conclusion from examining the synthetic-data plot.

For all four independent variables, the plots of synthetic residuals versus synthetic independent variables in Figure 6 contain patterns like those in the real-data plots. For income, the residuals in both plots decrease as income increases. The data analyst might suspect that units with large incomes strongly influence the estimated regression line, so that a model using $\log(I)$ in place of I may fit better. Unlike the actual incomes, the synthetic incomes do not include values over 500,000. As mentioned in Section 3.1, the remote server administrator could include synthetic income values in this range before putting the system online. The

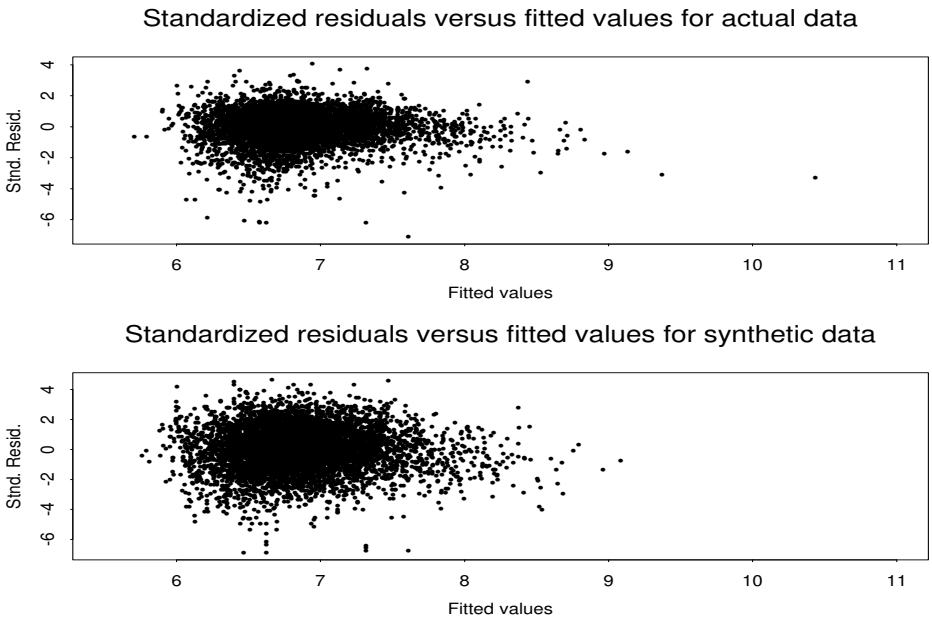

Figure 5: Plots of standardized residuals versus fitted values for the real data and the synthetic data for the regression model in Equation 4.

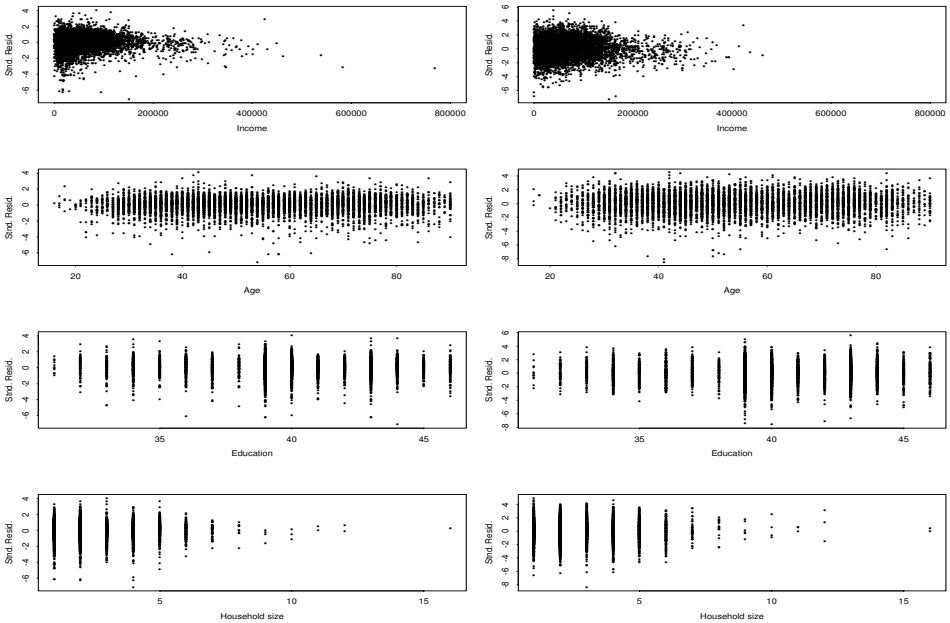

Figure 6: Plots of standardized residuals versus fitted values for the real data and the synthetic data for the regression model in Equation 4. Real-data plots are on the left, and corresponding synthetic-data plots are on the right.

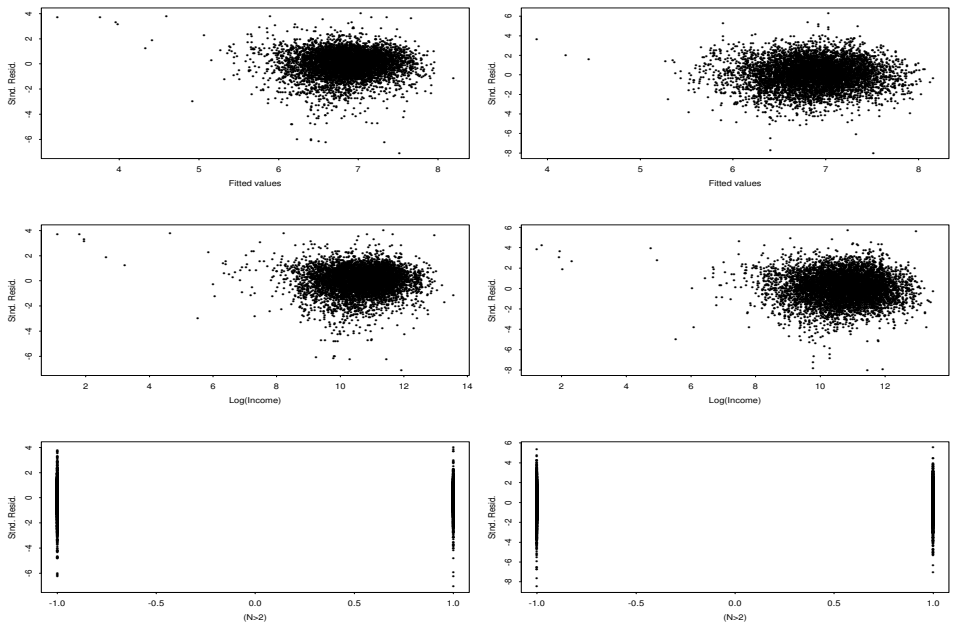

Figure 7: Plots of standardized residuals versus fitted values for the real data and the synthetic data for the modified regression model. Real-data plots are on the left, and corresponding synthetic-data plots are on the right.

plots involving age and education do not indicate any lack of fit. The plots involving household size suggest it may be profitable to replace **N** with the indicator variable $(N > 2)$.

Figure 7 displays plots of standardized residuals versus fitted values for a regression that uses $\log(\mathbf{I})$ in place of **I**, and $(N > 2)$ in place of **N**, in (4). To save space, only the plots involving the fitted values, income, and number in household are shown. Once again, the real-data and synthetic-data plots contain similar patterns. The coefficient of determination, R^2, increases by about 1.5% after these transformations, indicating slightly better fit. Such improvement is evidence that releasing synthetic diagnostics can help users obtain better-fitting models when compared to releasing only coefficients and standard errors.

6 Concluding Remarks

The arguments for remote regression servers are compelling. With remote servers, there is no need to release any real data, and users can obtain valid inferences for a wide class of analyses. Similar potential exists for the synthetic data approach of Rubin (1993): release microdata that are simulated from probability distributions. This idea, and extensions to it, has been examined by several authors, including Kennickell (1997); Fienberg *et al.* (1998); Abowd and Woodcock (2001); Raghunathan *et al.* (2003); Reiter (2002, 2003a,b); Franconi and Stander (2003); Muralidhar and Sarathy (2003); and Polettini (2003). One advantage of the regression server over releasing synthetic microdata is that inferences are based on actual rather than simulated data. A disadvantage is that users must work remotely instead of having synthetic data on their own computers.

The remote server diagnostics developed here perform well in several simulation studies. This suggests that, by incorporating these diagnostics into remote servers, administrators can provide remote server users with some way of checking whether their models fit the data. Of course, as with all disclosure avoidance methods, synthetic diagnostics should not be adopted carelessly. Remote server administrators should test for disclosures by running their own regressions before opening the system to the public, particularly when some variables have regions with sparse data. Administrators also may want to use noise with variance other than one when generating synthetic residuals. Increasing this variance on average increases the absolute distances between the synthetic and real-data residuals, thereby improving confidentiality protection; however, subtle trends in the residual plots can get swamped by increased noise, thereby potentially weakening diagnostic utility. Similarly, decreasing the noise variance can improve utility at the expense of potentially weakening confidentiality protection. Administrators can arrive at an acceptable noise variance by using the regression output and synthetic diagnostics in simulated attacks involving the real data of interest.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and*

Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.

Bustros, J. (2000). Access to microdata files at Statistics Canada. In *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 61–68.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Tech. rep., U.S. National Institute of Statistical Sciences.

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.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

Franconi, L. and Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing* forthcoming.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman & Hall.

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.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

Mantel, H. and Nadon, S. (1999). Dummy file creation for the remote access program of the National Population Health Survey. In *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 181–186.

Muralidhar, K. and Sarathy, R. (2003). A theoretical basis for perturbation methods. *Statistics and Computing* forthcoming.

Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing* forthcoming.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* forthcoming.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003a). Inference for partially synthetic, public use microdata sets. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Reiter, J. P. (2003b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Tech. rep., Institute of Statistics and Decision Sciences, Duke University.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Schouten, B. and Cigrang, M. (2003). Remote access systems for statistical analysis of microdata. *Statistics and Computing* forthcoming.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Wegman, E. J. (1972). Nonparametric probability density estimation. *Technometrics* **14**, 533–546.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.