False Discovery Rate in Large-Scale Data Error Localization: Assessing Survey Editing Methods' Performance

Paul J. Smith[1], Chin-Fang Weng[2], Eric V. Slud[3]
[1]University of Maryland, College Park (Retired)
[2]U.S. Census Bureau (Retired)
[3]U.S. Census Bureau & University of Maryland, College Park

Abstract
Statistical data editing means identifying potential response errors in the data. Without regard to how flagged observations will be modified, data editing is subject to two types of errors: labeling a correct observation as erroneous and not identifying an incorrect value. There is no statistical criterion to decide how many observations should be edited. Over-editing can increase data errors, degrade data quality, change the data structure and increase costs. Error localization consists of a separate test on each observation, where the null hypothesis states that the observation is error free and the alternative states that the observation is erroneous. The False Discovery Rate (FDR) is the fraction of false-positive findings among those deemed to be erroneous. Because FDR control is related to the number of edited observations, imposing an FDR requirement specifies the number of outliers to be edited, thereby controlling overediting. In this presentation, we experiment with a local FDR method on error localization and evaluate its performance and the performance of well-established editing methods on simulated data.

Key Words: data editing, response errors, outlier detection, multiple hypothesis tests, simulation

1. Introduction

Data editing is the process of detecting and correcting invalid or unlikely values in a reported data set due to respondents' *reporting errors*, or data processing *keying errors*. For example, a survey might ask for salary earned in 2019, but the respondent provides the salary earned in 2018. As another example, despite a survey asking for a dollar value rounded to the nearest thousand, an interviewer might enter an unrounded dollar value. Occasionally, the institutional data coding system may have human errors, for example, coding a missing item as 9,999,999,998 instead of 9,999,999,999 so that the incorrect code is interpreted as a genuine data value. We use the term *response errors* to generally classify all such errors. In a survey setting, usually data editing is performed after receiving raw responses and prior to *nonresponse adjustments* or *unit imputation*. Correcting response errors is important, because editing results significantly affect the later stages of imputation and estimation. Survey response errors degrade the survey data quality and may lead to incorrect inferences. Data editing methods should reduce non-sampling error, in turn reducing final estimation bias and increasing estimation precision.

Table 1 illustrates the effect editing can have on survey estimates, based on artificial data resembling real world occupational survey data. In the Table, there are estimates of the number of government employees and their associated coefficients of variation, both before and after editing. There are huge – by any definition – differences before and after editing between values of both the estimated totals and their coefficients of variation.

Table 1: Potential Data Editing Effect

	Data before Editing		Data after Editing	
Occupation	Est. Total	CV (%)	Est. Total	CV (%)
A	227,000,652,590	100.2	431,962	9.3
B	142,556	65.4	51,664	17.4
C	293,742	43.9	148,417	10.0
D	10,100,310,701	99.2	216,622	6.6

Source: Artificial data resembling real world occupational survey data.

The results in Table 1 look promising. In Table 1 the coefficients of variation after editing are much smaller and of more usual magnitudes. The reduced coefficients of variation indicate that massive response errors were removed, and more plausible values replaced them. For example, the original totals for Occupations A and D are wrong by orders of magnitude. Each of these may have been a data entry error in which two totals were erroneously concatenated into one entry.

Editing can be expensive and time consuming. The estimated cost of editing at national statistical institutions may be as high as 40 percent of the total survey budget (Granquist and Kovar, 1997; National Academies 2018, ch. 6). Editing approaches have evolved from *manual editing* (directly contacting respondents) to *automatic editing* (mathematical modeling) or *record linkage* (sharing information among agencies). See de Waal, Pannekoek & Scholtes (2011). Establishment surveys may contain millions of observations/records. Thus, the editing approach switches to *selective editing* (only large errors should be edited) (Norberg et al.,2010; Granquist and Kovar, 1997; Latouche and Berthelot, 1992; Hidiroglou and Berthelot, 1986). Therefore, efficiency and accuracy of editing procedures should be a matter of concern for statistical organizations.

1.1 Data Characteristics, Editing Process, and Editing Errors

Survey data, especially economic data, usually are nonnormal, skewed, and with long tails. Most observations are correct, but a small portion of response errors may occur. Frequently the underlying observations with response errors are small. Some categories of data are more difficult than others to collect; thus, within a survey the data quality varies among categories. Data characteristics also vary from year to year. For example, in years 2007, 2008, and 2009 economic data were substantially affected by the situation of the economy.

There are several approaches to editing. We focus on identifying response error in individual observations by using the same variable from consecutive years, i.e., prior period versus current period, as information to identify possible response errors. One version of this approach uses robust regression of current data on prior data, assuming the prior data is error-free. Extremely large positive or negative residuals are regarded as evidence of possible response error. Another editing method is to transform the data so that possible response errors stick out as outliers and can be identified (Hidiroglou & Berthelot 1986). A third method is to fit a normal distribution model to the transformed data and use outlier detection methods based on the normal distribution (Di Zio & Guarnera 2013).

There are two main steps in the editing process: *error localization*, which is the process of detecting potential errors, and *error value replacement* (item imputation), which is the process of replacing erroneous values with corrected values. Editing errors occur during both steps in the editing process. At the error localization step, editing errors occur when incorrectly flagging observations, i.e., *Type I errors* and *Type II errors*. At the error value replacement step, editing errors occur when imputing

values for flagged observations, i.e., *imputation bias*. In our effort to understand the performance of editing methods, we limit our study to the error localization step only.

1.2 False Discovery Rate (FDR) to Assess the Accuracy of Error Localization

Error localization in survey data editing is the problem of testing whether or not an observation in a data set contains a response error. Each of *N* observations is examined by some editing algorithm, and some observations are flagged as possibly erroneous. For each observation, the null hypothesis of no error is tested against the alternative of response error. This problem is analogous to genomics, in which two groups (say, healthy and diseased) are compared by comparing the activity of each of several thousand genes, separately gene by gene.

When *N* is very large, in the thousands or larger, the classical Bonferroni method of experiment-wise (or family-wise) error control makes it almost impossible to reject a false null hypothesis. Controlling the sample false discovery rate (FDR) is the preferred method in such large multiple testing situations.

For survey data, large observations are often more important than small observations and therefore larger response errors are more important than smaller response errors. For large values, committing a Type I error is more serious than committing a Type II error. The classical Bonferroni method for controlling family-wise Type I error is impractical for survey data editing situations. An alternative to Bonferroni adjustment is to control the false discovery rate (FDR), the expected proportion of type I errors among observations flagged as errors (Benjamini & Hochberg 1995; Efron 2010). The following table displays the definition of sample False Discovery Rate in a multiple hypothesis situation.

Table 2: False Discovery Rate to Access the Accuracy of Error Localization

		Predicted Class		
		Response Errors	Correct Values	
Actual Class	Response Errors	True Positive (TP)	False Negative (FN) (Type II error)	TP/ (TP + FN) = Sensitivity
	Correct Values	False Positive (FP) (Type I error)	True Negative (TN)	TN/ (FP + TN) = Specificity
		FP/ (TP + FP) = FDR		

1.3 Research Interests and Research Questions

The survey data editing problem is not well defined. In theory and in practice we do not know exactly what the true error-free survey data looks like and we don't know the number of response errors in the observed data and the nature of those errors. In the real data there are no "correct answers" to check an editing method's performance and no "test data" on which to conduct cross validation. We should make as few assumptions as possible about the data, since there is no way to verify any detailed assumptions about data generation or response error distributions, however plausible they seem. Experience may suggest vague, qualitative understanding of data quality, response errors and effectiveness of editing, but we can never see the truth. For these reasons we must rely on simulation as the only way to study the editing problem.

Many editing methods have been developed in the literature. We are interested in understanding these various methods' statistical performance, that is, their error rates particularly as quantified in terms of FDR. To our knowledge, the topic of editing error rates has not been widely addressed in the data editing literature, except when detailed mathematical models are assumed for data distributions and error genesis (Ghosh-Dastidar & Schafer (2006); Di Zio & Guanera (2013)). We

view data editing as hypothesis testing, one test for each observation. The null hypothesis is that the observation under test is free of error. Survey data with many observations leads to a multiple hypothesis testing problem.

We focus on quantitative periodic survey data and simulate various data scenarios, restricting attention to the error localization step. Among many editing approaches, we focus on selective editing methods, which aim to identify large data errors only. We are also interested in tuning parameters, which play an important role in the performance of selective editing methods.

Our research questions are:

1. Viewing error localization as a multiple hypothesis testing problem, what are the experimentwise type I error rates of the editing methods under varying data scenarios?
2. Does control of False Discovery Rate (FDR) provide a more useful performance criterion for editing methods than experimentwise type I and II error rates?
3. How do the tuning parameter values affect the performance of editing methods?
4. Experimental study of a new editing method, "Local FDR Control," proposed first in a genomic context (Efron, 2010)

2. The Challenge of Editing a Periodic Survey

2.1 Response Error Structure

The structure of a periodic survey with response error contains two sources of randomness: random fluctuation of growth rates and response error. Suppose that we observe a quantitative response Y and a predictor X. Let the true model (possibly after transformation) be written in general form as

$$Y = f(X) + \epsilon \qquad (1)$$

where X = data from the previous survey, known without error,
Y = true data from the current survey,
f = fixed, unknown function,
ϵ = random fluctuation of growth rate with zero mean.

With response error, the observed value Z is

$$Z = Y + \delta = f(X) + \epsilon + \delta \qquad (2)$$

where δ = response error. Similar, but fully parametric and mathematically rigorous, models for data with response error were proposed by Ghosh-Dastidar & Schafer (2006) and by Di Zio & Guarnera (2013).

The response error δ has a zero-inflated distribution. One does not know in advance which, if any, observations have response errors. The presence or absence of response errors must be modeled probabilistically. Most δ values are zero but $E(\delta) \neq 0$. So f is not identifiable. When δ is nonzero, we cannot separate δ from ϵ. Therefore, error localization is an under-identified problem. Possibly ϵ depends on X. Large magnitudes of ϵ increase the difficulty of identifying δ.

The description of the model in equations (1) and (2) is satisfied in our simulation model. It captures several of the features underlying the nonparametric formulation of the contaminated-data problem that is presented more fully in Appendix A and in Section 4.1, including mixtures of components with very different size and tail-behavior. However, the model is incomplete because the joint distribution of ϵ, δ and Y is not fully specified. Formulating such a model requires many assumptions about conditional probabilities, and such assumptions cannot be verified from observed data. At this point, we merely point out how complicated a general model may be. A specific instance of a fully specified model along these lines, used in simulations, is given in Section 4.1.

2.2 Survey Distribution and Correlation

Data contaminated by response error has some unknown mixture distribution that is unlike any standard distribution. Many real-world survey datasets should be regarded as mixtures without adequate information about the mixture components. Non-normality, even after transformation, is a common phenomenon (Weng, 2015).

Let X, Y, and Z be the variables described above from two successive runs of a periodic survey. The magnitudes of δ and the numbers of nonzero δ form various data patterns and determine data set quality. The ε and δ are unobservable, but correlations provide a partial picture of ε and δ in a data set. We should expect corr(X, Z) to be smaller than corr(X, Y), because δ dilutes the (X, Y) association. As the magnitude of ε increases, corr(X, Y) and corr(X, Z) both become weaker. A Larger $|\varepsilon|$ increases the difficulty of identifying δ. In the survey world, one can observe corr(X, Z), but not corr(X, Y). When editing data, corr(X, Z) provides only partial qualitative information about the presence and magnitude of response error.

2.3 Contaminated Growth Rate

The *growth rate*, Y_i/X_i, would be a way to visualize ε, but Y_i is unobservable. For survey data, one can only observe the *contaminated growth rate*, Z_i/X_i, in which ε and δ are lumped together. It is not unusual to observe that in economic surveys growth rates are more variable among small units (that is, units with small X_i) than among large units.

3. Editing Methods

The performance of error localizing methods depends on the design of the method and the number of observations to be edited. Their performance might also vary according to the error data structure. In Simulation I below, we focus on two methods: Hidiroglou-Berthelot (HB) (Hidiroglou and Berthelot, 1986) and Robust Regression (RR) (Huber, 1973). Neither of the two methods requires assumptions about the underlying distribution of the data. However, RR relies on linear regression. Simulation II compares also the new method of Local False Discovery Rate (Local FDR) (Efron, 2010).

3.1 Hidiroglou-Berthelot (HB) Method

Survey data may be longitudinal, that is, may include period-to-period data for the same units. *Differences* in reported values between time periods are then an important factor in determining the likelihood of an error. In survey data, units may not all have the same *importance*, which is another factor in the error localization process. For example, units with larger values may contribute more to estimates of totals and variance, and therefore, should receive more consideration. Similar reasoning applies to estimation of small units with special characteristics — those units should be checked first, since their observations contain critical, irreplaceable information. Hidiroglou and Berthelot (1986) first converted the concepts of difference and importance into formulas for survey editing.

The Hidiroglou-Berthelot (HB) method is a selective editing method that uses current and historical information to identify the most influential suspect values. For two occasions t and $t+1$, the data are $\big(x_i(t), x_i(t+1)\big),\ i = 1, \cdots, n.$ Let the estimated overall growth rate be

$$\tilde{\tilde{R}} = \frac{\sum_{i=1}^{n} x_i(t+1)}{\sum_{i=1}^{n} x_i(t)} = \sum\nolimits_{i=1}^{n} I_i r_i, \qquad (3)$$

where $I_i = I_i(t) = x_i(t) / \sum_{i=1}^{n} x_i(t)$, is a measure of influence of unit i, and
$r_i = x_i(t+1)/x_i(t)$, r_i in [0, ∞), is a measure of growth for unit i.

The pairs (r_i, I_i), i=1,…n, are a 1 to 1 transformation of the original data. It is implicitly assumed that all $x_i(t)$ are positive. In practice, observations with zeros would be removed from the data for special treatment and not submitted for HB editing.

In order to capture the outliers among the r_i, the HB method transforms the distribution of r_i to a less skewed or more bell-shaped one as follows:

$$s_i = \begin{cases} 1 - \dfrac{\tilde{R}}{r_i} = \dfrac{r_i - \tilde{R}}{r_i} < 0 & \text{if } r_i < \tilde{R} \\ \dfrac{r_i}{\tilde{R}} - 1 = \dfrac{r_i - \tilde{R}}{\tilde{R}} \geq 0 & \text{if } r_i \geq \tilde{R} \end{cases} \tag{4}$$

where $\tilde{R}$ is the sample median of the ratios r_i, and s_i ranges over $(-\infty, +\infty)$. Note that when $r_i < \tilde{R}$, s_i is a nonlinear transformation and when $r_i \geq \tilde{R}$, s_i is a linear transformation. In (2), the small r_i (near 0) are transformed into large negative s_i and the large r_i (near ∞) are transformed into large positive s_i.

To get back to the original scale and to control the size of x_i, which can range over several orders of magnitude, another data transformation is needed:

$$E_i = s_i\{\max\big(x_i(t), x_i(t+1)\big)\}^U, \tag{5}$$

where $0 \leq U \leq 1$, and E_i is called the "effect."

Finally, an outlier is defined as a value lying outside the interval $\tilde{E} - Cd_{Q1}, \tilde{E} + Cd_{Q3}$, where d_Q is a quartile of the sample effect distribution and $\tilde{E}$ is the median effect. The tuning parameter $C > 0$ is chosen to give a desirable length of the acceptance region. The exponent U is another tuning parameter, frequently set to 1/2. In what follows, we always set $U = 1/2$.

The HB method identifies outliers by combining two quantities: the difference criterion (r_i, current value divided by previous value) and the importance criterion (I_i, observation on unit i divided by total observations on all units). The difference criterion, r_i, explicitly appears in the computing formula (2) and through s_i in (3). The importance criterion I_i implicitly appears in the computing formula (3) in U, through the proportionality $I_i = x_i(t)/\sum_{i=1}^{n} x_i(t)$ within E_i , so that E_i is proportional to a combination of "difference" and "importance" quantities:

$$E_i \propto s_i \max\{I_i(t-1), I_i(t)\}^U$$

The parameter U determines the tradeoff between s_i and I_i in the definition of outliers. Changing the value of U will change the distribution of E_i. The larger the value of U, the larger the effect for bigger x_i. When U = 0, the whole computation only depends on the period-to-period growth ratio. When U = 1, the calculations are conducted on the original scale of $\max\big(x_i(t), x_i(t+1)\big)$ as well as s_i. Note that when $U < 1$, the importance criterion of a larger observation is downweighted. Since s_i and $\max\big(x_i(t), x_i(t+1)\big)$ have different scales and distributions, it is difficult to determine the ideal value of U. Moreover, there is no optimality criterion to determine U. To determine the value of U, users experiment with several values to obtain "reasonable" results.

Unlike some editing methods, the HB method does not require distribution theory for its computation. However, the HB method attempts to symmetrize the distribution of E_i so that the interval $[E_M - C\big(d_{Q1}\big), E_M + C\big(d_{Q3}\big)]$ can serve as an acceptance region, that is, a sort of hypothesis test. Thus, U serves as the parameter for adjusting the heavy-tailed distribution of E_i. When U is small, the kurtosis of E_i is lower and vice versa. This means the choice of C also depends on U. The HB method uses C to determine the cutoff point for outliers, but C is chosen arbitrarily, so that the number of outliers may be large and unpredictable.

3.2 Robust Regression (RR) Method

Survey data typically result from complex sampling designs (e.g., stratified, clustered), so that their distributions are not normal. However, regression may be regarded as just a collection of algorithms to fit a linear function to a data set. Its validity is based on a structural assumption, but not on a distributional model. For example, for point estimation alone, ordinary least squares (OLS) need not assume a normal error distribution. Here, the goal of estimation is error identification rather than inference on the regression parameters.

Survey data editing resembles methods that regression analysts use to identify outliers or influential points in an experimental dataset. They have developed methods such as box plots, Studentized residuals, and Cook's distance to detect unusual data points. The robust regression (RR) M-estimate (Huber 1973) has been widely used for outlier detection in regression problems (Rousseeuw and Leroy, 1987**).** There were studies performing editing and imputation simultaneously, in which robust estimation was involved (Little and Smith, 1987; Thompson, 2007). Di Zio, et al. (2007) studied the capability of outlier detection methods, including RR, in an actual business survey, after transforming the data to the log scale. Instead of ordinary least squares regression, we employ the RR M-estimate as an editing tool. RR can be thought of as a reweighted least squares regression which downweights extremes in the sample and is therefore less affected by contaminated data. No distributional assumptions are required by robust regression if the goal is estimation. However, if an inference or confidence interval is required, some distributional assumptions are needed.

Let $X = [x_{ij}]$ denote an $n \times p$ matrix, let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ be a given n-vector of responses, and let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ be an unknown p-vector of parameters to be estimated. Instead of minimizing the sum of squares as in least squares regression, M-estimation defines a weight function such that the estimating equation becomes

$$\sum_{i=1}^{n} w_i (y_i - X_i{}'\beta) X_i' = 0, \qquad (6)$$

where $w_i = \min\{1, k/|y_i - X'_i\beta|\}$ is the weight given to the *i*-th observation and *k* is a tuning parameter. The parameter *k* is chosen so that if the data are normally distributed, the efficiency of the Huber estimator, compared to the ordinary least squares (OLS) estimator, is $Var(\hat{\beta}_{OLS})$ / $Var(\hat{B}_{RR}) \approx .95$. Typically, *k* is about 1.5 to 2.0. The weights depend on the residuals, and the residuals on the weights. Therefore, equation (4) is solved using **I**teratively **R**eweighted **L**east **S**quares (IRLS).

When the $\boldsymbol{y}$ response vector contains outliers, robust regression downweights these points while fitting a regression line. We want to identify data points that are far from the "true" line, that is, the line estimated without the effect of errors. Robust regression provides such a line, and then we infer that extreme residuals from the robust regression line are possible errors. In the RR editing method, one computes the residuals:

$$d_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i), \qquad (7)$$

where $\hat{b}_{\mathbf{0}}$ and $\hat{b}_1$ are the robust estimates of the regression coefficients. By flagging the $100\alpha\%$ largest absolute residuals, say 1 percent of the data file, the number of edits is well controlled. In our application, we perform robust regression on the original scale.

3.3 Local FDR Method

As before, we have observations z_i, i = 1, ..., N, which are current data, some of which may be erroneous We don't know which observations contain error or the magnitude of error if present.

Controlling local FDR (Efron 2010) begins with estimating the unobserved *Y* based on regression of Z on *X*. The regression method employed is Least Trimmed Squares (LTS), a resistant robust regression algorithm which attempts to remove the effect of outliers from the fitted regression line.

Next, the densities of the predicted *Y* values (under LTS) and *Z* are estimated via smoothed histograms, denoted respectively as $f_0(u)$ and $f(u)$. The histograms have many narrow bins containing

only a small number of observations. For each z_i, the ratio $f_0(z_i)/f(z_i)$ is calculated. Small values of this ratio suggest that z_i is improbable under the i-th null hypothesis and that z_i includes response error. One flags all observations satisfying $f_0(z_i)/f(z_i) < q$, where $0 < q < 1$. The quantity q is a tuning parameter selected by the user.

4. Simulation I: FDR, Editing Methods, and Data Scenarios

The performance of error localizing methods depends on the design of the method and the number of observations we intend to edit. Their performance might also vary according to the error data structure.

4.1 Generating the simulated data

The simulated data mimics real world data, such as ASPEP (Annual Survey of Public Employment & Payroll, U.S. Census Bureau), which has similar characteristics to other continuous, longitudinal economic data. The data are non-normal and highly skewed so that a small number of large units make major contributions to estimates of totals. The growth rates may vary substantially, depending on the size of the unit. Large differences may exist between raw and edited data.

Let X, Y, and Z be the same variable from a periodic survey. Using R software, we generated samples of size 1,000 consisting of independent triplets (X_i, Y_i, Z_i), $i = 1,\ldots,1000$.

We generated the X data as a sample of 1000 independent observations x_i, i=1, ... , 1000, from a mixed lognormal distribution with two components. Recall that if X is lognormal, then $U = \ln(X)$ has a normal distribution with mean μ and standard deviation σ. The median of $X = \exp(\mu)$. The parameters are called the log-mean and log-standard deviation. Twenty percent of the X data (the "large" component) had log-mean $\mu = 0.2$ and log-standard deviation $\sigma = 1.25$. These X values tend to be large and have a highly skewed distribution. The other 80% of the X data (the "small" component) have log-mean $\mu = -0.75$ and log-standard deviation $\sigma = 1.1$. Their distribution is concentrated near zero but still highly skewed. The parameters of X were chosen empirically to make the X distribution resemble that of a real world survey variable.

To generate Y, we created 1000 lognormal growth rates $r_{i=} y_i/x_i$, i=1, ... , 1000. Both components had log-mean = 0, or componentwise median growth rate of 1. For each scenario in the next section, the large component had a small log-standard deviation and therefore fairly stable growth rates. (See Table B1 of the Appendix.) The small component had a larger log-standard deviation and therefore variable growth rates. Finally, we calculated $y_i = r_i x_i$ for each i.

Multiplicative response errors m_i were applied randomly to about 10% of the y_i to create $z_i = m_i y_i$. The remaining z_i were identical to y_i. The conditional probability that $z_i \neq y_i$ is

$$P[z_i \neq y_i \mid x_i, y_i] = 0.001 + 0.10/(1+\exp(-5+0.008*(x_i - 380))).$$

These parameters were chosen to ensure that errors were most likely to occur when x_i is small. The magnitude of an error m_i, given that an error occurs, is randomly selected from a lognormal distribution, independent of the true (x_i, y_i). The parameters of this lognormal distribution varied from one scenario to another and did not depend on whether x_i came from the large or small component of the X distribution.

The distributions of X and the of the response error probabilities are fixed throughout the simulation. However, by altering the parameters of the growth rates and error magnitudes, we can obtain a range of values of corr(X,Y) and corr(X,Z). Our simulation always is based on the same distribution structure, but we can simulate a broad range of variability in growth rates and error magnitudes. This enables us to examine editing methods over a wide range of data scenarios.

The two-component model described above is a detailed specification of the general nonparametric model described in Section 2.1. We regard this model as more realistic than a single component model, based on our experience with real data sets. It can serve as a useful test bed for new

algorithms as well as a model which might be analyzed mathematically in studies of existing editing procedures.

4.2 Generating Data Scenarios

From our experience, survey data quality varies not only between surveys, but also within a survey. Within a survey, some category of data has good quality and can have corr(X,Z) around 0.90 or higher, but some can have corr(X,Z) only around 0.70. With the same periodic survey, some periods of data can have corr(X,Z) around .90, but other periods can have corr(X,Z) could be around .60. It is the magnitude of error size and the numbers of errors that drive corr(X,Z) up and down.

Varying the data scenarios in the simulation reflects the variety of data features in real world periodic surveys. We identify six scenarios of interest, depending on the variability of growth rates and the magnitudes of response errors. Large corr(X,Y) corresponds to stable growth rates across the population, while small corr(X,Y) reflects highly variable growth rates. Large corr(X,Z) reflects small response errors and stable growth, while small corr(X,Z) indicates some combination of variable growth rates and large response errors.

As described in Section 3.3, by adjusting the parameters of the growth rates, r_i, and the magnitudes of the response errors, m_i, we can achieve a variety of desired combinations of corr(X,Y) and corr(X,Z).

We examined six data scenarios: corr(*X*, Z) was set to three levels (high, medium, and low), and corr(*X*,*Y*) was set to two levels (high, low). Table 3 shows the six data scenarios being studied. The lognormal parameters defining the scenarios are tabulated in Table B2 of the Appendix.

Table 3: Data Scenarios

	corr(*X*,*Y*) ≈.**95**	corr(*X*,*Y*) ≈.**81**
corr(*X*,***Z***)	.90	.73
corr(*X*,***Z***)	.73	.50
corr(*X*,***Z***)	.50	.28

4.3 Results and Discussion

Table 4 shows the results from Simulation I. This simulation investigates the interactions between data scenarios and editing methods. Note that in these simulations, RR has been constrained to flag the same number of observations as the HB method. The sample FDR is obtained by:

$$FDR = \frac{\text{False Positives}}{\text{False Positives} + \text{True Positives}} = \frac{\text{Type I Errors}}{\text{Number of Errors Flagged}}$$

FDRs differ among data scenarios: For the HB method, 10% < FDR < 61%. For the RR method 30% < FDR < 81%. The HB method is superior to RR method in every scenario. The lowest FDR occurred in the easiest scenarios, numbers 2 and 3, in which corr(*X*,*Y*) = 0.95 and corr(*X*,*Z*) was = 0.50 or 0.73. The highest FDR occurred in Scenario 4 (the most difficult), with corr(*X*,*Y*) = 0.81, and corr(*X*, *Z*) = 0.73.

Obviously, the design of an editing method plays a role in the editing accuracy. The RR method uses the original *X* and *Z* values and takes the largest robust regression residuals as possible response errors. On the other hand, the HB method multiplies the transformed growth rate and transformed max(*X*, *Z*) values to create "effect" values. Finally, HB uses the extreme values of the effects to flag possible response errors. By transforming the data, HB increases the effect's size, hence increasing its statistical power, especially when *X* is small.

Data scenarios also affect the editing accuracy. Scenario 3 is the easiest case to edit. Its corr(X,Y) = .95 corresponds to smaller ϵ, and corr(X,Z) = .50 to larger $\epsilon + \delta$. In the latter scenario, it is easier to identify response errors. In Scenario 4, corr(X,Y) = .81 implies that ϵ is larger, and corr(X,Z) = .73 implies the relative magnitude of δ is small, in which case it is difficult to identify errors.

From our experience, data generated under Scenario 4 more closely resembles real periodic survey data than in the other Scenarios. Note that, when corr(X,Y) = .81, all the FDRs are very high, from 36% to 81%,. These two editing methods do not perform well in this situation.

Table 4: Simulation I Results – Data Scenarios vs. Editing Methods

Data Scenario: (corr(X,Y); corr(X,Z)	Number of Errors Flagged	HB		RR	
		Type I Errors	FDR	Type I Errors	FDR
1 (0.95; 0.90)	19	6	32%	12	63%
2 (0.95; 0.73)	33	4	12%	17	30%
3 (0.95; 0.50)	42	4	10%	21	50%
4 (0.81; 0.73)	31	19	61%	25	81%
5 (0.81; 0.50)	43	18	42%	29	67%
6 (0.81; 0.28)	50	18	36%	32	64%

5. Simulation II: False Discovery Rate, True Positive, False Positive, and Turning Parameters

The Local FDR method (Efron 2010), introduced in Section 3.3, is newly applied here to the context of statistical editing. Artificial data were generated by selectively contaminating 2,185 publicly traded NYSE (New York Stock Exchange) yearly closing stock prices. In this section two very different data sets were chosen: year 2010-2011 with corr(X, Y) = 0.95 and year 2006-2007 with corr(X, Y) = 0.85. We want to get a sense of what the data look like. We also want to understand how the same editing methods perform under different scenarios.

5.1 Generating the simulated data

The goal of this simulation is to examine the performance of editing methods on a real world data set. This is only feasible if we know the true values of X and Y. Typical periodic surveys only publish the edited values of the data, which may contain unidentified response or imputation errors. Edits may be based on earlier survey results, but those results may also contain unidentified response errors.

One data source which is not affected by response errors is stock market data. The stock prices in financial publications are exact, so we can regard X and Y as known. Of course, without response error, Z is identical to Y, but we can generate errors and create a simulated version of Z.

It turns out that the distributions of annual closing prices of New York Stock Exchange (NYSE) securities are similar to those of economic variables measured in government and private surveys, based on examination of scattergrams and correlation coefficients. Our simulation strategy is to select several two-year periods of NYSE closing prices and generate errors to create simulated (X,Y,Z) data. We edit the Z data for year t using the error-free X data of year (t - 1) and compute various error criteria.

We applied this simulation technique to the 2006-2007 closing prices and to the 2010-2011 closing prices. We looked at the 2006-2007 NYSE closing prices because in 2007 the recession had a major effect on NYSE stock prices. In contrast, the stock market was more stable in 2010-2011. Our simulated errors were more easily detectable in the latter years.

In each simulation we assigned errors at random, with P[error] = 0.1. The error probabilities were assigned independently of the x_i. The errors were multiplicative and followed a scaled lognormal distribution with log-mean=0 and log-standard deviations of 1.0 for the 2010-2011 case and 0.5 for the 2006-2007 case. Using different log-standard deviations was necessary because stock prices were much lower in 2007 than in 2011.

5.2 Simulation II: Results & Discussion

5.2.1 False Discovery Rate

Table 5 displays results for a single iteration (Z data is simulated) with corr(X,Y) = .95 and corr(X,Z) = 0.60 for Years 2010-2011. The sample size is 2185, and a total of 222 response errors were generated. Three methods, HB (C = 20, U = 0.5), RR (a = 0.04), and Local FDR, (q = 0.4), flagged 77, 88, and 87 errors, respectively. The FDR for HB was 7%, for RR it was14%, and for Local FDR it was 62%. The HB is the best.

In the same way the Year 2006-2007 data are created with corr(X,Y) = 0.85 and corr(X,Z) = 0.78. The sample size is 2185 and a total of 218 response errors were generated. Three methods, HB (C = 20, U = 0.5), RR (a = 0.02), and Local FDR (q = 0.45), flagged 36, 37, and 40 errors, respectively. The FDR for HB was 31%, for RR was 41%, and for Local FDR was 53%. The HB is the best.

The tuning parameters determine the numbers of observations to flag. Increasing the number of flagged observations increases the risk of Type I errors. HB has two tuning parameters, C and U; however, we do not know how to choose the parameter values to fit a particular data set. We pick parameter values (C = 20, U = 0.5) to fit both Years 2010-2011 and Years 2006-2007. HB flagged 77 observations for Year 2010-2011 but flagged 36 observations for Years 2006-2007. This feature of automatically adjusting the number of flagged observations makes HB particularly useful. The RR method has a weight function with tuning parameter α. It is not clear how to determine the tuning parameter values to control these flag numbers. The number of observations flagged by HB determined the tuning parameter RR. The same method was used to determine (approximately) the tuning parameter q used for local FDR control.

There are some possible explanations for the poor performance of Local FDR as an editing method. First, regressing Z on X to estimate Y introduces estimation error. (This estimation error would be worse if X were measured with error in a real world situation.) Second, comparing smoothed histograms of $\hat{Y}$ and Z is inefficient, because histograms are inefficient density estimators. Third, some of our editing scenarios are inherently difficult to edit. Local FDR seems oversensitive to data quality.

Table 5: Years 2010-2011 and Years 2006-2007: False Discovery Rates

	Years 2010-2011			Years 2006-2007		
Methods	HB	RR	Local FDR	HB	RR	Local FDR
Total number of Errors	222	222	222	218	218	218
Number of errors flagged	77	88	87	36	37	40
False Positive	6	12	54	11	15	21
True Positives	71	76	33	25	22	19
FDR	7%	14%	62%	31%	41%	53%

5.2.2 True Positives

We take a closer look in Figure 1 and Figure 2 at how well the editing methods localized the true errors. Both figures are presented in the same manner: gray dots are error free data, black dots are erroneous data, and red dots are erroneous data that have been edited and yield true positive results.

In Figure 1, upper left panel, 2,185 publicly traded NYSE (New York Stock Exchange) year 2010-2011 closing stock prices were plotted; 222 observations had injected errors. Black dots are erroneous data. Errors mainly occur for small stock prices and the error point (378, 3420) has been excluded from the graphs in order to maintain a readable scale. Gray dots are error-free data. Note the clear linear trend for error-free data.

In Figure 1, upper right panel, HB edit results are presented. Red dots are correctly flagged errors, i.e., true positives. Black dots are unflagged errors. Lower left panel, RR edit results are presented. Lower right panel, Local FDR edit results are presented. HB, RR, and Local FDR identified all extreme outliers, that is, all errors larger than 500. HB and RR behave similarly and flag many smaller errors and inliers. Local FDR identifies some smaller errors in an inconsistent fashion.

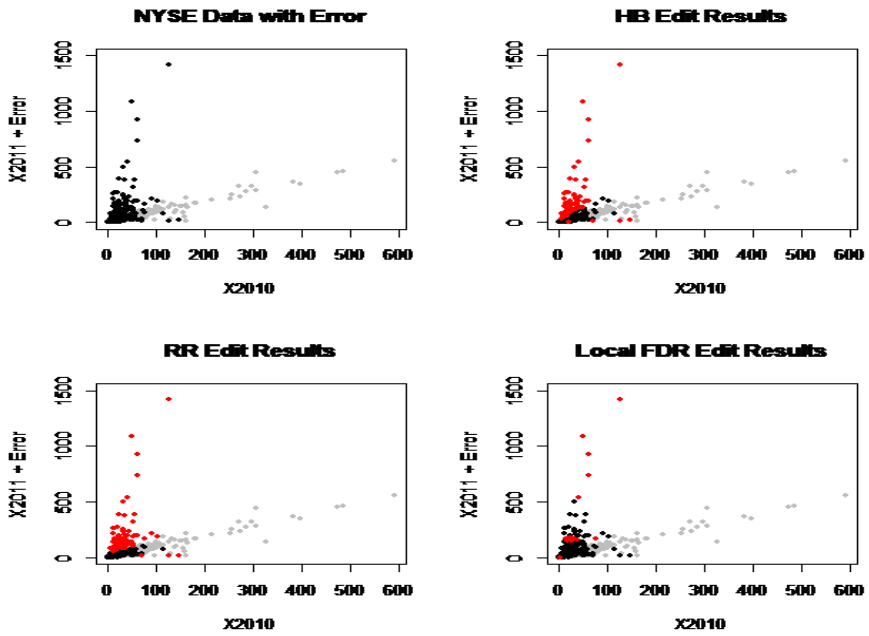

Figure 1: Year 2010-2011 Localizing True Errors

In Figure 2, upper left figure, 2,185 publicly traded NYSE (New York Stock Exchange) year 2006-2007 closing stock prices were plotted; 218 observations had injected errors. Black dots are erroneous data. Errors mainly occur for small stock prices. Gray dots are error-free data. Unlike Figure 1, there is no clear linear trend for error-free data.

Upper right figure: HB edit results are presented in this figure. Red dots are true positives. Black dots are unflagged errors. Lower left figure: RR edit results are presented. Lower right figure: Local FDR edit results are presented. HB and RR identified most extreme outliers, but HB identified additional small outliers in terms of year 2006 or year 2007. Local FDR identifies all erroneous values in the range in year 2007, around 100 to 300.

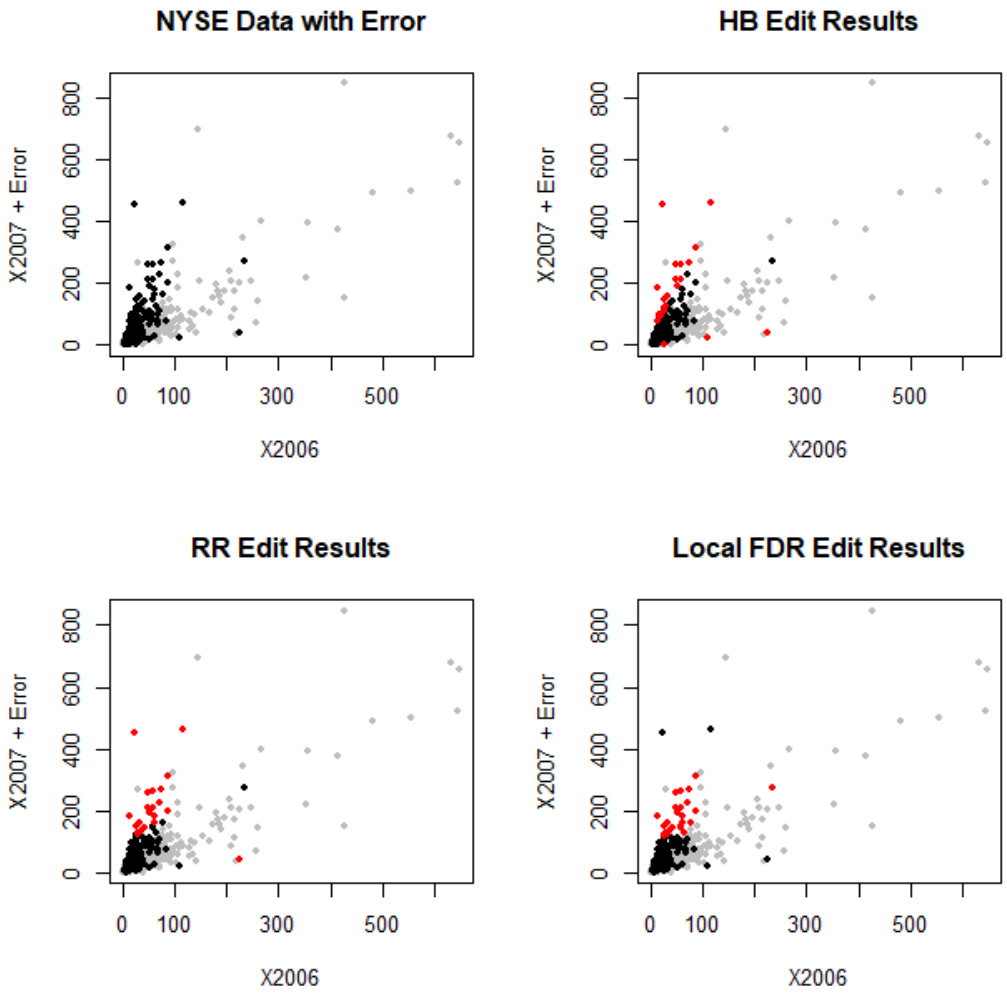

Figure 2: Year2006-2007 Localizing True Errors

5.2.3 False Positives

Figure 3 and Figure 4 display false-positive errors. Both figures are presented in the same manner: gray dots are error free data, black dots are erroneous data, and red dots are false positive observations.

In Figure 3, the upper left figure displays 2,185 yearly closing stock prices with 222 observations containing injected errors. Black dots are erroneous data. Errors mainly occur for small stock prices and the error point (378,3420) does not appear on the graphs. Gray dots are error-free data.

Upper right figure: HB edit results are plotted in this figure. Lower left figure: RR edit results are plotted in this figure. Lower right figure: Local FDR edit results are plotted in this figure. HB, RR, and Local FDR all committed Type I errors. But the three methods flagged different observations. Some Type I errors have large magnitudes. HB and RR flagged points that seem far from the general linear trend. Local FDR does not seem to follow an interpretable pattern.

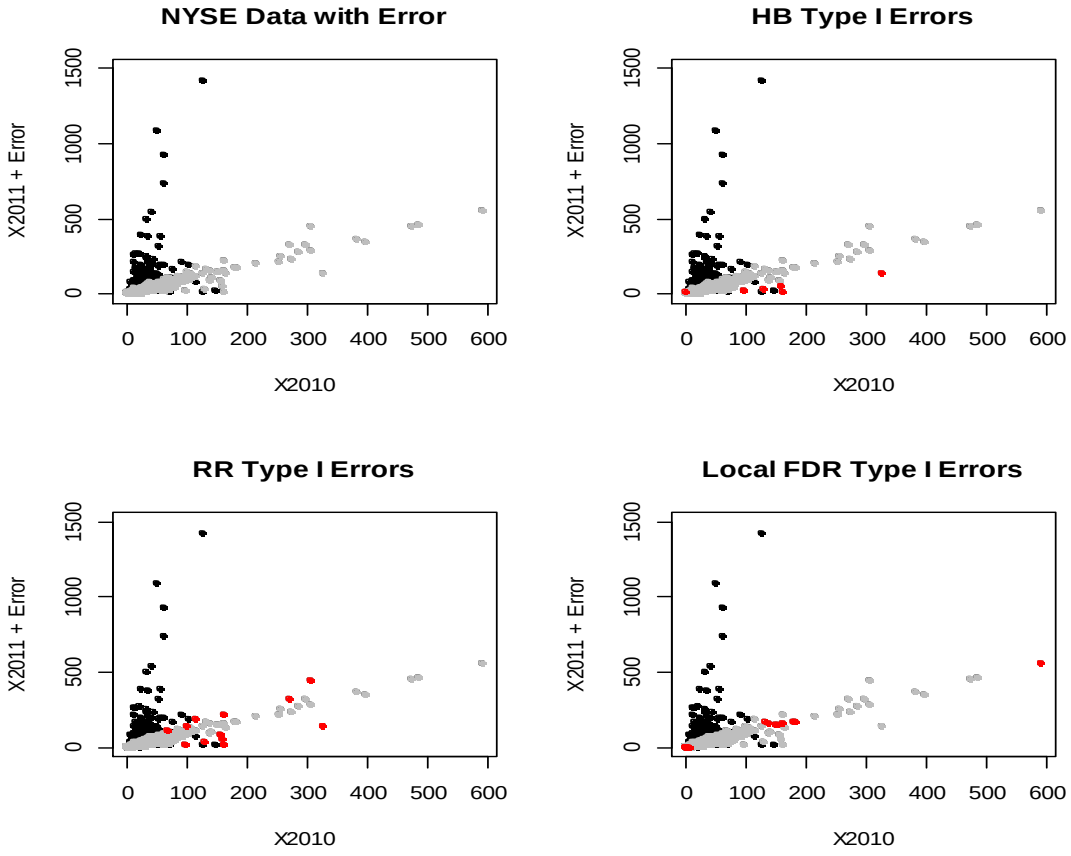

Figure 3: Year2010-2011 Location of False positive

In Figure 4, the upper left figure: 2,185 yearly closing stock prices with 218 observations containing injected errors. Black dots are erroneous data. Errors mainly occur for small stock prices. Gray dots are error-free data.

Upper right figure: HB edit results are presented. Lower left figure: RR edit results are presented. Lower right figure: Local FDR edit results are presented. HB, RR, and Local FDR all committed Type I errors. But the three methods flagged different observations. For HB the false positive points are located along the lower edge of the general linear trend. Should one prevent HB from

overediting? RR flags points where the values for both years 2006 and 2007 are large. Does it mean RR may be useful in the high corr(X,Y) situation? Local FDR, again, focuses on the points with year 2007 values between 100 to 300.

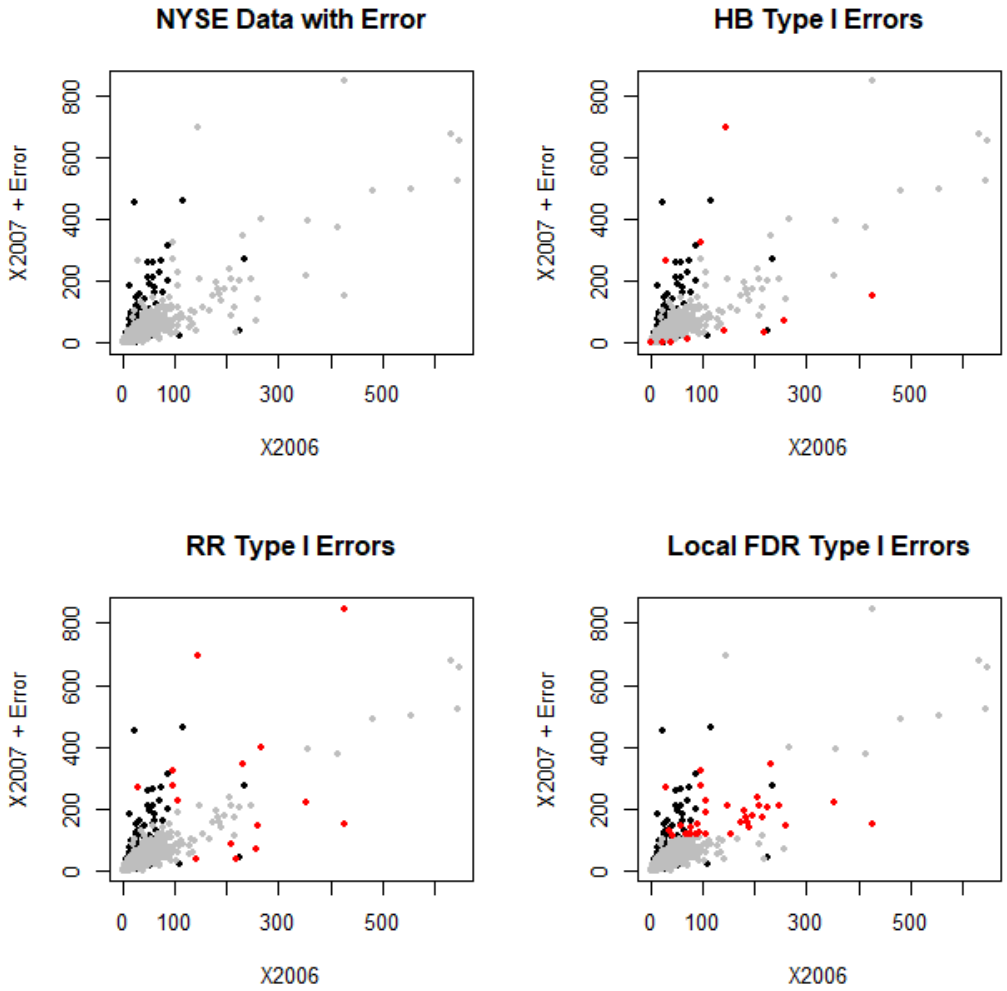

Figure 4: Year2006-2007 Location of False positive

6. Discussion and Future Research

6.1 Summary and Discussion

We have created an innovative model for simulating editing of a periodic survey. In this model, we create subsets of "Large" and "Small" units whose values follow different lognormal distributions. The year-to year growth rates also follow different distributions, capturing the fact that smaller units are more likely to be diverse and more variable in their growth. In our model, response errors occur at random but are more likely to occur in smaller units, reflecting experience with actual editing problems. This phenomenon is reflected in our simulations by defining the probability that any specific unit may have response error as a decreasing function of its size.

Our model is flexible and reasonably faithful to the characteristics of real-world editing data. We regard this model as a useful test bed for future studies of statistical editing.

Our simulations compare the performance of the Hidiroglou-Berthelot (1986) and robust regression-based (RR) editing methods over a variety of data scenarios. These editing schemes assume the availability of a previous error-free data set $X = [x_1, \ldots, x_n]$ to help in editing an observed current data set $Z = [z_1, \ldots, z_n]$, where a small proportion of observed z_i contain response error. The scenarios are described in terms of correlation coefficients and together describe a variety of levels of data quality.

For each scenario, we generated 10,000 replicated size $n = 1000$ samples of (X,Y,Z) and summarized the numbers of false detections (Type I errors), overlooked response errors (Type II errors) and false discovery rates (the ratio of erroneous error detections to the total number of detections, see below).

In each of these scenarios, we found that the HB method was superior to RR, whether judged by numbers of Type I or Type II errors or by false discovery rates. The performance of both editing methods varied with the scenario. Editing errors were least frequent when corr(X,Y) was high and corr(X,Z) was low and most frequent when corr(X,Y) was low and corr(X,Z) was high.

As mentioned above, we summarized false discovery rates (FDR) for each simulation. The FDR is an alternative error criterion which reflects the balance between correct and incorrect "detections." We also investigated local FDR rate, which can be interpreted as the estimated probability that z_i is error-free, given that z_i was flagged as an erroneous observation. In our simulations local FDR was difficult to control, at least partly because we approximate the unknown proportion of error-free observations by 1. This is a conservative approximation, but perhaps too conservative.

6.2 Future Research

The performance of any editing method depends on certain tuning parameters. For example, the HB method employs parameters C, which controls the number of flagged observations, and U, which rescales the original data via a power transformation. How does the choice of these parameters affect editing errors? What other quantities, aside from error rates, should characterize effective editing? Selective editing (De Waal, Pannekoek & Scholtus, 2011) relies on a *score function*, which measures the change of an estimator (such as an estimated total) as a given observation is deleted or replaced by an imputation. Control of a score function also depends on the tuning parameters of an editing method. We plan to investigate how error rates depend on tuning parameters using simulation.

In our simulations we saw that the HB method consistently outperformed robust regression editing schemes after adjusting for the number of observations to be flagged as possibly erroneous. We plan to study the HB and RR methods analytically, assuming our modeling assumptions hold. We conjecture that HB extracts more information from the data than RR, but this needs proof.

It may also be true that robust regression would perform better if one first transformed the data, possibly by a log transformation. We will investigate this question and see whether RR might perform comparably to HB after transforming the data.

A rigorous mathematical analysis of our model may also indicate potential bounds on the statistical efficiency of editing data that satisfy the model. Such optimality results would apply to real world data which roughly satisfies our modeling assumptions and would thereby provide useful practical guidance to data analysts.

Our computation of local FDR depends on approximations of density ratios, but better computational algorithms can be found in the literature. In particular, since histograms are inefficient estimators of densities, we will examine the effect of other density estimators that are

more reliable in the tails of the distribution, at least in some important cases. After all, one expects that extreme values of Z are evidence of response error.

References

Benjamini, Y. & Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *J. Royal Statist. Soc. Ser. B*, **57** 289-300.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley and Sons.

Di Zio, M., Guarnera, U., Luzi, O., and Tommasi, I. (2007): "Detection of potentially influential errors in statistical survey data." Presented at *Intermediate Conference of Italian Statistical Society*. Specialized Session: Venice.

Di Zio, M., Guarnera, U. (2013), A Contamination Model for Selective Editing. *Journal of Official Statistics 29*, pp.539-555.

Efron, B. (2010), *Large Scale Inference*. Cambridge University Press.

Ghosh-Dastidar, B., and J. L. Schafer (2006). "Outlier Detection and Editing Procedures for Continuous Multivariate Data." *Journal of Official Statistics 22*, pp.487 – 506.

Granquist and Kovar, 1997; "Editing of Survey Data: How Much Is Enough?" In: *Survey Measurement and process Quality*, L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin, eds. John Wiley & Sons, New York, pp. 415-435.

Hidiroglou, M.A. and Berthelot, J.M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys." *Survey Methodology*, 12, 73-83.

Huber, P.J. 1973. "Robust regression: Asymptotics, conjectures and Monte Carlo." *Ann. Stat.*, **1**, 799-821.

Latouche, M., and J.M. Berthelot (1992), "Use of a Score Function to Prioritize and Limit Reconnects in Editing Business Surveys." *Journal of Official Statistics 8*, pp. 389-400.

Little, R. J. and Smith J. (1987), Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association 82*, pp. 58-68.

National Academies of Sciences, Engineering, and Medicine. (2018). *Reengineering the Census Bureau's Annual Economic Surveys.* Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25098.

Norberg, A. et al. (2010), *A General Methodology for Selective Data Editing*. Statistics Sweden.

Rousseeuw P. J., Leroy A. M. (1987) *Robust Regression and Outlier Detection*, Wiley, New York.

Thompson, K.J. (2007). "Investigation of Macro Editing Techniques fpr Outlier Detection in Survey Data." Paper presented at ICES-III.

Weng, C.F. (2015), "Bayesian Multiple Imputation of Zero Inflated Count Data". Joint Statistical Meetings, Survey Research Methods Section. In JSM Proceedings, Survey Research Section. Alexandria, VA: American Statistical Association.

Appendix

Appendix A. Mathematically Rigorous Specification of Model in Section 2.1

A full specification of the model

$$Y = f(X) + \epsilon \quad \text{(A1)}$$

$$Z = Y + \delta \quad \text{(A2)}$$

requires a specification of the joint distribution of (X, ϵ, δ), the formula for f, and a probabilistic model for each of these terms. In addition, the probability model for which components of Z are erroneous must be specified. Often a preliminary transform of the data (such as a log transform)

may be needed to attain this form. In what follows, we assume that (X, Y, Z) are the outcome of such a transformation.

In the editing problem, $X = [x_i,..., x_n]$ is usually thought to be known without error. The components of x_i may be multidimensional (as in Ghosh-Dastidar & Schafer, 2006). The x_i are conditionally independent, given some covariates c_i which may not have been observed.

The ϵ_i are conditionally independent random variables, given x_i. Hence Y is also a vector of conditionally independent random variables, given (x_i, ϵ_i) . The distribution of ϵ_i may depend on x_i.

The occurrence of response errors in $Z = [z_i, ..., z_n]$ is described by the indicators

$$J_i = \begin{cases} 1 & \text{if a response error occurs} \\ 0 & \text{otherwise} \end{cases} \qquad (A3)$$

We allow $P[J_i = 1 \mid x_i]$ to depend explicitly on x_i. The error magnitudes, if errors occur, are i.i.d. random variables D_i, i=1,...,n, and are totally independent of (X, ϵ). Therefore $\delta_i = J_i\ D_i$ and $\delta_i = 1$ iff $z_i \neq y_i$.

Our Simulation 1 follows the structure described above after carrying out log transformations on (X, Y, Z). The vector X is drawn from a two-group mixed lognormal distribution. The group membership indicators c_i are unobservable. They represent the "Large" vs. "Small" groups described in Appendix B below. The multiplicative group-specific ratios $r_i = y_i/x_i$ correspond to $\epsilon_i = \log(y_i) - \log(x_i)$. The function f is linear on the log scale. The probability $P[z_i \neq y_i]$ is a decreasing logistic function defined in Sec. B.1. The multiplicative response errors m_i take the form $\exp(J_i\ D_i)$, given that a response error occurs.

Appendix B. Details of Simulations

Here we describe our simulation methodology and specific distributional parameters for simulated data. The simulated X, Y and Z variables are generated on the original scale, and errors are lognormal values which multiply a randomly selected set of Y values to create the observed Z variables. Recall that only (X, Z) are observable. Editing methods attempt to identify observations with $z_i \neq y_i$, based on observing only (X, Z).

Appendix B.1 Simulation 1

We begin by generating X as a mixture of two scaled lognormal components, a "Large" component of 200 observations and a "Small" component of 800 observations. Recall that a lognormal variable is defined as a random variable whose logarithm is normally distributed with parameters (μ,δ), the log-mean and log-standard deviation. The parameters of the X distribution are tabulated below.

Table B1

Component	log-mean	log-std	scale factor
Large Component	0.20	1.25	100
Small Component	-0.75	1.10	100

Given X, the Y values are created by multiplying each x_i by a random scaled lognormal growth rate r_i. The growth rates are distributed differently in the Small and Large components. They will depend on the desired $\text{corr}(X, Y)$ in Table 3 of Section 4.2. See the table below.

Given Y, a randomly selected set of Y values will be multiplied by a scaled lognormal error m_i to obtain z_i. The probability that a y_i will be multiplied by an error is

$$P[\text{error in } z_i \mid x_i] = P[z_i \neq y_i \mid x_i] = 0.10/(1 + \exp(-5 + 0.008(x_i - 380))).$$

If y_i is not selected, we set $z_i = y_i$. Note that the occurrence of response error depends only on the size of x_i and not the component from which x_i was drawn. The m_i are independent and identically distributed and independent of (x_i, y_i). Their distribution will vary, depending on the desired corr(X, Z) in Table 3 of Section 4.2.

The following table specifies the distributional parameters for each of the six correlation scenarios in Table 3.

Table B2

Scenario	1	2	3	4	5	6
corr(X, Y)	0.95	0.95	0.95	0.81	0.81	0.81
corr(X, Z)	0.90	0.73	0.50	0.73	0.50	0.28
Large group						
log-mean	0	0	0	0	0	0
log-std	0.055	0.055	0.055	0.090	0.090	0.090
Small group						
log-mean	0	0	0	0	0	0
log-std	0.50	0.50	0.50	0.85	0.85	0.85
Errors						
log-mean	0	0	0	0	0	0
log-std	1.00	1.31	1.62	1.01	1.48	1.83
scale	0.90	0.90	0.90	0.95	0.95	0.95

Notice that corr(X, Y) only depends on the growth rate, but corr(X, Z) depends on both the growth rate and error parameters.

For each of the six scenarios, 10,000 replicated data sets were generated and then edited using HB (C=20, U=1/2) and RR (alpha=0.4). The resulting Type I and Type II rates and the sample FDR were summarized.

Appendix B.2 Simulation 2

In Simulation 2 we used end-of-year closing prices of securities listed on the New York Stock Exchange as the *X* and *Y* variables. Our data included 2185 stocks out of the 2189 listed firms; four firms were dropped from the analysis because their prices exceeded the other 2185 by several orders of magnitude.

The year-to-year correlations between closing prices ranged from 0.85 (for years 2006-2007) to 0.97 (for years 2013-2014). We chose to analyze years 2006-2007, the beginning of the Great Recession, and years 2010-2011, when the market had stabilized after the recession. In each case the earlier year was treated as *X* and the following year was treated as *Y*. We did not treat the stock closings as a mixture of small and large stocks, as in Simulation 1.

The *Z* variables were created by randomly selecting about 10% of the *Y* values and multiplying the selected y_i by independent scaled lognormal variables m_i to create z_i. If y_i was not selected, we set $z_i = y_i$. The presence of error and the magnitude of errors, if any, were independent of *X* and *Y*. The parameters of the scaled lognormal errors are tabulated below.

	log-mean	log-std	scale factor
Years 2006-2007	0	1	0.9
Years 2010-2011	0	1	0.9

The resulting (*X,Y,Z*) data were then edited as described in Sec. 5.2.