Data Modeling for a Simulation Study of the Quarterly Financial Report Estimator

Donald M. Luery
U.S. Census Bureau, Washington, DC 20233
donald.m.luery@census.gov

ABSTRACT

U.S. Census Bureau staff conducted a simulation study into alternative estimators for the Quarterly Financial Report (QFR). The QFR collects income statement and balance sheet data quarterly from samples of manufacturing, mining, wholesale trade, and retail trade companies. QFR data was first published in 1947 by the Federal Trade Commission and was transferred to the Census Bureau in 1982. The QFR uses a post-stratified estimator that does not directly use the sample weights. The post-strata totals are unknown so that they are estimated by projecting forward sample estimates (weighted) from previous years. Several research projects have been conducted looking into aspects of this estimator but none have been a comprehensive investigation looking at all aspects of the sample design and estimation. In 2003 and 2004, the Census Bureau conducted such an investigation through a simulation study. This involved generating time series data for each of the companies in an artificial population. This paper discusses modeling the QFR data and generating the time series.

Keywords: Time series, Mixed model, Markov chain, Student-t distributed noise

1. Introduction

The Quarterly Financial Report (QFR) program is a quarterly survey that provides up-to-date aggregate statistics on the financial results and position of U.S. corporations. The QFR publishes estimated statements of income and retained earnings, balance sheets, and related financial and operating ratios for the domestic operations of manufacturing, mining and trade corporations. The statistical data are classified by industry and by asset size. The primary users of the QFR are governmental organizations charged with economic policy-making responsibilities. QFR data have "principal economic indicator" status and are essential to the calculation of such key national economic performance measures as the Gross Domestic Product and Flow of Funds Accounts. (QFR Web Page)

[*] **Disclaimer:** This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical and methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

The basic sample design and estimator used for the QFR have remained unchanged since the 1950's. The sampling unit is a corporation. The QFR has a rotating panel design in which eight panels are in the survey for any given quarter. Each year, the QFR draws a new 'half sample' for the noncertainty strata from the annual corporate tax returns from the most recent tax year and splits it into four panels. Each quarter, the QFR introduces one of the four new panels and drops the oldest panel from the previous year's half sample. The QFR stratifies each half sample by industry (called sample industry) and asset class, a classification into six size groups primarily by assets. The selection is with equal probability within each stratum. (Sands, 1984)

The QFR estimator of a total differs from a traditional design-based estimator because the QFR does not base the weight for a sample corporation on its initial probability of selection. Instead, the QFR assigns weights based on a post-stratification defined by the industry reported by the corporation during enumeration (called the enumeration industry) and the original asset classes. The weight for a corporation equals the ratio of the estimated total number of corporations in its post-stratum to the number of sample corporations in the post-stratum. The total number of corporations in a post-stratum is not known so it must be estimated. This is accomplished using a moving average of estimated post-strata totals that are estimated using the sample weights. The moving average is over the estimated totals for the current year and previous year or previous two years (Sands, 1992).

The history of the development of the estimator is not known. After the QFR program migrated to the Census Bureau from the Federal Trade Commission in December 1982, there were several efforts to study the properties of the estimator (Trager and Zarrett, 1993) and compare it to a fixed weight, design-based estimator (Chapman and Biemer, 1984). These efforts indicate that the QFR estimator is biased (with respect to repeated sampling) but may have a smaller variance than the fixed weight estimator. Phil Kott (1990) developed a combined model and design-based approach that suggests that the QFR estimator could be unbiased (in a combined model and design-based sense), if the specified model is correct.

In 2003, a team was formed to investigate the QFR estimator and to make recommendations on changes to

this estimator as needed. The two most important issues addressed by this team (Caldwell, 2005) are:

Issue 1: 'Some sample units are classified in one industry when selected for sample, but in another industry when reporting… Should the estimator's weighting reflect the original probability of selection, or the sample frequency in the (post-) strata at the time of reporting?'

Issue 2: 'The population changes in size over time. Meanwhile, the sample rotation groups [panels], selected according to the population at a specific time, typically are of different sizes, and can change in size further due to random and nonrandom effects. How should the estimator's weighting account for the uncertainty in the size of the population?'

The team decided to conduct a simulation study. The advantage of this would be that 'truth' would be known and all major features of the design and estimation could be simultaneously represented. For the simulation study to produce valid results, the artificial populations must mimic the salient features of the QFR populations in the distribution of the data in any quarter and, particularly, the change in company data over time.

Along with the current estimator, the study investigated nine other estimators that varied in the estimation of the post-strata totals. As in the current estimator, the final weights for companies did not use their sample weights but were the same for each company in a post-stratum. The study also investigated the simple expansion estimator that used the sample weights without post-stratification.

The study selected five variables for the analysis – sales; inventory; net property, plant and equipment (NPPE); net income before taxes (NIBT); and net income after taxes (NIAT). The study simulated sixty quarters of data that reflected for each company the dynamics and variability of the QFR data. The modeling and simulation had two components

- Data simulation for quarters 1 and 2 (Caldwell, 2005)
- Change model – The change models simulated the dynamic change in the data over time. The variables sales, inventory and NPPE, which are strictly non-negative, and NIBT and NIAT, which can take on negative as well as positive values, are inherently different. Different modeling strategies were used for the non-negative variables versus NIBT and NIAT. The model for NIBT and NIAT can be found in (Caldwell, 2005).

This paper discusses the modeling and simulation of the data for sales, inventory and NPPE. Section 2 presents the models used to simulate the quarterly change for three variables – sales, inventories, and NPPE. Section 3 presents an evaluation of the models, section 4 discusses some aspects of the generation of the simulated data, and section 5 presents an evaluation of the simulated data. Section 6 presents a few concluding remarks.

2. Change Modeling for Sales, Inventories, and NPPE

The purpose of the change model was to project forward for 58 additional quarters the time series for each company created in the simulation study. The change model models the change for companies from one quarter to the next. We do this by modeling the log of the ratio of current quarter to previous quarter. Eight quarters of QFR data were available to develop the models. There are three components to the change model for sales, inventory, and NPPE. We modeled each study variable separately. The three models examined were:

1. Zero observation model,
2. Zero/non-zero change model, and
3. Non-zero change model.

These are summarized next and the latter two described in more detail below.

Zero observation model. The log of zero does not exist so that a change model based on logs cannot predict a zero observation. Inspection of the QFR data indicated that if a value of an item for a company was zero in one quarter then it was usually zero in the other quarters. The model for zero observations is – if the value is zero for the first quarter, all subsequent quarters would be zero. If the first value is non-zero then the following two model components are used.

Zero/non-zero change model. The non-zero change model is a model for continuous variables and, as such, simulated values of zero change have probability zero of occurring. However, small companies can often have zero changes and these zero changes do not occur at random. That is, whether a quarterly change is zero or not depends on whether the changes for the previous quarters are zero or not. Because of the shortness of the data series available for modeling – from one up to seven changes per company, models in which the probability of change depended on more than the most recent change were not investigated. A model where the probability of a zero change only depends on whether the previous quarterly change is zero or not is a Markov chain. We used a Markov chain model to determine whether a change will be zero or whether we use the non-zero change model to create the value of a change. The zero/non-zero change model was used only for inventory and NPPE for the asset classes 03 and 07, the asset classes for the smallest companies, because there were few zero changes for the larger companies and for sales.

The table 1 shows the percentages of zero changes for each variable for assets classes 03 and 07.

Separate models by asset classes would be able to capture dependency on size. The effect of industry on the probability of a zero change was explored through nominal logistic regressions for each of these two asset classes. These analyses did not detect significant industry effects.

It might be expected that companies within an asset class would have different transition probabilities in the Markov chain model. Some companies might persist with zero changes more often than other companies. This was explored by a hierarchical, random effects model using a Dirichlet distribution, which is conjugate for the multinomial distribution that was used in the modeling of the transition probabilities. See section 2.1 for a detailed discussion of this model.

Non-zero change model. We used a first-order auto-regressive model for the non-zero change model, that is, a change is proportional to the previous change plus a random disturbance. Instead of modeling the random disturbances using a normal distribution, we used a *t*-distribution. Analysis of the distribution of non-zero changes showed that large (non-outlier) changes occur much more frequently than would be expected from a normal distribution. A *t*-distribution can be used to model this feature of the data. We established upper and lower bounds for the relative change that decreased with the size of the previous quarter's level. For example, a company with high sales in the previous quarter would be limited to a smaller relative increase than a company with low sales and vice versa. See section 2.2 for a detailed discussion of this model.

2.1 Zero/non-zero Change Model

The zero/non-zero change model is a two-state random effects Markov chain model. Let $z_{it} = 1$ if the change is nonzero and 0 otherwise for company i at time t. Let the joint distribution of $z_{i,t-1}$ and z_{it} be $p(z_{i,t-1} = j, z_{it} = k) = \pi_{jk}(i)$ for $j, k \in \{0, 1\}$ and let $\pi_0(i) = \pi_{00}(i) + \pi_{01}(i)$ and $\pi_1(i) = \pi_{10}(i) + \pi_{11}(i)$ be the marginal probabilities of being in states 0 and 1. The transition probabilities in the two-state Markov chain are then

$$p(z_{it} = k \mid z_{i,t-1} = j) = p_{jk}(i) = \pi_{jk}(i) / \pi_j(i).$$

The probabilities, $\pi_{jk}(i)$ for $j, k \in \{0, 1\}$, will be modeled as random company effects, that is, each company will have a different set of probabilities which are generated from a common distribution. The $\pi_{jk}(i)$ are modeled as Dirichlet with parameters $(a_{00}, a_{01}, a_{10}, a_{11})$. Redefine these parameters as $\mu_{jk} = a_{jk} / \tau$ where $\tau = a_{00} + a_{01} + a_{10} + a_{11}$. The first two moments for $\pi_{jk}(i)$ are mean $= \mu_{jk}$ and variance $= \mu_{jk}(1-\mu_{jk})/(\tau+1)$. τ is comparable to a sample size and small τ indicates a substantive company effect. The Dirichlet distribution has the form

Table 1. Percentages of Zero Changes in Log Ratio

Asset Class	Sales	Inventory	NPPE
03	1.37	23.68	13.06
07	0.22	11.04	4.51

$$p(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}; a_{00}, a_{01}, a_{10}, a_{11}) = \frac{\Gamma(\tau)}{\prod_{j=0}^{1}\prod_{k=0}^{1}\Gamma(\mu_{jk}\tau)} \prod_{j=0}^{1}\prod_{k=0}^{1} \pi_{jk}^{\mu_{jk}\tau - 1}.$$

The Dirichlet distribution was used to generate the $\pi_{jk}(i)$ for a company. The transition probabilities $p_{jk}(i)$ for a company were determined from these $\pi_{jk}(i)$ and the evolution of whether a change will be zero or not were based on these transition probabilities. The initial change was determined by the data simulation for the first two quarters (Caldwell, 2005).

An extension of this model whereby the probability of a zero change might depend on the value of the change from the previous quarter (not just whether it was zero or not) was not explored.

To fit the parameters π_{00}, π_{01}, π_{10}, π_{11} and τ, let for company i,

$x_{00}(i)$ = # of changes from zero to zero,
$x_{01}(i)$ = # of changes from zero to non-zero,
$x_{10}(i)$ = # of changes from non-zero to zero,
$x_{11}(i)$ = # of changes from non-zero to non-zero,

and

$$n_i = x_{00}(i) + x_{01}(i) + x_{10}(i) + x_{11}(i)$$

then the $(x_{jk}(i), j, k \in \{0, 1\})$ are multinomial with parameters $(\pi_{jk}(i), j, k \in \{0, 1\}, n_i)$.

Combining the model for the $x_{jk}(i)$ with the Dirichlet model for the $\pi_{jk}(i)$, we have that the $(x_{jk}(i), j, k \in \{0, 1\})$ are distributed Dirichlet-multinomial with parameters $(\mu_{jk}, j, k \in \{0, 1\}, \tau, n_i)$. The probability distribution function for the Dirichlet-multinomial is

$$p(x_{jk}(i), j, k \in \{0, 1\}; \mu_{jk}, j, k \in \{0, 1\}, \tau, n_i) =$$

$$\binom{n_i}{x_{ij}(i), j, k \in \{0, 1\}} \frac{\Gamma(\tau)}{\prod_{j=0}^{1}\prod_{k=0}^{1}\Gamma(\mu_{jk}\tau)} \frac{\prod_{j=0}^{1}\prod_{k=0}^{1}\Gamma(x_{jk}(i) + \mu_{jk}\tau)}{\Gamma(n_i + \tau)}$$

SAS™ PROC NLP was used to fit this model.

2.2 Non-zero Change Model

2.2.1 Preliminary Analysis of Log Ratio Change of Sales, Inventory and NPPE

The distribution of the log ratio change for these series were initially investigated through histograms, normal quantile plots, and estimates of moments. The initial exploration was by asset class. This analysis showed substantial kurtosis for all of the series after removal of a few extreme outliers and little skewness for sales and

Table 2. Skewness and Kurtosis for Log Ratio Change

Variable	Moment	03	07	08	14	16	18
Sales	Skewness	–0.35	–0.11	–0.54	–0.20	–0.13	–0.72
	Kurtosis	6.10	7.10	16.44	8.81	14.90	25.13
	Degrees of Freedom	4.98	4.85	4.36	4.68	4.40	4.24
Inventory	Skewness	–0.20	0.10	–0.27	–0.44	0.16	–0.67
	Kurtosis	8.55	8.17	8.38	8.99	17.05	29.91
	Degrees of Freedom	4.70	4.73	4.72	4.67	4.35	4.20
NPPE	Skewness	0.67	1.59	1.37	1.60	1.75	–1.35
	Kurtosis	10.15	21.07	18.24	25.03	43.47	85.50
	Degrees of Freedom	4.59	4.28	4.33	4.24	4.14	4.07

inventory and moderate skewness for NPPE. The table 2 shows the skewness and kurtosis by asset class for each of the variables after removing extreme changes. The asset classes are 03, 07, 08, 14, 16, and 18 where 03 is the asset class for the smallest companies, 16 is the asset class for the largest non-certainty companies, and 18 is the certainty asset class. Large kurtosis can be modeled by a *t*-distribution. For a *t*-distribution, the kurtosis $\kappa = 6/(\nu - 4)$ where ν is the degrees of freedom. In terms of the kurtosis, the degrees of freedom $\nu = 4 + 6/\kappa$. The comparable degrees of freedom are also shown in the following table.

For a normal distribution, the skewness and kurtosis would be zero and the degrees of freedom infinite. T-distributions with degrees of freedom 30 or greater are considered approximately normal. All of these distributions show marked departures from normality.

2.2.2 Tree Partition

Since one of the effects that the study was trying to emulate was the reclassification of industries, it was felt that a change model should reflect both sample industry and enumeration industry along with asset class. Including either one or the other might under represent in the simulated data the variation caused by the reclassification. One approach to modeling would be to stratify the data by asset class, sample industry, and enumeration industry and fit different models in each partition. However, the sample sizes in these would be relatively small for estimating well the parameters in the change model. One way to improve the parameter estimation would be to borrow strength from the other industries.

To this end, SAS™/Enterprise Miner was used to form tree partitions of the data. The partitioning variables would be asset class, sample industry, and enumeration industry. The target variable for forming the partitions was the absolute value of the residual where the residuals were obtained from an initial run of the mixed model described below. Partitions were formed so that they differed by the average absolute residuals. In this way, differences in variation among companies in different asset classes, sample industries, and enumeration industries could be explained by the model. During the partitioning, an attempt was made to create partitions within a single asset class because asset class was not subject to change in the survey and substantial variation would be expected for different sizes companies. Subsequent to an initial manual partitioning using Enterprise Miner, partitions with few observations were combined with other partitions so that there would be approximately a minimum of 800 observations in each partition. This minimum sample size was felt to ensure adequate sample to estimate well the parameters in the change model. The partitions were combined if they had similar average absolute residuals and standard deviations of the absolute residual. After this, the following mixed model was estimated separately in each partition.

2.2.3 Mixed model

Initially, an autoregressive (AR) 1 model (Box and Jenkins, 1976) was fit to the log ratio change data for each of the partitions. The AR coefficients were usually negative indicating that the expected change in the current quarter would be in the opposite direction from the change in the previous quarter. Examination of the residuals and absolute residuals indicated, in general, negative correlations of each of these with the log of the previous quarter's data. The former would indicate that a company change in the current quarter would be less when the previous quarter's estimate was large and greater when it was small. The latter would indicate that the spread in the log ratio change (relative change in the original scale) would be less when the previous quarter's estimate was large and greater when it was small. Preliminary models were fit to examine this using loglinear variance models (Harvey 1976, Carroll and Ruppert 1988) in JMP. These models have the general form

mean model: $E(y) = \mathbf{X}b$

variance model: $\log(Variance(y)) = \mathbf{Z}\lambda$.

Based on this preliminary modeling, the following mixed model was selected to model the QFR data. It was estimated separately for each partition.

$$y_{it} = \mu_{it} + e_{it}$$

where

i = company,

t = time,

$\mu_{it} = r_i y_{i,t-1} + \alpha + \beta \log(x_{i,t-1})$

x_{it} = reported data for company i and quarter t,

$y_{it} = \log (x_{it} / x_{i,t-1})$, log ratio change,

Table 3. Dirichlet Model for Inventory

Variable	Asset Class	Parameters					Transition Probabilities	
		μ_{00}	μ_{01}	μ_{10}	μ_{11}	τ	Zero to Non-zero Change	Non-zero to Non-zero Change
Inventory	03	0.107 (0.011)	0.112 (0.011)	0.107 (0.011)	0.675 (0.018)	2.935 (0.405)	0.511	0.863
	07	0.049 (0.006)	0.054 (0.006)	0.072 (0.007)	0.825 (0.011)	2.935 (0.405)	0.523	0.919
NPPE	03	0.044 (0.007)	0.081 (0.010)	0.084 (0.010)	0.791 (0.015)	4.170 (0.813)	0.650	0.904
	07	0.016 (0.004)	0.031 (0.005)	0.034 (0.005)	0.918 (0.008)	4.170 (0.813)	0.658	0.964

r_i = random autoregressive effect for company i with mean ρ and variance v,

e_{it} = t-distributed random effect for company i and time t with degrees of freedom df, mean 0, and variance parameter σ_i^2 where

$$\log\left(\sigma_i^2\right) = a_\sigma + b_\sigma \log\left(x_{i,t-1}\right),$$

and β and b_σ represent the expected inverse relationships of the log ratio and $\log\left(\sigma_i^2\right)$ with $\log\left(x_{i,t-1}\right)$.

The model fitting was conducted in two steps. First, the linear model was fit using SAS™ PROC MIXED. The residuals were output and the variance model described above for e_{it} was fit using SAS™ PROC NLP.

3. Evaluation of the Models

3.1 Zero/non-zero Change Model

Table 3 provides the parameters and transition probabilities for the Dirichlet model for inventory and NPPE. The standard errors for the parameter estimates are in parentheses.

τ is the equivalent of the sample size. The small τ's indicate a substantive company effect. Each $\pi_{jk}(i)$ is equal to a gamma ($\mu_{jk}\tau$, 1) random variate divided by the sum over j and k ($\mu_{jk}\tau$, 1) random variates. Johnson and Kotz (2000) show that if Y_0, Y_1, … Y_m have a joint Dirichlet distribution with parameters θ_j (j = 0,1, … , m) then each $Y_j = X_j \Big/ \sum_{i=0}^{m} X_i$ where the X_j are χ^2 with $2\theta_j$ degrees of freedom. It is easy to show that $X_j = 2Z_j$ where Z_j is gamma (θ_j, 1). When $\mu_{ij}\tau$ is small, the gamma random variates are skewed and highly kurtotic. For inventory and asset class 07, for example, $\mu_{00}\tau$ is about 0.15, the skewness is about five and the kurtosis is about 40. This led to microscopic gamma random variates and very unstable $\pi_{jk}(i)$. τ was arbitrarily multiplied by 10 to increase stability. This was an expedient solution to this problem. If time had permitted, further investigation would have been desirable. A subsequent analysis, while preparing this paper, based on a simulation from the gamma distributions indicated that this adjustment might not have been necessary.

3.2 Mean Component of the Mixed Model

Tables 4, 5, and 6 provide summaries of the modeling for each of the variables. The square root of v is given instead of v to aid in the comparison with ρ.

The autoregressive coefficient is usually negative for sales (15 out of 16) and inventory (11 out of 11) as

Table 4. Parameter Analysis for Sales Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
ρ	16	–0.374	–0.210	0.056	15	15
$\sqrt{v}$	16	0.073	0.235	0.465	13	14
α	16	–0.043	0.183	0.646	9	9
β	16	–0.095	–0.023	0.003	7	11

Table 5. Parameter Analysis for Inventory Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
ρ	11	–0.235	–0.167	–0.038	8	9
$\sqrt{v}$	11	0.194	0.265	0.353	9	11
α	11	–0.015	0.097	0.339	6	7
β	11	–0.046	–0.012	0.001	5	7

Table 6. Parameter Analysis for NPPE Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
ρ	16	–0.072	0.110	0.249	9	10
$\sqrt{v}$	16	0.163	0.273	0.544	12	15
α	16	–0.067	–0.027	0.060	1	3
β	16	–0.008	0.003	0.009	1	3

expected but it was positive 15 out of 16 times for NPPE. They are not large but the p-values indicate they are significantly different from zero. The p-values for the mixed model variance parameter, ν, are also significant indicating that there is substantive variation among companies. The values are large enough to indicate that companies can have both positive and negative autocorrelations with a bias toward negative correlation for sales and inventory and positive correlation for NPPE. The slope coefficient, β, was negative all but once for sales and inventory and showed significant p-values though less often than that of ρ and ν. The coefficients are, in general, not large and might have been removed from the model. The slope and constant, α, are highly negatively correlated (–0.9) and the constant would have been removed along with the slope. The slope and constant for the NPPE models were not significant and should have been omitted.

3.3 Variance Component of the Mixed Model

The tables 7, 8, and 9 provide summaries of the modeling for each of the variables in the variance component of the mixed model.

The slope coefficients, b_σ, are almost always negative for sales (15 out of 16) and always negative for inventory and NPPE as was expected. The p-values indicate that they are significantly different from zero in general. The degrees of freedom are all small and always less than four. Thirteen out of sixteen times the degrees of freedom for NPPE were estimated to be less than two indicating a *t*-distribution without a finite second moment.

The mixture of *t*-distribution was also fit for NPPE. The results are not shown here but they were less successful. Generally, either one of the degrees of freedom in a mixture was estimated with a very large standard deviation or the degrees of freedom were not substantively different. The means were generally near zero indicating that skewness was not identified by this modeling.

Table 7. Variance Parameter Analysis for Sales Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
a_σ	16	–2.747	–0.767	0.153	9	11
b_σ	16	–0.196	–0.119	0.042	10	13
df	16	1.928	2.977	3.898	16	16

Table 8. Variance Parameter Analysis for Inventory Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
a_σ	11	–2.190	–0.950	0.092	9	10
b_σ	11	–0.291	–0.125	–0.010	10	10
df	11	1.767	2.501	3.514	11	11

Table 9. Variance Parameter Analysis for NPPE Mixed Model

Parameter	No. of Nodes	Min	Med	Max	No. p-values ≤ 0.01	No. p-values ≤ 0.05
a_σ	16	–3.267	–1.924	–0.566	16	16
b_σ	16	–0.426	–0.154	–0.035	16	16
df	16	1.235	1.862	2.186	15	16

4. Generation of the Simulated Data

This section discusses a few issues in the generation of the simulated data series.

4.1 Generating the Random Autoregressive Effect r_i

The autocorrelation coefficient, r_i, must fall in the interval (–1, 1). Using normal random variates to generate a company's autocorrelation coefficient might generate a coefficient outside of this range. In order to generate reasonable r_i, it was assumed that a transformation of the r_i into the (0, 1) interval had a Beta distribution. The parameters in the Beta distribution were found through the method of moments. The expected value of r_i is ρ and its variance (v) was estimated from the mixed model. Define $q_i = ½(r_i+1)$ then q_i falls in the interval (0, 1) and has expectation $Q=½(\rho+1)$ and variance $V=¼v$. q_i was modeled using a Beta distribution with positive parameters a and b. The expectation and variances for a Beta distribution are

$$E(q_i)=\frac{a}{a+b} \text{ and } Var(q_i)=\frac{ab}{(a+b)^2(a+b+1)}.$$

By matching the moments, the Beta parameters are

$$a=\frac{Q}{V}\left(Q(1-Q)-V\right) \text{ and } b=\frac{(1-Q)}{V}\left(Q(1-Q)-V\right).$$

Substituting for Q and V we have that

$$a=\frac{1}{2}\frac{1+\rho}{v}\left(1-\rho^2-v\right) \text{ and } b=\frac{1}{2}\frac{1-\rho}{v}\left(1-\rho^2-v\right)$$

For the method of moments to work, $\rho^2 + v < 1$ which was found to be true in all cases. The q_i are generated from the Beta distribution with these parameters and $r_i = 2q_i–1$.

4.2 Generation of a *t* Random Variate

The *t*-distribution is a mixture of a normal and a gamma distribution. Let $y|\mu$, t be distributed $N(\mu, t^{-1})$ where t is the precision, i.e., replace σ^2 by t^{-1} in the standard parameterization. This parameterization is more convenient for generating a *t*-distributed random variate when a gamma random number

generator is available. (If we had used σ^2 as the random variable instead of t then σ^2 would have had an inverse gamma distribution.) Let t be a random variable where $t = \sigma^{-2}\chi^2/df$ where χ^2 is distributed chi-square with df degrees of freedom and σ^2 is a constant. A chi-square distribution is a gamma distribution with parameters $\alpha = ½\ df$ and $\beta = 2$. It follows that $t\,|\,df,\sigma^2$ is distributed gamma with parameters $\alpha = ½\ df$ and $\beta = 2(df\,\sigma^2)^{-1}$ so that $Et = \sigma^{-2}$ and $\text{var}(t) = 2\sigma^{-4}/df$. To generate a t random variate y, generate a gamma random variate z from a $\Gamma(½df,\ 1)$ and let $t = \beta z = (2z/df)\,\sigma^{-2}$. Generate a t random variate $y = \mu + t^{-½}u$ where u is distributed $N(0, 1)$.

4.3 Estimation of x_{it}

The naïve estimate of $x_{it} = x_{i,t-1}\exp(y_{it})$ has a positive bias. First, $\exp(y_{it})\,|\,r_i,t,x_{i,t-1},y_{i,t-1}$ is log normal because $y_{it}\,|\,r_i,t,x_{i,t-1},y_{i,t-1}$ is $N(\mu_{it},t^{-1})$. Thus $E(\exp(y_{it})\,|\,r_i,t,x_{i,t-1},y_{i,t-1}) = \exp(\mu_{it}+½t^{-1})$. Finally,

$$E(x_{it}\,|\,r_i,t,x_{i,t-1},y_{i,t-1}) = E(x_{i,t-1}\exp(y_{it})\,|\,r_i,t,x_{i,t-1},y_{i,t-1})$$
$$= x_{i,t-1}\exp(\mu_{it}+½t^{-1}).$$

A biased corrected estimate is $x_{it} = x_{i,t-1}\exp(\mu_{it}-½t^{-1})$.

4.4 Controlling Extreme Changes

The above-generated extreme changes fell outside the changes found in the data. This can be seen in table 10 for log ratio change sales for asset class 03. The table shows that the simulation generated a change outside the lower range for the QFR data. Other variables and asset classes showed extreme changes in the simulated data both above and below the observed ranges of the QFR data. This can have the consequence of generating level estimates well above the range of the observed data or generating microscopic level estimates. Inspection of the data indicated that the size of the largest positive changes decreased as the level, $x_{i,t-1}$, increased and the size of the largest negative changes decreased as the level, $x_{i,t-1}$, decreased. This section describes the development of upper and lower bounds for the change in the log ratio.

Table 10. Extreme Observations for Log Ratio Sales for Asset Class 03

QFR Data		Simulated Data	
Lowest	Highest	Lowest	Highest
-3.13927	1.52364	-10.52255	1.30225
-2.59650	1.66575	-2.15833	1.41523
-2.56110	1.72238	-1.98180	1.41787
-2.12124	1.84124	-1.98180	1.46206
-1.77834	2.01292	-1.78077	2.48730

See Figure 1 for a plot of the log ratio change data for sales for asset class 03 and the upper and lower bounds.

The upper and lower bounds are linear functions of the log of the level of the variable from the previous period. Let $z_{it} = \log(x_{it})$ and $z_{i,t-1} = \log(x_{i,t-1})$ be the logs at times t and t–1. The upper bound for the change, y_{it} is $ub_i = a_u + b_u z_{i,t-1}$ and the lower bound is $lb_i = a_\ell + b_\ell z_{i,t-1}$.

The linear functions for the upper and lower bounds were developed by plotting the log ratio change (y_{it}) versus the log of the variable from the previous time ($z_{i,t-1}$). Using data from all industries within an asset class, upper and lower bounding points along the range of $z_{i,t-1}$ were selected by hand excluding observed outliers. For each asset class, separate regression lines were fit to these bounding points to generate the above regression coefficients. Because all of the points used to estimate the regressions were considered acceptable and using the estimated regressions lines would now identify 'half' of these bounding points as outliers, a constant of 0.1 was added to the upper bound and 0.1 was subtracted from the lower bound. This constant was chosen so that the adjusted bounds, $ub_i = 0.1 + a_u + b_u z_{i,t-1}$ and $lb_i = -0.1 + a_\ell + b_\ell z_{i,t-1}$, included almost all of the bounding points.

The time and unit subscripts are not needed for the following development and will be omitted. The upper and lower bounds were applied to the adjusted log ratio change $y' = y - ½\tau^{-1}$. If the generated random variate y' was greater than ub, it was reduced to ub and, if it was

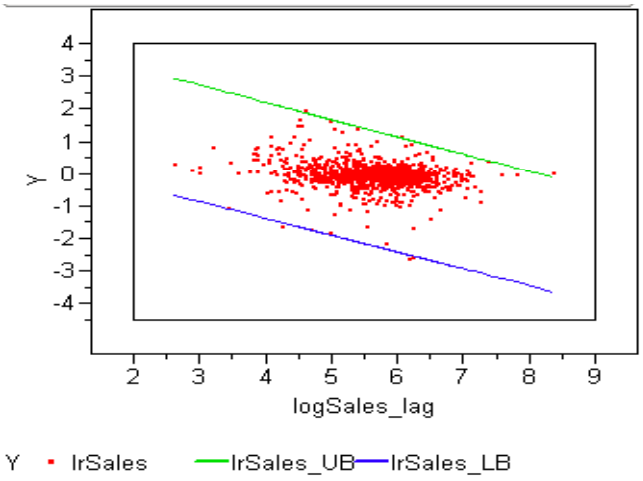

Figure 1. Upper and Lower Bounds for Log Ratio Sales for Class 03

less than lb, it was increased to lb. These bounds would have the effect of biasing the simulated data. The following was used to correct for this bias.

Define the bounded adjusted log ratio change y'' to be

$$y'' = \begin{cases} ub & \text{if } y' \geq ub \\ y' & \text{if } lb < y' < ub \\ lb & \text{if } y' \leq lb \end{cases}$$

Now,

$$E\left(\exp\left(y''\right)\middle|\tau\right) = \int_{-\infty}^{\infty} \exp\left(y''\right) P\left(y \mid \mu, \tau\right) dy$$

$$= \exp\left(ub\right)\left(1 - \Phi\left(\left(ub - \mu + \tfrac{1}{2}\tau^{-1}\right)\tau^{\frac{1}{2}}\right)\right)$$

$$+ \exp\left(\mu\right)\begin{pmatrix} \Phi\left(\left(ub - \mu - \tfrac{1}{2}\tau^{-1}\right)\tau^{\frac{1}{2}}\right) \\ -\Phi\left(\left(lb - \mu - \tfrac{1}{2}\tau^{-1}\right)\tau^{\frac{1}{2}}\right) \end{pmatrix}$$

$$+ \exp\left(lb\right)\Phi\left(\left(lb - \mu + \tfrac{1}{2}\tau^{-1}\right)\tau^{\frac{1}{2}}\right).$$

A bias corrected estimate is

$$x'' = x\exp\left(y''\right)\exp\left(\mu\right)\Big/ E\left(\exp\left(y''\right)\middle|\tau\right).$$

4.5 Bounding the Level Estimates

Despite the bounding of extreme changes the generated level estimate at time t, $x_{it} = x_{i,t-1}y_{it}''$, still showed values noticeably outside the observed range of the QFR data. One final adjustment was used to correct for this. If x_{it} was greater than exp(MAX) then it was set to exp(MAX). If it was less than exp(MIN) then it was set to exp(MIN). MAX and MIN were by asset class and were identified using the QFR data omitting data not in the test-enumerated industries since the data in the study-enumerated industries had a smaller range than the data as a whole. These bounds were rounded and adjusted so that they were monotonically increasing with asset class. No adjustment was made for the biasing due to this bounding.

5. Evaluation of Simulated Data

A simulated dataset paralleling the QFR dataset was used to evaluate the adequacy of the simulated data. For each company, a simulated series was created starting with the initial value for the company in the QFR dataset. The above change models were then used to generate the artificial series for the number of quarters that the company appeared in the QFR dataset. The initial values were then dropped because they would be the same in both datasets. Histograms and the first four moments were compared for asset classes crossed by test enumeration industries. Examination of the histograms and moments indicated that the original and simulated data had similar distributions. This is summarized in the following table for the moments. The table shows the coefficients of regressions of the moments from the simulation study as the dependent variable and the moments from the original data as the independent variable. The intercept was not included in the regressions. A coefficient of one would indicate that the simulated moment was on average equal to the moment from the original data. The table shows that in almost all cases the coefficients were less than one. For skewness and kurtosis except for the sales' kurtosis, the coefficients were close to one indicating that these aspects of the variation in the data were reasonably represented in simulated data.

Table 11. Analysis of Moments of Simulated Data

Variable	Mean	Standard Deviation	Skewness	Kurtosis
Sales	0.875	0.854	0.935	0.869
Inventory	0.875	0.879	1.027	1.030
NPPE	0.893	0.887	0.972	0.941

6. Discussion

The modeling captured the variation over time and between companies in the QFR data and the simulated data closely approximated the distributions of the variables sales, inventory, and NPPE. In retrospect, some streamlining of the models would have been appropriate and a more unified approach for modeling some of the components would have been preferable.

7. References

Box, G. and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Caldwell, C., et al. (2005), "Investigation of Alternative Estimators for the Quarterly Financial Report," draft Internal U.S. Census Bureau Report.

Carroll, R.J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, New York: Chapman and Hall.

Chapman, D.W. and Biemer, P.B. (1984), "A Comparison of Two Estimators for the Quarterly Financial Report Survey," Internal U.S. Census Bureau Report.

Kott, P. (1990), "Post-stratifying with Sample Data Only: Can Using the Obvious Model Help?," *Proceedings of the American Statistical Association, Survey Research Methods Sections.*

Johnson, N.L. and Kotz, S. (2000), *Continuous Multivariate Distributions*, 2nd Ed., New York: Wiley

Harvey, A.C. (1976), "Estimating Regression Models with Multiplicative Heteroscedasticity," *Econometrica*, 44, 461–465.

Sands, M.S. (1984), "Explanation of Allocation Procedure Used for the Quarterly Financial Report (QFR) Sample," Internal U.S. Census Bureau Memorandum.

Sands, M.S. (1992), "QFR Estimator and Nonresponse Adjustment," Internal U.S. Census Bureau Memorandum.

Trager, M.L. and Zarrett, P.N. (1993), "Quarterly Financial Report (QFR) Research: Investigation of the Properties of the Current QFR Estimator," Internal U.S. Census Bureau Memorandum, QFR-21-A-16.