Multiple Imputation in the Annual Survey of Manufactures

T. Kirk White**

Jerome P. Reiter*

*Duke University
jerry@stat.duke.edu

**U.S. Bureau of the Census
tkw2@duke.edu

PRELIMINARY AND INCOMPLETE[1]

Introduction

Missing data due to item non-response is a pervasive problem in social and economic data. Often economists simply throw out observations with missing data. Estimates based on the remaining samples can be inefficient or even biased. A large and growing literature in statistics show to impute values for missing data, and a literature dating back at least to Rubin (1987) has shown how to use multiple imputation to estimate the additional uncertainty introduced into the data due to imputation. Methods of single imputation, such as plugging in the industry mean or a ratio estimate, lead to underestimation of uncertainty in many analyses. This paper will apply a particular version of multiple imputation, Raghunathan et al.'s (2001) sequential regression multiple imputation (SMRI), to a particular dataset, the Annual Survey of Manufactures (ASM). The goal is to improve inferences for a commonly used confidential economic dataset. But the SMRI method is applicable much more generally, and thus our empirical results may be of interest to researchers using any dataset that contains missing or imputed items.

Although the SMRI method is due to Raghunathan et al. (2001), here we briefly describe the method and motivate its use. SMRI is a multivariate technique for imputing missing values using a sequence of regression models. The basic idea is to impute X_1 from a regression of X_1 on $(X_2, X_3, etc.)$, impute X_2 from a regression of X_2 on $(X_1, X_3, etc.)$, impute X_3 from a regression of X_3 on $(X_1, X_2, etc.)$, and so on. The regression models are specified to match the distribution of the outcome variable. For example, use a multinomial logistic regression for a multinomial variable, a logistic regression for a binomial variable, and a linear regression for a continuous variable with normally distributed errors. An advantage of this strategy is that it is generally easier to specify plausible conditional models than plausible joint distributions. A disadvantage is that the collection of conditional distributions is not guaranteed to correspond to a proper joint distribution, particularly when the models use different conditioning sets.

Two other advantages of the SMRI method relative to the current Census imputation methods are transparency and flexibility. The Census Bureau routinely imputes values for missing data, but it is often not clear to researchers using the microdata how this imputation was done. And, while the imputation methods used by the Census Bureau (and other data collection agencies) may be appropriate for Census

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The research in this paper was conducted while the first author was an employee of the U.S. Census Bureau at the Triangle Census Research Data Center. This paper has been screened to ensure that no confidential data are revealed.

Bureau goals such as industry tabulations, they may not be appropriate for researchers using confidential microdata. Use of SMRI as a general method for imputation enables researchers to tailor the imputation to their particular application.

Tim Dunne (1998) and other researchers have noted that identifying and dealing with missing and imputed data are important problems for researchers using the Census of Manufactures (CMF) and Annual Survey of Manufactures (ASM) microdata. Until very recently, the ASM and CMF data available in the Census Research Data Centers (RDC) contained no item-level flags to identify imputed data. Dunne (1998) documented ways to identify imputed data in the CMF and ASM. In one industry frequently studied by economists, roughly 40% of the data appears to be imputed (Collard-Wexler, 2007).[2] To the extent that economists deal with missing data issues at all, they typically throw out observations identified as imputed (Foster, Haltiwanger, and Syverson, 2007; Collard-Wexler 2007; and many others). Throwing out observations with missing or imputed data can lead to biased estimates and deflated standard errors (Little and Rubin, 2002). On the other hand, treating imputed data as if it were complete data will typically lead to confidence intervals that are "too narrow", i.e., the confidence intervals will understate the amount of uncertainty in the estimates due to missing data**.**

The U.S. Census Bureau uses a variety of methods to impute data in the ASM and CMF. Based on our analysis of the item-edit flags in the 2002 CMF and the 2003-2005 ASM, one of the most common methods appears to be "cold-deck." The cold-deck method typically involves using data from other records, perhaps in the same industry, sometimes from earlier years, to impute missing data items. While we have made inquiries at the Census Bureau through the appropriate channels, at the time of this writing we have not been able to ascertain precisely how the Census Bureau implements the cold-deck method in the ASM and CMF. Nevertheless, we can say for certain that only one imputation per missing item is recorded in the ASM and CMF data available in the RDCs. From a researchers' perspective, having only one imputation per missing data item will typically cause the researcher to understate the amount of uncertainty in her estimates.

This paper uses new data and new methods to address both problems mentioned above: identifying and dealing with missing data in the ASM/CMF. The Census Bureau does many different types of edits of the data it collects. The Census Bureau classifies some types of edits as "imputed," and others as "not imputed." The 2002 CMF and 2003-2005 ASM data contains item edit flags that identify imputed items, and the name of the method used to impute them. For any data item for which the item edit flag is in the "imputed" category, we consider that data item "missing." We apply the sequential regression imputation method (Raghunathan et al., 2001) to multiply impute missing values. Then we compare our method to a version of an imputation method currently used by the Census Bureau. First we estimate a model of missingness based on the data. We use this model to create missing data, that is, we "poke holes" in the records with complete data. Then we create multiple imputations for these artificially missing values using the sequential regression method, and we create single imputations using a version of the cold-deck method (described below). We find that industry means based on our multiple imputations tend to be closer to the true mean (based on the real data) than estimates from the data imputed (singly) using the cold-deck method. Further, we find that our estimated confidence intervals tend to be wider than the confidence intervals from the single cold-deck imputed data, and our confidence intervals are more likely to cover the confidence intervals from the real data.

Data

As mentioned above, we use the 2002 Census of Manufactures (CMF) and the 2003-2005 Annual Surveys of Manufactures, since these years of the data have item-level edit flags. Roughly a third of the plants in the manufacturing data come from Administrative Records (AR) data. These plants are not sent a survey form. We follow most of the economics literature and drop these plants from our sample. The long-run goal of the project is to estimate plant-level and aggregate total factor productivity (tfp), so we are primarily interested in missing/imputed data in variables that are typically used to compute tfp: the total value of shipments (TVS), the total cost of materials (CM), production workers hours (PH), production

[2] Roughly one third of the data in the CMF comes from establishments which have fewer than 5 employees and are not sent a survey form. Data for these plants comes from Administrative Records (AR). The 40% figure in Collard-Wexler refers to non-AR records. Likewise, throughout this paper, when we refer to percentages of imputed or missing data, we are referring to percentages of non-AR records.

worker wages (WW), total salaries and wages (SW), expenditures on electricity (EE), capital expenditures on buildings (CBE), and capital expenditure on machinery (CME). For each of these variables there is a 3-digit item edit flag explaining whether the recorded value was reported and/or edited or imputed in some way. As mentioned above, we treat as "missing" any data item which the Census Bureau item edit flags classify as "imputed." We then proceed to create our own imputations for these "missing" data items. Table 1 shows the rates of "missingness" for our variables of interest across all non-AR manufacturing plants in each year of our sample. Payroll (total salaries and wages) can be found in administrative records, and the Census Bureau does not classify these payroll items as "imputed" (and thus we do not consider them missing). This may account for the low rate of missingness reported for this variable in Table 1. Among non-AR plants in the 2002 CMF, 44% of the CBE items are "raked." This means that the sum of reported detail items (such as capital expenditure on building and capital expenditures on machinery) do not balance to the reported total (such as total capital expenditures). The Census Bureau changes the details proportionally so that they add up to the reported total. These "raked" items are not considered imputations, and thus we do not count them as missing. This accounts for the low rates of missingness reported in Table 1 for the 2002 capital expenditures.

	2002	**2003**	**2004**	**2005**
Total Value of Shipments	28	29	27	26
Total Cost of Materials	42	35	33	34
Plant Hours (production workers)	39	30	27	27
Worker wages	19	25	22	22
Total Salaries and wages	0	1	0	0
Electricity expenditures	46	30	30	30
Capital expenditures (buildings)	0	36	27	28
Capital expenditures (machinery)	0	36	27	31
Sample size	215,683	64,417	55,645	57,155

Table 1: Percentages of missingness/imputation among all non-AR plants in the 2002 Census of Manufactures and the 2003-2005 Annual Surveys of Manufactures, according to item edit flags.

The Table 2 reports the standard deviations of the percentage missing for each 5-digit NAICS manufacturing industry for each year of our sample. The table shows that the rates of missingness vary considerably across industries within the manufacturing sector.

Variable	**2002**	**2003**	**2004**	**2005**
Total Value of Shipments	14	11	10	9
Total Cost of Materials	12	13	11	10
Plant Hours (production workers)	10	11	9	10
Worker wages	13	11	8	8
Total Salaries and wages	0	1	1	1
Electricity expenditures	14	12	11	10
Capital expenditures (buildings)	0	13	10	10
Capital expenditures (machinery)	0	13	10	10

Table 2: Standard deviations of 5-digit NAICS industry percentage missing, among non-AR plants in the 2002 Census of Manufactures and the 2003-2005 Annual Surveys of Manufactures, according to item edit flags.

An Imputation Model

As mentioned above, Raghunathan et al. (2001) develop a multivariate technique for multiply imputing missing values, SMRI, that uses a sequence of regression models. The basic idea is to impute X_1 from a regression of X_1 on $(X_2, X_3, etc.)$, impute X_2 from a regression of X_2 on

$(X_1, X_3, etc.)$, impute X_3 from a regression of X_3 on $(X_1, X_2, etc.)$, and so on. The regression models are specified to match the distribution of the outcome variable. For example, the user can specify a multinomial logistic regression for a multinomial variable, a logistic regression for a binomial variable, and a linear regression for a continuous variable with normally distributed errors.

In preliminary analysis we found that most of our variables have large positive first order autocorrelations, but insignificant autocorrelations at higher lags. The exceptions were the capital expenditure variables, which also have significant second-order autocorrelations. Thus we hypothesize that an imputation model using just current and one-period lagged values might do a good job. Using just OLS, we tried several different regression specifications for plants with complete data. Based on the R-squared of the regressions our preferred specification is:

$$\ln Y_{ijt} = \alpha_{0j} + \alpha_{1j} \ln X_{1ijt-1} + ... + \alpha_{kj} \ln X_{kijt-1} + \alpha_{k+1,j} (\ln X_{1ijt-1})^2 + ... + \alpha_{2k,j} (\ln X_{kijt-1})^2 + \varepsilon_{ijt} \quad (1),$$

where Y_{ijt} is any of our variables listed in tables 1 and 2, for plant i in year t; the X's are lagged values of all of these variables, including the variable Y; and ε is an error term. We ran this regression for 86 4-digit NAICS industry groups and found that the R-squared of the regression exceeded 0.90 for all variables and all industries.

Using the specification in equation (1) as a starting point, we then applied the method of Raghunathan et al. (2001) using IVEware, code based on the SAS macro language. IVEware implements the sequential regression approach, conditioning on all variables (which in our case includes squared terms) in the specified models. We want to ensure that our industries are homogenous enough so that using the same imputation model for all plants in the industry makes sense. At the same time we need to keep the industry grouping coarse enough to keep the analysis of many industries feasible. To balance these two aims we assume that all establishments in the same 5-digit NAICS industry can use the same imputation model. For this preliminary analysis we allow for a year dummy variable, but otherwise we assume the parameters of the model are the same across all years within the same industry. Thus our full imputation model is:

$$\ln Y_{ijt} = \beta_{j0} + \beta_{j1} \ln X_{1ijt} + ... + \beta_{jk} \ln X_{kijt} + \beta j_{j,k+1} (\ln X_{1ijt})^2 + ... + \beta_{j,2k} (\ln X_{kijt})^2 + \delta_{jt} + \beta_{j,2k+1} \ln Y_{ij,t-1} + ... + \beta_{j,3k} \ln X_{kijt-1} + \beta_{j,3k+1} (\ln Y_{ij,t-1})^2 + ... + \beta_{j,4k} (\ln X_{kijt-1})^2 + \varepsilon_{ijt} \quad (2),$$

where *i* indexes the plant, *j* is the industry, *t* is the year, and *k* indexes the explanatory variable. Note that lagged values of the dependent variable appear as predictors. The parameter δ_{jt} is an industry-specific year dummy. The SMRI procedure first imputes initial values for all missing data, drawn from models estimated with the complete data. The procedure then cycles through all the variables, replacing missing values based on equation (2). That is, for variable Y, the imputations are drawn from the posterior predictive distribution defined by the regression in (2), where the parameters of the regression have non-informative prior distributions. This involves estimating the parameters in equation (2) using the current version of the completed data, then randomly drawing a value of Y using the drawn parameter values. At each draw, the procedure imputes new values for the originally missing values, using the imputed values of covariates from previous iterations. In practice we iterate 10 times for each imputation and keep 20 imputations for each missing value. Thus we do 200 imputations for each missing value and keep 20.

Comparison to Cold-deck Imputation

To assess the performance of the sequential regression imputation method for missing values in the ASM and CMF, we compare our results to results based on a version of "cold-deck" imputation. The single regression method can handle missing values in any of the observed variables and it allows for different types of missing data patterns. In particular, it does not require monotone missingness.[3] However, to keep

[3] IVEware does assume that the missing data mechanism is ignorable.

the comparison simple, we focus on samples with missing items in only one variable at a time. Specifically, we begin by selecting plant-year observations in which the Total Cost of Materials (CM) variable may or may not be missing, but all other variables in table 1 are reported on the survey form (and not imputed).[4] Although the item-edit flags only exist for the years 2002-2005, we want to use lagged values of the variables as predictors. Thus for year 2002 CMF observations, we included lagged values from plants in the 2001 ASM. Using this sample, we estimate a logit model of missingness for each 5-digit NAICS:

$$\begin{aligned}\Pr(I(CM_{ijt}) = 1) = \Lambda(&\gamma_{j0} + \gamma_{j1} \ln X_{1ijt} + ... + \gamma_{jk} \ln X_{kijt} \\ &+ \gamma_{jCM} \ln CM_{ij,t-1} + \gamma_{j,k+1} \ln X_{1ijt-1} + ... + \gamma_{j,2k} \ln X_{kijt-1} + \varepsilon_{ijt})\end{aligned} \quad (3),$$

where *I()* is the indicator function: $I(CM_{ijt})=1$ if CM_{ijt} is *observed* and $I(CM_{ijt})=0$ if CM_{ijt} is missing; Λ is the cdf of the logistic distribution; and X_{1ij} through X_{kij} are all the variables in table 1 except the Total Cost of Materials. We keep the predicted probabilities from (3) for each plant. Then we select only the plant-year records in our sample with "complete" data, meaning all the variables in table 1 are observed, including CM and one-year lags of all variables. For each complete data record, we take a draw from a Bernoulli distribution with probability equal to that record's predicted probability from (3). Based on the Bernoulli trial we set CM to missing or we keep the observed CM. Basically we are "poking holes" in the CM variable in the complete data.

Having created artificially missing data from our complete records, we do two types of imputation: (i) multiple imputation using sequential regression and (ii) single imputation using a cold-deck method. For the purpose of our comparison and to avoid disclosure issues we select the 89 5-digit NAICS industries that have more than 300 complete plant-year records. To ensure some degree of comparability of plants within an industry we throw out industries with 9's in the NAICS code, since these tend to be catch-all categories (e.g., 33399="All Other General Purpose Manufacturing"). This leaves us with 66 industries and 261 industry-years (three industry-years did not have enough complete records). The appendix has a complete list of the industries in our sample.

For the multiple imputations using the sequential regression method, we used a model in the form of equation (2). IVEware allows the user to perform stepwise regressions: for each industry the program adds explanatory variables to the regression specification one at a time until the R-squared of the regression increases by less than a specified number (we choose 0.01). Thus in general the regression specification is different for different industries. We found that on average our imputations were closer to the real data if we excluded the capital expenditure variables.[5] We construct 20 imputations for each missing value and then use Rubin's (1987) combining formulas to compute confidence intervals which reflect not only the uncertainty from sampling, but also the uncertainty in our estimates due to the missing data.

For comparison, we also impute single values for the artificially missing Total Cost of Materials (CM) data using a simple ratio method, which we also refer to as a "cold-deck" method. For each artificially missing CM item, we impute $CM\_imp_{ijt}=TVS_{ijt}*(CM_{jt}/TVS_{jt})$, where TVS_{ijt} is plant *i*'s total value of shipments in year t, CM_{jt} is the mean cost of materials in industry j in year t and TVS_{jt} is the mean value of

[4] For the current preliminary paper, we report results only for the Total Cost of Materials imputations. We plan to compare our imputations to relevant alternative imputation methods for all the main ASM variables. For some variables, other imputation methods are more common than the cold-deck method. For example, for electricity expenditures (EE), the most common imputation method used by the Census Bureau seems to be plugging in the "industry average." Thus for the electricity variable we plan to compare our imputation results to results where the industry average is substituted for missing values.

[5] In principle capital expenditures should have explanatory power for the cost of material inputs and other variables. Capital expenditures differ from other variables in that capital investment is more "lumpy" (Doms and Dunne, 1998; Power, 1998; Sakellaris, 2004), with many observed 0 values. In principle, the IVEware program can handle variables with mixed discrete/continuous distributions like these capital expenditure variables. However, so far, including the capital expenditure variables has caused the imputations to be far from the real data. In future work we plan to look into this further.

shipments in industry j in year t. We use this method as our benchmark for comparison for three reasons: (i) it is simple and transparent; (ii) in most industries at some of the CM items that are flagged as imputed by the "cold-deck" method seem to be imputed this way—i.e., many cold-deck-flagged CM observations in a given industry have the same CM/TVS ratio; (iii) at the time of this writing we have not yet been able to find out how the other CM items flagged as "cold-deck" were imputed. Table 3 presents a comparison of our results using the sequential regression multiple imputation method and the single value ratio method.

	Mean	s.d.	25th Percentile	Median	75th percentile
True mean minus MI mean	**-418**	**1843**	**-458**	**-75**	**24**
True mean minus Cold-deck mean	**-732**	**3305**	**-652**	**-228**	**22**
R_MI	**0.98**	**0.02**	**0.98**	**0.99**	**1.00**
R_cold-deck	**0.95**	**0.06**	**0.93**	**0.96**	**0.98**
W_MI	**1.04**	**0.14**	**1.00**	**1.01**	**1.04**
W_cold-deck	**1.02**	**0.16**	**0.99**	**1.00**	**1.01**

Table 3. Comparison of industry-year means for sequential regression multiple imputation versus the single value ratio method for the Total Cost of Materials, thousands of dollars. All the statistics in the table are computed from industry-year means; thus each statistic represents many plant-level observations. See text for explanation of R and W statistics.

The first row of table 3 shows the across-industry distribution of the difference between the industry-year mean computed from the real (complete) data and the combined industry-year mean from our 20 implicates using the sequential regression imputation method. The first column of the first row the mean across all our industries of the difference between the industry means from the real data and the industry means from our multiply imputed data. This cell shows that *on average* the industry-year means of the Total Cost of Materials from our preferred imputations are about $418,000 higher than the industry-year means from the real (complete) data. This is perhaps not a great performance, but it is significantly better than the $718,000 upward bias in the estimates from the single imputation ratio method. The first two rows of the second column show that the standard deviation of the difference between the means from our preferred imputations versus the real data is significantly smaller than the standard deviation of the mean differences from the single value ratio method. At the median of the distribution of mean differences, our method dominates the single value ratio method: a $75,000 difference versus a $228,000 difference.

In principle all single imputation methods suffer from the fact that the confidence intervals of any estimates from singly imputed data tend to understate the amount of uncertainty in the estimate due to missing data. Therefore we would like to compare confidence intervals for estimates computed from our multiply imputed data to confidence intervals from the real data and from the singly imputed data. For each of our 20 implicates, we compute a 95% confidence interval for each of the 261 industry-year means. Then, for each industry-year we used Rubin's (1987) combining formulas to combine the confidence intervals from our 20 implicates. Then we compute the 95% confidence intervals for each industry-year mean from the real data. For each industry-year, we find the intersection of the (combined) confidence interval from our multiple imputations with the confidence interval from the real data. Define R_MI_j as the length of this intersection divided by the length of the confidence interval from the real data.
This measures how much the confidence intervals from our multiple imputations overlap with the C.I.'s from the real data. The third row of table 3 reports the mean, standard deviation, median and 25th and 75th percentiles of the distribution of these ratios. The results shows that the C.I.'s from our multiply imputed data overlap more than 90% of the C.I. from the real data across the distribution of industry-years. The fourth row of table 3 shows the distribution of the analogous ratio for the cold-deck imputations. The C.I.s from the cold-deck imputations cover less of the C.I.'s from the real data, but not much less.

We want to check that the coverage of our C.I.'s is not driven by having unreasonably large C.I.'s from the multiple imputations. For each industry-year we compute the ratio of the width of the (appropriately combined) C.I.'s from the 20 implicates to the width of the C.I. from the real data. A ratio close to 1 means that a C.I. is not "too" large. The fifth row of table 3 shows the mean, standard deviation,

median and 25th and 75th percentile of the distribution of this ratio across industry-years. The width of the C.I.'s from our multiply imputed data is typically quite close to width of the C.I. from the real data. The sixth row of the table shows the distribution of the analogous ratio for the cold-deck imputations. Comparing these to row 5, the C.I.'s from the single cold-deck imputations tend to be smaller than the C.I.'s from the multiple imputations, but not much smaller.

Tentative Conclusions and Next Steps

As emphasized throughout, the results presented here are preliminary. The sequential regression multiple imputation method produces less biased results that single value ratio imputation method. The difference between the confidence intervals is in general not as large as we expected. One reason for this may be our sample selection criteria. To avoid disclosure issues we chose industries with many complete records. The vast majority of these industries have fewer than 10% missing values in the Total Cost of Materials variables, and, by construction all of the other variables in table 1 are observed in our sample. Tables 1 and 2 show that for the typical manufacturing industry far more than 10% of its data items are missing. When the rates of missingness are higher, one would expect the confidence intervals from single imputation to be much smaller than the true data intervals, and the correctly combined confidence intervals from multiply imputed data to better reflect the uncertainty.

The results presented here only include imputations for missing values in one variable, the Total Cost of Materials. In future work we plan to apply the sequential regression method for imputations of all the variables in tables 1 and 2, and possibly other variables, and to use these imputed data to compute plant-level total factor productivity (tfp). Plant-level tfp is often estimated from something like the following equation:[6]

$$\ln tfp_{ijt} = \ln TVS_{ijt} - (\alpha_{0j} + \alpha_{kj} \ln K_{ijt} + \alpha_{lj} \ln L_{ijt} + \alpha_{mj} \ln CM_{ijt}) \quad ,$$

where K is a measure of the plant's capital stock (usually constructed from capital expenditures and assets), L is a measure of labor inputs (constructed from plant hours, production workers' wages, and total salaries and wages), and the other variables and indices are as described above. The cold-deck ratio method described in the main text forces the ratio of materials to shipments to be the same for all imputed observations in the same industry and year. Thus one might expect the ratio method to understate the amount of dispersion in plant-level tfp. Dispersion in within-industry plant-level tfp is an important feature of the U.S. manufacturing data (Abraham and White, 2006; Collard-Wexler, 2007; others). Since the sequential regression imputation method conditions on all the observations, it can potentially capture more of the dispersion of tfp seen in the observed data.

While we have focused on the implications of missing and imputed data for researchers using the microdata via the Census Research Data Centers, this research may also have direct implications for Census Bureau programs. While we reiterate that these results are preliminary, the first two rows of table 3 seem to indicate that sequential regression multiple imputation estimates are less biased than estimates from data imputed use the ratio method, even for simple industry-level means. We hope to investigate these implications further when and if we find out exactly how the Census Bureau does "cold-deck" imputations in these data.

[6] Typically, the dollar-valued variables would also be deflated so that the measures are in real (constant dollar) terms; other variables such as energy inputs or two types of labor or capital inputs might also be added to the specification, and the coefficients might be allowed to vary over time; or proxy variables might be used as in Olley and Pakes (1996) or Levinsohn and Petrin (2003). In all of these cases, the principle is the same: sequential regression allows for more flexibility to capture the assumed relationship between the variables than simple ratio methods do.

Appendix. Table A.1 lists the names and 5-digit NAICS codes of the 66 industries used to compute the statistics in Table 3 (in descending order of number of complete observations).

NAICS code	Name	NAICS code	Name
32311	Printing	32733	Concrete Pipe, Brick, and Block Mfg
32221	Paperboard Container Mfg	33361	Engine, Turbine, and Power Transmission Equipment Mfg
33351	Metalworking Machinery Mfg	33151	Ferrous Metal Foundries
32732	Ready-Mix Concrete	33641	Aerospace Product and Part Mfg
33232	Ornamental and Architectural Metal Products	32612	Plastics Pipe, Pipe Fitting, and Ulaminated Profile Shape Mfg
33441	Semiconductor and Other Electronic Component	31311	Fiber, Yarn, and Thread Mills
31161	Animal Slaughtering and Processing	33152	Nonferrous Metal Foundries
33271	Machine Shops	32518	Other Basic Inorganic Chemical Mfg
33281	Coating, Engraving, Heat Treating, and Allied Activities	32512	Industrial Gas Mfg
33231	Plate Work and Fabricated Structural Product Mfg	32312	Support Activities for Printing
33211	Forging and Stamping	32551	Paint and Coating Mfg
32412	Asphalt Paving, Roofing, and Saturated Materials	31142	Fruit and Vegetable Canning, Pickling, and Drying
32121	Veneer, Plywood, and Engineered Wood Product Mfg	32616	Plastics Bottle Mfg
32111	Sawmills and Wood Preservation	33243	Metal Can, Box, and Other Metal Container Mfg
32191	Millwork	33131	Aluminum Production and Processing
33721	Office Furniture (Including Fixtures) Mfg	32621	Tire Manufacturing
31151	Dairy Product (Except Frozen) Mfg	32212	Paper Mills
33272	Turned Product and Screw, Nut and Bolt Mfg	33637	Motor Vehicle Metal Stamping
33712	Household and Institutional Furniture Mfg	32615	Urethane and Other Foam Product (Except Polystyrene) Mfg
31111	Animal Food Mfg	33111	Iron and Steel Mills
31181	Bread and Bakery Product Mfg	33221	Cutlery and Handtool Mfg
32611	Plastics Packaging Materials and Unlaminated Film	31122	Starch and Vegetable Fats and Oils Mfg
32721	Glass and Glass Product Mfg	32712	Clay Building Material and Refractories
33531	Electrical Equipment Mfg	31141	Frozen Food Mfg
33341	Ventilation, Heating, A/C Mfg	33251	Hardware Mfg
32521	Resin and Synthetic Rubber Mfg	33291	Metal Valve Mfg
33621	Moter Vehicle Body and Trailer Mfg	32213	Paperboard Mills
33711	Wood Kitchen Cabinet and Countertop Mfg	31321	Broadwoven Fabric Mills
33261	Spring and Wire Product Mfg	31121	Flour Milling and Malt Mfg
31211	Soft Drink and Ice Mfg	32223	Stationery Product Mfg
32541	Pharmaceutical and Medicine Mfg	32561	Soap and Cleaning Compound Mfg
33331	Commercial and Service Industry Machinery Mfg	33661	Ship and Boat Building
32222	Paper Bag and Coated and Treated Paper	33311	Agricultural Implement Mfg

References

Abraham, Arpad and T. Kirk White. 2006. The Dynamics of Plant-level Productivity in U.S. Manufacturing. Center for Economic Studies Working Paper. CES-WP-06-20.

Collard-Wexler, Allan. 2007. Productivity Dispersion and Plant Selection in the Ready-Mix Concrete Industry. Working Paper.

Doms, Mark and Tomothy Dunne. 1998. Capital Adjustment Patterns in Manufacturing Plants. *Review of Economic Dynamics* 1, 409-429.

Dunne, Timothy. 1998. CES Data Issues Memorandum: 98-1. Center for Economic Studies Data Issues Memo.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2005. Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? Center for Economic Studies Working Paper. CES-WP-05-11.

Levinsohn, James and Amil Petrin. 2003. Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies*. 70(2). April. 317-342.

Little, R. J. A. and Rubin, D. B. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Olley, Stephen and Ariel Pakes. 1996. The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*. November.

Power, Laura. 1998. The Missing Link: Technology, Investment, and Productivity. *Review of Economics and Statistics*. May. Vol. 80, No. 2, pages 300-313.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*. Vol 21. No. 1.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sakellaris, Plutarchos. 2004. Patterns of Plant Adjustment. *Journal of Monetary Economics*. Vol. 51. pp. 425-450.