

Appendix C.

Sampling and Estimation Methodologies

The estimates in this report are based on two stratified simple random samples. The ACE-1 sample consists of 44,108 companies with paid employees (determined by the presence of payroll) in 1998. The ACE-2 sample consists of 15,000 businesses without employees. The two sample populations received different survey forms (see Appendix D for an example of each survey form).

The scope of the survey was defined to include all private, nonfarm, domestic companies. Major exclusions from the frame were government-owned operations (including the U.S. Postal Service), foreign-owned operations of domestic companies, establishments located in U.S. Territories, establishments engaged in agricultural production (not agricultural services), and private households.

The 1999 Standard Statistical Establishment List (SSEL) was used to develop the 1999 ACE-1 sample frame. The SSEL is the U.S. Census Bureau's establishment-based database. The database contains records for each physical business entity with payroll located in the United States, including company ownership information and prior-year administrative data. In creating the ACE-1 frame, establishment data in the SSEL file were consolidated to create company-level records. Employment and payroll information was maintained for each six-digit North American Industry Classification System¹ (NAICS) industry in which the company had activity. Next, payroll data for each company-level record were run through an algorithm to assign the company, first to an industry sector (i.e., manufacturing, construction, etc.), then to a subsector (three-digit NAICS code), then to an industry group (four-digit NAICS code), then to an industry (five-digit NAICS code), and finally to an ACES industry code based on the industry. The resulting sample frame contained slightly more than 5.5 million companies.

The 1999 ACE-1 sampling frame consists of a certainty portion and a noncertainty portion. The 15,838 companies with 500 or more employees were selected with certainty. The remaining companies with 1 to 499 employees were then grouped into 133 industry categories. Each industry was then further divided into four strata. Since capital expenditures data were not available on the sampling

frame, prior-year payroll was used as the stratification variable. The stratification methodology resulted in minimizing the sample size subject to a desired level of reliability for each industry. The expected relative standard errors (RSEs) ranged from 1 to 3 percent.

The ACE-2 sample frame was selected from four categories of small businesses.

- Companies with no payroll and no employees on March 12 in the prior year, but with characteristics indicating possible employment during the survey period.
- Companies which had received an Employer Identification (EI) number within the last 2 years, but for which no payroll, employment, or receipts data have yet been received.
- Nonemployer corporations and partnerships.
- Nonemployer sole proprietorships with sales or receipts of \$1,000 or more.

Each of these four categories was treated as a separate stratum. The source of the first two categories of businesses was the 1999 SSEL; the source of the second two categories was the 1998 Nonemployer Database. Companies within each stratum were selected using a simple random sample. From a universe of about 17.1 million businesses, approximately 15,000 businesses were selected

ESTIMATION

Each company selected for the survey has a sample weight which is the inverse of its probability of selection. All sampled companies within the same stratum and industry grouping have the same weight. Weights were increased to adjust for nonresponse. The coverage rate for all companies was 89.3 percent. The coverage rate is calculated by multiplying 100 by the ratio of the capital expenditures of all reporting companies weighted by the original sample weights, to the capital expenditures of all reporting companies weighted by the adjusted-for-nonresponse sample weights. Weight adjustment, publication estimation, and (RSE) estimation are described in the following subsections.

Weight Adjustment

For estimation purposes, each company was placed into 1 of 4 response-related categories:

1. Respondents.

¹North American Industry Classification System (NAICS) – United States, 1997. For sale by National Technical Information Service (NTIS), Springfield, VA 22161. Call NTIS at 1-800-553-6847.

2. Nonrespondents.
3. Not in business.
4. Known duplicates.

A company was considered a respondent or nonrespondent based on whether the company provided sufficient data in items 1 or 2 of the ACE-1 survey form for the ACE-1 segment or item 1 of the ACE-2 survey form for the ACE-2 segment. Companies that went out of business prior to 1999 and duplicates were dropped from the survey. Companies that went out of business during the survey year were kept in the sample and efforts were made to collect data for the period the company was active.

ACE-1 segment. The following discussion assumes 665 strata (strata designation $h = 1, 2, \dots, 665$) which are based on 133 industries, each containing five strata (including the certainty stratum).

The original stratum weights (W_h) were adjusted to compensate for nonresponse. The adjusted weight is computed as follows:

$$W_{h(\text{adj})} = W_h * \frac{(P_{hr} + P_{hn})}{(P_{hr})}$$

where,

$W_{h(\text{adj})}$ is the adjusted stratum weight of the h^{th} stratum

$W_h = \frac{N_h}{n_h}$ is the original stratum weight of the h^{th} stratum

N_h is the population size of the h^{th} stratum

n_h is the sample size of the h^{th} stratum

P_{hr} is the sum of total company payroll for respondent companies in stratum h

P_{hn} is the sum of total company payroll for nonrespondent companies in stratum h

ACE-2 segment. The ACE-2 segment initially was stratified into four strata based on the four small business categories mentioned above. The stratum consisting of “companies with no payroll and no employees on March 12 in the prior year, but with characteristics indicating possible employment during the survey period” was poststratified into two strata. The stratum “companies which had received an Employer Identification (EI) number within the last 2 years, but for which no payroll, employment, or receipts data have yet been received” was poststratified into three strata. In both instances, the poststratification was based on updated administrative-record data that were not available at the time the sample frames were created. This method resulted in seven strata (strata designation $h = 1, 2, \dots, 7$). The stratum population sizes, sample sizes, response counts, and stratum weights for

the five strata resulting from the poststratification were modified accordingly. For these five strata, the following formulas use these modified sizes and weights; for the remaining two strata, the formulas use the original stratum sizes and weights.

The stratum weights (W_h) were adjusted to compensate for nonresponse. The adjusted weight is computed as follows:

$$W_{h(\text{adj})} = W_h \left(\frac{n_h}{r_h} \right) = \frac{N_h}{r_h}$$

where,

$W_{h(\text{adj})}$ is the adjusted stratum weight of the h^{th} stratum

$W_h = \frac{N_h}{n_h}$ is the stratum weight of the h^{th} stratum

N_h is the population size of the h^{th} stratum

n_h is the sample size of the h^{th} stratum

r_h is the number of respondents in the h^{th} stratum

Note: A statistical procedure was used in reweighting extreme outliers to minimize the mean square error of the estimates. Mean square error accounts for both sampling variability and bias.

Publication Estimation

Publication cell estimates were computed by obtaining a weighted sum of reported values for companies treated as respondents. For those strata undergoing nonresponse adjustment, the estimates for X_j are biased, since this method assumes that nonresponse is not a purely random event. No attempt was made to estimate the magnitude of this bias.

ACE-1 segment. The ACE-1 estimates were derived as follows. Each estimated cell total, \hat{X}_j , is of the form

$$\hat{X}_j = \sum_{h=1}^{665} \sum_{i \in h} = (W_{h(\text{adj})} * X_{(j),i,h})$$

where,

$W_{h(\text{adj})}$ is the adjusted weight of the h^{th} stratum

$X_{(j),i,h}$ is the value attributed to the i^{th} company of stratum h , where j is the publication cell of interest.

Note: Although a company was assigned to and sampled in one ACES industry, it could report expenditures in multiple ACES industries. When this occurred, the reported data for all industries were inflated by the weight in the sample industry.

ACE-2 segment. The ACE-2 estimates were derived as follows:

$$\hat{X}_j = \sum_{h=1}^7 \sum_{i \in h} (W_{h(\text{adj})} * X_{(j),i,h})$$

where,

$W_{h(\text{adj})}$ is the adjusted weight of the h^{th} stratum

$X_{(j),i,h}$ is the value attributed to the i^{th} company in stratum h , where j is the publication cell of interest (note, since no industry level estimates are derived for ACE-2 companies, this j will always represent a national-level cell estimate).

Relative Standard Error Estimation

The Relative Standard Error (RSE) is the Standard Error (SE, and denoted by $\hat{\sigma}$ in the formulas) divided by the estimate. It provides a measure of the variation of the data relative to the estimate being made.

The SE is the square root of the variance of the estimated cell total. To estimate the variance, it is necessary to estimate the variance contribution of each of the individual noncertainty strata. For the ACE-1 estimates, there are $h=532$ individual noncertainty strata; for the ACE-2 estimates, there are $h=7$ individual strata. For the combined ACE-1 and ACE-2 national-level estimates presented in Table C-1a, the variance is estimated by summing the corresponding ACE-1 and ACE-2 variance estimates. The variance is estimated by:

$$\hat{\sigma}^2(\hat{X}_j) = \sum_h (N_h * (W_{h(\text{adj})} - 1) * s_{(j),h}^2)$$

where, N_h and $W_{h(\text{adj})}$ are as defined above, and

$$s_{(j),h}^2 = \left(\sum_{i \in h} \frac{X_{(j),i,h}^2}{(r_h - 1)} \right) - \left(\frac{(\sum_{i \in h} X_{(j),i,h})^2}{r_h * (r_h - 1)} \right)$$

where,

$X_{(j),i,h}$ is as defined above

r_h is the number of respondents in stratum h

Finally, the relative standard error of the estimated total, \hat{X}_j , the value appearing in the tables (presented as percentages) is computed as

$$RSE(\hat{X}_j) = \left(\frac{\hat{\sigma}(\hat{X}_j)}{\hat{X}_j} \right) * 100$$

RELIABILITY OF THE ESTIMATES

The data shown in this report are estimated from a sample and will differ from the data which would have been obtained from a complete census. Two types of possible errors are associated with estimates based on data

from sample surveys: sampling errors and nonsampling errors. The accuracy of a survey result depends not only on the sampling errors and nonsampling errors measured but also on the nonsampling errors not explicitly measured. For particular estimates, the total error may considerably exceed the measured errors.

Sampling Variability

The sample used in this survey is one of many possible samples that could have been selected using the sampling methodology described earlier. Each of these possible samples would likely yield different results. The RSE is a measure of the variability among the estimates from these possible samples. The RSE accounts for sampling variability but does not account for nonsampling error or systematic biases in the data. Bias is the difference, averaged over all possible samples of the same design and size, between the estimate and the true value being estimated.

The RSEs presented in the tables can be used to derive the SE of the estimate. The SE can be used to derive interval estimates with prescribed levels of confidence that the interval includes the average results of all samples:

- intervals defined by one SE above and below the sample estimate will contain the true value about 68 percent of the time,
- intervals defined by 1.6 SE above and below the sample estimate will contain the true value about 90 percent of the time,
- intervals defined by two SEs above and below the sample estimate will contain the true value about 95 percent of the time.

The SE of the estimate can be calculated by multiplying the RSE presented in the tables by the corresponding estimate. Note that RSEs in this publication are in percentage form. They must be divided by 100 before being multiplied by the corresponding estimate. For example, using data from Tables 4a and 4b, the SE for total nondurable manufacturing capital expenditures would be calculated as follows:

$$\hat{\sigma}(\hat{X}_j) = \left[\frac{RSE(\hat{X}_j)}{100} \right] * X_j = \left(\frac{2.3}{100} \right) * \$79,845 \text{ million} = \$1,836$$

The 90-percent confidence interval can be constructed by multiplying 1.6 by the SE, adding this value to the estimate to create the upper bound, and subtracting it from the estimate to create the lower bound.

$$\hat{X}_j \pm [1.6 * \hat{\sigma}(\hat{X}_j)]$$

Using data from Table 4a, for nondurable manufacturing capital expenditures, a 90% confidence interval would be calculated as:

$$\$79,845 \text{ million} \pm 1.6(\$1,836) = \$79,845 \pm \$2,938 \text{ million}$$

Nonsampling Error

All surveys and censuses are subject to nonsampling errors. Nonsampling errors can be attributed to many sources: inability to obtain information about all companies in the sample; inability or unwillingness on the part of respondents to provide correct information; response errors; definition difficulties; differences in the interpretation of questions; mistakes in recording or coding the data; and other errors of collection, response, coverage, and estimation for nonresponse.

Explicit measures of the effects of these nonsampling errors are not available. However, to minimize nonsampling error, all reports were reviewed for reasonableness and consistency, and every effort was made to achieve accurate response from all survey participants.

Coverage errors may have a significant effect on the accuracy of estimates for this survey. The SSEL, which forms the basis of our survey universe frame, may not contain all businesses. Also, businesses that are contained in the SSEL may have their payroll misreported.

1998 RESTATEMENT

The 1998 estimates presented in this report are a restatement of the 1998 SIC-based estimates. The 1998 estimates were restated to account for the following:

- Revisions to the 1998 SIC-based estimates
- Restating of the 1998 SIC-based estimates on a NAICS basis
- Change in the 1999 ACE-1 definition

The revisions made to the SIC-based estimates reflect a downward revision of \$2.7 billion. These revisions were due to corrections in the 1998 SIC-based data. After the revisions were made to the SIC-based estimates, the reported SIC-based codes were recoded to NAICS-based industry codes. The recoding process is described below. Once the data were recoded into NAICS-based industry codes, new estimates and variances were derived using the methodology previously described for the 1999 estimates. The new estimate of capital expenditures for companies with employees was then adjusted upward by approximately \$20 billion to account for the change in the ACE-1 sample frame definition. Details of this adjustment are described below.

Recoding to NAICS-Based Industry Codes

Reported Standard Industrial Classification (SIC) industries for 1998 were recoded to the North American Industry Classification System (NAICS) in the following manner:

1. Single location companies were recoded to a NAICS-based industry using industry classification information from the 1997 Economic Census. This process accounted for approximately 20.1 percent of the total restated estimate on a NAICS basis.

2. Multiple location companies that reported a single SIC-based industry in 1998 and a single compatible NAICS-based industry in 1999 were recoded to this 1999 NAICS-based industry for 1998 restating purposes. This process accounted for approximately 19.4 percent of the total restated estimate on a NAICS basis.
3. For multiple location companies not meeting the requirements of Step 2, every location of a sampled company was assigned a NAICS-based industry using a combination of information from the 1997 Economic Census and 1999 SSEL. Payroll for these locations was used to assign a NAICS-based industry code to each SIC-based industry with capital expenditures in 1998. This process accounted for approximately 60.5 percent of the total restated estimate on a NAICS basis.

Change in the ACE-1 Sampling Frame Definition

The 1998 ACE-1 sampling frame consisted of companies with at least one paid employee on March 12. Companies with payroll but no employees on March 12 were in the 1998 ACE-2 frame. In 1999, these companies were moved to the ACE-1 frame. In order to compare the 1998 estimates with the 1999 estimates, the 1998 data were adjusted upward by approximately \$20 billion (the portion of the 1998 ACE-2 estimate represented by companies with payroll and no employment). NAICS-based industry level estimates were computed by using the distribution of similar companies in the 1999 ACES sample. For example, if 20 percent of the 1999 estimate of new structures for companies with payroll and no employment was in coal mining, then 20 percent of the 1998 new structure's estimate was allocated to coal mining. The final restated estimate is as follows:

$$X_{ik} = X'_{ik} + p_{ik} \cdot X_k$$

where,

- X'_{ik} initial NAICS-based estimate in industry I and item k (i.e., new structures, new equipment...)
- p_{ik} percent of the 1999 item k estimate in NAICS industry I
- X_k initial NAICS-based estimate for item k

The final variance is estimated by:

$$\hat{\sigma}^2(X_{ik}) = \hat{\sigma}^2(X'_{ik}) + p_{ik}^2 \hat{\sigma}^2(X_k)$$

where:

- $\hat{\sigma}^2(X_j)$ variance of the initial NAICS-based estimate
- p_{ik} percent of 1999 item k estimate in industry I