Estimating Technology Adoption and Aggregate Volumes from U.S. Payments Surveys in the Presence of Complex Item Nonresponse

Geoffrey R. Gerdes* May X. Liu*

Abstract

In recent years, the Federal Reserve has surveyed banks to collect information on the number and value of check payments, and to track the adoption of new electronic check clearing methods that replace traditional paper-based methods. The data requested from each respondent consists of a hierarchy of nested totals and subtotals. The ability of banks to report data to this voluntary survey tends to vary, however, which leads to a complex pattern of item nonresponse. In light of this, independent estimation of aggregates by ratio estimator using only the reported data created violations of adding-up and other logical constraints among survey items. To overcome this problem, we investigated various item imputation methods, including multiple imputation methods. But these methods tended to overestimate the proportions of banks in the population that had adopted the new clearing technologies. Accordingly, we adjusted the methods to account for that bias.

Key Words: Item nonresponse, imputation, ratio estimator

1. Introduction

The Federal Reserve has surveyed depository institutions (banks) to estimate national aggregate payment volumes for checks, debit cards, the automated clearinghouse, and cash withdrawal volumes from ATMs every three years since 2000. In this paper we look only at the check portion of the survey, and focus on a method for obtaining national estimates of 1) the volumes of checks that are processed with different methods and flow through different channels, and 2) the proportion of banks that use the different methods.

Because of the availability in the marketplace of innovative new electronic image-based methods of interbank check processing, the 2007 survey asked banks to report these volumes along with volumes of traditional processing methods. For a particular bank, the sum of these methods should add up to total interbank check volume. Over half the sampled banks responded to this voluntary survey but extensive item nonresponse resulted in a reported proportion of some items below one-third. For example, almost all respondents were able to provide the aggregate total number and value of checks they received and paid, but some faced difficulties in obtaining the detailed channel and processing method volumes for those checks.

The quality of information we can use to estimate a national total drops, of course, as the number of reported items drops. With item nonresponse, estimating each item independently by using only the reported data leads to violations of adding-up constraints. One approach that would retain the adding-up property would be to drop incomplete responses. Clearly we want to retain as many responses as possible, however, to use all the available information, even if it is partial information. To maximize the use of all responses for estimation while ensuring the adding-up of estimated totals, we employed imputation methods. The imputation methods we used take advantage of the information from correlations between variables and logical constraints.

*Federal Reserve Board of Governors, Washington, DC 20551. Please send correspondence to May.X.Liu@frb.gov. Opinions are the authors and do not necessarily reflect those of the Board of Governors or its staff.

There are various ways that banks may receive and pay a check. When a bank reports that it has received and paid a non-zero volume of checks through the one of the new methods, we say that the bank has *adopted* the new methods. We can estimate the proportion of banks that adopted from the set of banks that completely responded to the item that represents the new methods (that item is called Electronic). We can also estimate the proportion of banks that adopted the new methods by using an imputed dataset, but care must be taken because standard imputation methods may not properly handle both a volume imputation and an adoption imputation.

In fact, compared to the estimates using only the complete responses, the one-stage volume imputation method we used for the 2007 estimates would have led to an overestimate of the adoption proportion, even while providing a reasonable estimate for total volumes by using a ratio estimator. This is because the original one-stage method of imputation was not conditional on adoption, and may be thought of as a weighted sum of a volume imputation assuming the bank had adopted and a volume imputation assuming the bank had not adopted. Thus, the one-stage imputed volume is, of course, smaller than it would be for the adopter, and greater than it would be for the nonadopter (who would have a zero volume). Further, the one-stage multiple imputation approach does not account for any error in the volume estimates that arises from ignoring the adoption decision.

To create more realistic imputations, we constructed an adoption indicator and incorporated it into a two-stage nested imputation framework to jointly estimate adoption along with volumes. We found that the two-stage method succeeds in producing imputed datasets that can be used to estimate both aggregate volume and adoption. In addition, we were able to introduce previously unexploited data correlated with the adoption decision into the first stage of imputation, which helps to improve the estimates.

2. Survey Design

The population for our 2007 survey comprised over 13,000 insured banks. For the sampling frame we treated affiliated banks as a single entity, and included all types of banks with insured transaction deposits over $1 million, including commercial banks, savings institutions, and credit unions. All these banks report a variety of balance sheet and income statement information on a periodic, usually quarterly basis. Two items of primary interest from these "call reports" are checkable deposits, which we call CHKD, and money market deposits, which we call MMDA.

Both CHKD and MMDA are the main source of funds used to pay checks, and are highly correlated with volumes of payor bank checks across banks. Most banking activity is concentrated in the largest banks, and so distributions of bank size, as measured by assets or liabilities (including most forms of deposits) are skewed to the right. To take advantage of the correlations and to account for the skewness we stratified the bank population by size and type and concentrated sampling probabilities into the strata with larger banks. (See [3] for more information.)

Bank customers deposit checks into their accounts and also write checks that draw on the funds that are available in their accounts. Banks collect funds associated with a deposited check from the bank on which it is drawn by delivering or "presenting" the check. The bank the check is drawn on is called the payor bank, and the checks it pays are called paid checks. The bank that receives the funds from the paying bank is called the collecting bank. Checks that banks present to each other are generally settled on the books of the Federal Reserve Banks or a correspondent bank.

Our survey respondents were asked to report their bank's paid checks. They were also asked to report the methods or channels through which these checks were received and

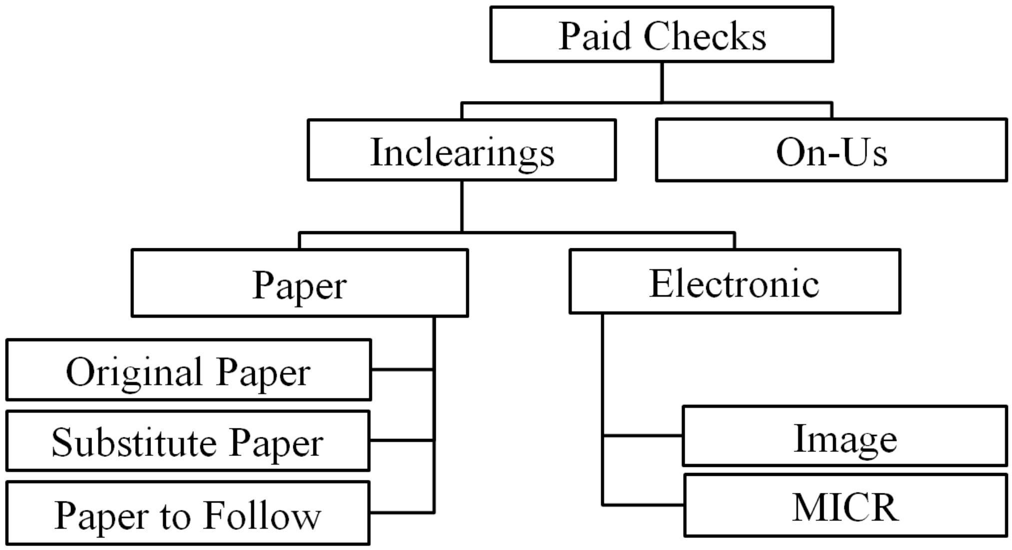

Figure 1: The methods through which checks were received and paid have a hierarchical structure.

paid. These methods were categorized on the questionnaire so that the paid check data form a hierarchy of nested totals and subtotals, as shown in Figure 1.

Paid checks can be divided into the categories of inclearings and "on us" Checks. In banking jargon, inclearings are checks that the payor bank receives from the collecting bank either directly or through one or more intermediaries or agents, such as a Federal Reserve Bank or a private clearinghouse. If the account of an individual that deposited a check, called the "payee," and the account of the individual that wrote the check, called the "payor," are at the same bank then the check is "on us" to the bank, and does not pass through the interbank clearing system.

Traditional industry practice and regulations based on state laws required the collecting bank to physically deliver the original paper check to the payor bank. If the payor bank was nearby, a collecting bank might send the check by courier directly to the payor bank or exchange checks with other banks at a local clearinghouse. If the payor bank was in another city, then an intermediary was usually used to transport the check by truck or by airplane. Checks that are received using these traditional methods are reported under Original Paper in the questionnaire. A small number of large banks had agreements to exchange electronic information about the checks drawn on each other using a method we call "paper to follow," but would generally deliver the original paper checks at a slower pace. Finally, a small number of banks had agreements to present checks by exchanging computer files containing limited data (account and routing number, dollar amount, check number, etc.) without delivering the paper checks. This method of receiving check information is called MICR, named after the line of account, routing number, and other information encoded with magnetic ink on the bottom of checks.

With changes in regulations governing check processing that resulted from the Check 21 law, collecting banks may now remove original paper checks from the clearing process and process them electronically without entering into agreements with payor banks [1]. The removal of the original paper check from the processing flow is called check truncation. Payor banks, however, may still require paper. In that case, the collecting bank must present a Substitute Check. If the payor bank agrees to receive electronic images, called Image

Pattern	Paid Checks	Inclearings	On-Us	Paper	Electronic	Image	MICR	Count
1	0	0	0	0	0	0	0	478
2	0	0	0	0	0	1	1	7
3	0	0	0	1	1	0	1	2
4	0	0	0	1	1	1	0	10
5	0	0	0	1	1	1	1	47
6	0	1	1	1	0	0	0	9
7	0	1	1	1	1	1	0	4
8	0	1	1	1	1	1	1	218

Table 1: The patterns of item nonresponse among the methods through which checks are received and paid. For simplicity, subcategories of Paper are not displayed. A “1” indicates that both month’s number figures were missing, while a “0” indicates that at least one of the two month’s figures was reported. The pattern is close to monotonic, meaning that if subtotals are reported, then their associated totals are generally also reported.

Exchange on the survey, then no paper need be involved in the collection process.

The costs and benefits of adopting electronic check image processing vary, are changing rapidly, and can be influenced by a variety of factors. Each bank chooses a time to adopt on the basis of the expected future costs and benefits of adopting at that time. It is easier to adopt electronic image deposit technology than it is to adopt electronic image receipt technology, and banks typically adopt image deposit before image receipt. For collecting banks who have adopted the image deposit technology, they may completely stop sending paper checks, or continue sending paper to some banks while sending electronic images to other banks. Payor banks that adopt electronic receipt technology will usually experience a transition period when both paper and electronic images are received. In the long run, all banks that process checks most likely will adopt electronic image processing methods.

As we said above, our survey collected information about paid checks, and we focus on Inclearings, which are interbank checks that are received and paid by a bank in its role as payor bank. In our questionnaire, Inclearings are divided into Paper checks–which should be the sum of Original Paper, Substitute, and Electronic Presentment (in banking jargon)–and Electronic (checks presented electronically)–which should be the sum of Image Exchange and MICR. The questionnaire also provided the ability for repondents to report totalsin cases where they could not report data on one or more subtotals.

For each of the items in the hierarchy, respondents were asked to provide four figures (number and value for March and April of 2007). Thus, in addition to the logical relationships implied by the hierarchy, each item had logical relationships within the four figures. For example, number-value pairs should not have a zero amount accompanied by a nonzero amount.

3. Survey Data

Initial responses contained logical errors as well as missing data. To obtain a dataset for analysis, responses were examined, and for any violations of identified logical constraints, respondents were contacted and, if appropriate, data edits were made. In most cases where logical inconsistencies could not be resolved, figures were considered missing. For the check estimates, we also dropped responses for which the Paid Check item had all four figures missing. In this paper we study the 775 commercial banks responses that remained after this data editing process.

The data display a complex pattern of item nonresponse. It is generally easier for the banks to report items higher up in the hierarchy. To help the reader understand item nonresponse in the check section of the survey, we provide the patterns for the number figures, where an item is considered missing only if both month's number figures are missing (Table 1). In the table, variables to the left are generally either at the same level or higher in the hierarchy. For example, Inclearings and On-Us are more completely reported than Paper and Electronic, and Electronic is more completely reported than Image and MICR. Note that for presentation purposes, the three subcategories of Paper are not displayed.

The pattern of missing check data is close to monotone, and the variables on the left tend to be more observed than those on the right. For example, Paid Checks are more observed than Inclearings or On-Us. Similarly, Electronic is more observed than Image or MICR. This pattern is because the "top-line" totals tend to be easier for banks to report than the subcategories below.

4. Imputation and Estimation Methods

4.1 One-Stage Imputation

Because of the close-to-monotone missing item structure, we impute items in hierarchical fashion, from top to bottom as in Figure 1 or left to right as in Table 1. Logical relationships tend to reduce the number of individual imputations that are necessary. When an imputation is made for Inclearings, for example, an imputation for On-Us is implied because Inclearings and On-Us must add up to Total Paid Checks. Within an item, the number figures tend to be more observed than the value figures, and numbers are thus imputed before values.

For imputations used in the 2007 estimates reported in previous Federal Reserve publications, such as [2], we used an iterative EM-algorithm-based linear approach designed to obtain maximum-likelihood estimates of model parameters, under the assumption that the the missing data mechanism is ignorable. That approach also produces the best linear predictors (imputations) of the missing values [4]. From these estimated models we generated imputations for multiply imputed datasets which included a random component that accounts for the unexplained portion of the models. These multiply imputed datasets were then used to estimate national totals.

At each step in the EM algorithm we ran simple linear regressions to model the expected relationship between a given missing figure (the dependent variable) and the "closest" reported figure (the independent variable).[1] The distance from a given missing figure was determined by first searching for a reported figure within the same item. Figures of the same type (number or value) are considered closer than those of a different type. If none are of the figures are available, then search for a related total.

Although the survey did not contain a "yes/no" question about whether the electronic methods had been *adopted* by a particular bank, a zero reported volume for Electronic clearly indicated that a bank had not yet adopted the new methods. Thus, one way of estimating the proportion of banks that adopted in the population is simply to construct an indicator based on the reported data from the Electronic item. Using only the reported data to estimate adoption may ignore valuable information. To maximize the use of available information and to get better estimates for adoption we would prefer to use imputed data instead.

However, using the regression-based approach discussed above leads to imputations centered on the mean of the dependent variable, and that do not reflect the true distribution of the data. Thus, adoption is always implied by the imputations. Using reported data from one of the strata in the study to illustrate this problem, Figure 2 depicts a scatterplot of the

[1] Because of perfect and near-perfect collinearity among the variables, multiple regression is problematic and also does not add much benefit.

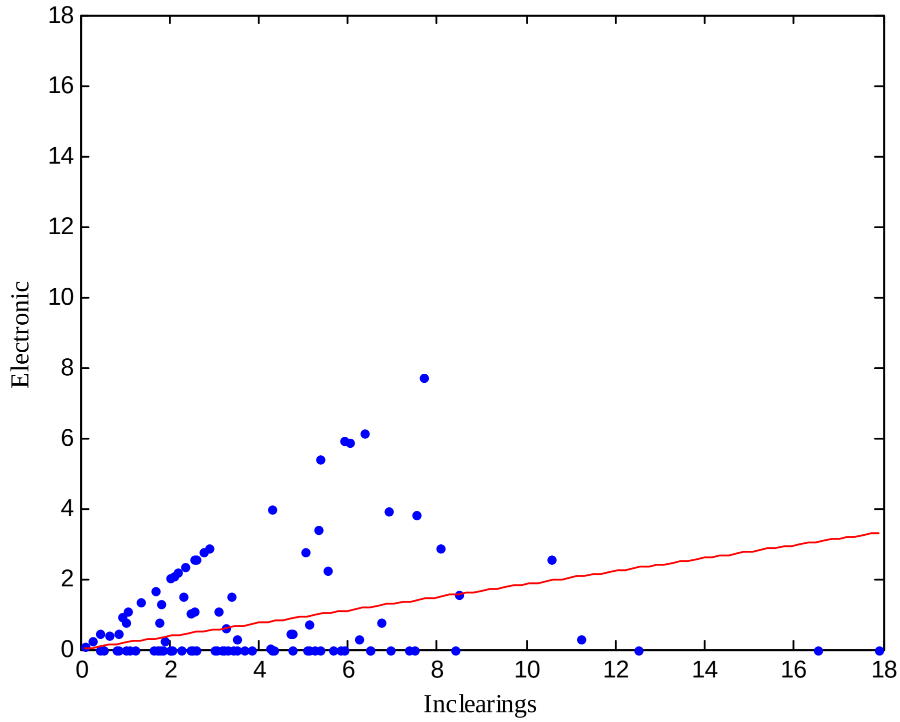

Figure 2: Scatterplot of reported Electronic against Inclearings from one of the strata in the survey, both in millions of checks per year. Inclearings is the sum of Electronic and Paper. Electronic is thus bounded by Inclearings and zero. The line, taken from a simple linear regression of Electronic on Inclearings, represents the expectation of Electronic for a given amount of Inclearings.

number of checks received and paid by electronic image (Electronic) on the y axis and total number of checks received from other banks (Inclearings) on the x axis. The figure also includes the line implied by a fitted regression of Electronic on Inclearings, which would be a typical regression used for imputation. The mean imputed amount of Electronic for a given amount of Inclearings falls on that line. Although most of the reported Electronic data appear on the boundaries of a 45 degree line where Electronic equals Inclearings or along the line where Electronic is equal to zero, the regression line is always positive.

Another way to see the problem is to examine the distribution of the proportion of the volume of Inclearings that are Electronic for only the reported data and compare it to the distribution of an imputed dataset (including imputed and reported data). Figure 3 displays a histogram of only the reported data and compares it to a histogram of the multiply imputed data using the one-stage regression imputation approach. Note that the distribution has a bimodal shape, reflecting the fact that banks typically prefer to receive all or nearly all of the checks they pay by either Paper or Electronic methods. Some banks, of course, reported significant shares of both Paper and Electronic volume. These banks are likely recent adoptees of electronic check image receipt, are in the midst of a transition period.

The histogram from the one-stage imputation method displays a left-of-center spike, where the proportions of Electronic and Inclearings lie between 20 and 30 percent. Correspondingly, the two modes that were displayed in the reported data are much smaller in the imputed data. Compared with the reported data, the imputed data from the one-stage imputations present an unrealistic representation of the percentage of banks 1) that have not adopted, 2) that are in a period of transition or have simply chosen to operate with both paper and image processing capability, and 3) that have adopted.

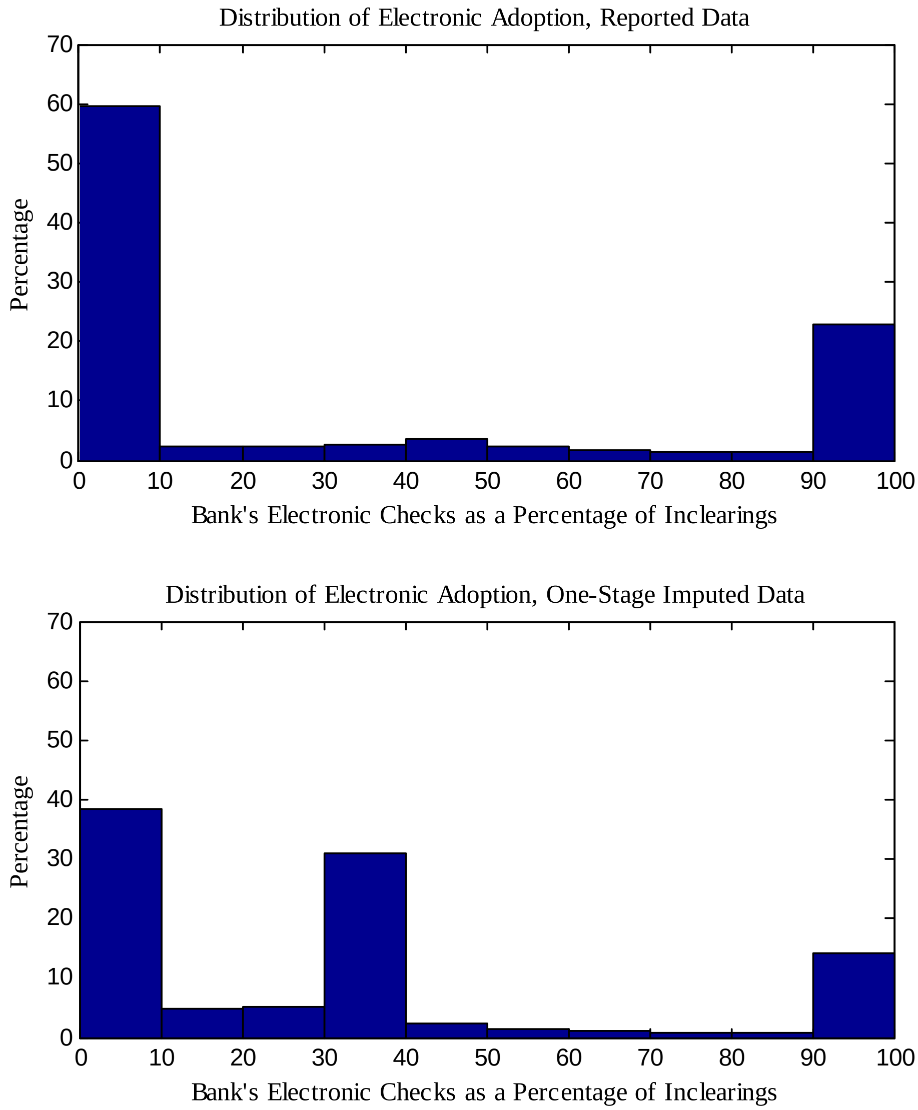

Figure 3: Comparison of the distribution of the proportion of the volume of Inclearings that are Electronic for only the reported data (top) to the distribution of imputed data using the one-stage method (bottom). The one-stage method does not produce data that appropriately reflect the true distribution.

It seems clear from these figures that the one-stage linear regressions result in blended imputations reflecting a weighted average of what volumes would be if the bank had either not adopted (zero Electronic) or had adopted (typically with Electronic equal or close to Inclearings). Thus, while a direct or one-stage imputation method could produce national volume estimates for the Electronic variable with desireable properties, it would clearly lead to overestimates of the number of banks that adopted the new electronic image processing methods.

4.2 Two-Stage Imputation

To overcome the problems with the one-stage imputation method, we devised an approach that first imputes the adoption decision, then imputes the volume. We started by constructing a binary adoption indicator based on whether a zero or positive volume for electronic image-based checks was reported. In the first-stage we used a logit model to impute a value of zero or one for the adoption indicator for banks with missing volume data for the Electronic variable, and then assigned zero volumes to the Electronic variable for those banks whose imputed adoption indicator was zero. In the second stage, for those banks whose imputed adoption indicator was equal to one, we used a linear regression approach as in the one-stage method to impute volumes. Unlike the one-stage method, however, linear regressions were restricted to the subset of banks for which the adoption indicator was equal to one.

The logit model, with the adoption indicator as the dependent variable, used a variety of population variables as independent variables that were convenient to obtain and that were found to be highly correlated with the adoption indicator. The variables we used included measures of bank size, information about whether a bank had adopted electronic check image deposit methods, and the proportion of other banks in the local market that had adopted electronic check image deposit methods.[2]

To create multiply-imputed datasets that account for the error introduced by the logit-based imputation method in the first stage, we used a bootstrapping approach for the adoption indicator. Each bootstrapped dataset was a resample of the banks for which adoption was observed using random draws with replacement. The number of random draws for each bootstrap resample was equal to the number of observed adoption indicators.

We estimated a logit model for each resample and, using the fitted parameters, we calculated the predicted probabilities for each case with a missing adoption indicator. Then we made random draws from the implied binomial distributions to impute the missing adoption indicators. Of course, an imputed zero for the indicator led to an imputation of zero for the volume of Electronic.

As noted above, in the second stage the linear regressions are run only on the observations for which the adoption indicator, either observed or imputed, is equal to one. Imputations in the second stage include a random error as in the one-stage method.

This approach solved the problems with the one-stage imputation discussed above. Figure 4 provides a histogram of the reported data and a histogram of the multiply imputed data from the two-stage method, allowing a comparison of the two distributions. Unlike the histogram for the one-stage method (shown in Figure 3) the histogram for the two-stage method is strikingly similar to the histogram of the reported data.

4.3 Estimation

4.3.1 Volume Estimates

To obtain aggregate estimates of volume for the population of commercial banks we used separate ratio estimators. We took advantage of the high correlation between the universally available size covariate checkable deposits (CHKD) and various check volumes measured in the study.

Let y_{hi} be the reported amount of the dependent variable of interest for the ith bank in stratum h and let x_{hi} be its population covariate, where $h = 1, \ldots, L$ and L is the total

[2] When a bank receives and pays a check using electronic image processing methods, it is typically coming from an electronic check image deposit from a local collecting bank. Before a payor bank adopts electronic image receipt it typically will have already adopted electronic deposit methods to collect checks.

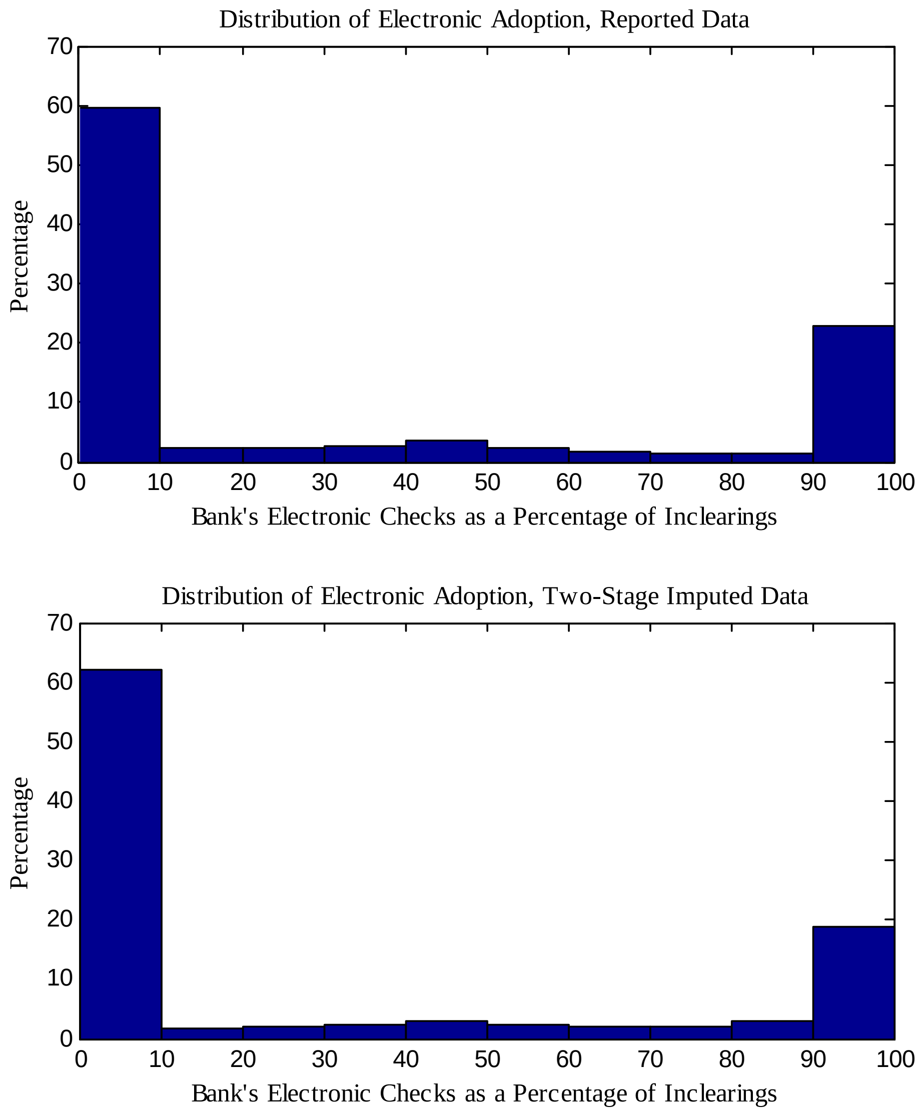

Figure 4: Comparison of the distribution of the proportion of the volume of Inclearings that are electronic for only the reported data (top) to the distribution of imputed data using the two-stage method (bottom). The two-stage method appears to reflect the true distribution.

number of strata while $n_h$is the number of respondents in stratum h. Then the ratio estimate for the population total $\hat{Y}_h$ of stratum h is given by the reported total multiplied by the ratio of the covariates in the population to the covariates from the respondents:

$$\hat{Y}_h = r_h X_h \equiv \frac{y_h}{x_h} X_h = y_h \frac{X_h}{x_h},$$

where $x_h = \sum_{i=1}^{n_h} x_{hi}$ and $y_h = \sum_{i=1}^{n_h} y_{hi}$ are the respondent total for the covariate and the dependent variable, respectively, $X_h = \sum_{i=1}^{N_h} X_{hi}$ is the population total of the covariate, and N_h is the total number of banks in the population.

The estimated standard error for $\hat{Y}_h$ is given by the following classical formula that accounts for the uncertainty arising from sampling:

$$\hat{\sigma}_{\hat{Y}_h} = \sqrt{\mathrm{var}(\hat{Y}_h)} = \left[\frac{N_h^2(1-f_h)}{n_h} s_h^2\right]^{1/2},$$

where $s_h = [\sum (y_{hi} - r_h x_{hi})^2/(n_h - 1)]^{1/2}$, $f_h = n_h/N_h$ is the sampling fraction and the factor $(1 - f_h)$ is the correction for a finite population.

4.3.2 Adoption Estimates

We are interested in obtaining an estimate of the proportion of the bank population that had adopted the technology to receive and pay checks electronically. Let

$$a_{hi} = \{1 \text{if respondent } i \text{ in stratum } h \text{ adopts, } 0 \text{ otherwise}\}.$$

Then the estimate of the proportion of banks in stratum h is

$$\hat{p}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} a_{hi}.$$

The national proportion is given by

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{n_h} N_h \hat{p}_h.$$

Standard errors for $\hat{p}_h$ are computed using a bootstrapping approach, where n_h random draws of the adoption indicator, with replacement, were made one-thousand times from the set of observed adoption indicators for stratum h. The standard error of $\hat{p}$ is then calculated by the standard deviation of the one-thousand different estimated proportions.

5. Comparison of Estimates and Conclusions

Imputation allows us to produce rectangular datasets that maintain adding-up and other logical constraints among the survey items, while also maximizing the use of information from observations with incomplete data. Using the two-stage method we described in this paper also allows us to produce imputations that can be used to simultaneously produce realistic estimates of total volumes and proportions of banks that have adopted new methods of receiving and paying checks.

Table 2 provides a comparison between the three different approaches we tried for estimating adoption proportions and volumes. In the reported case, we used only the data

Commercial Bank Estimates		Banks that Adopted %	Volumes					
			Inclearings		Electronic		Paper	
			# (mil.)	$ (bil.)	# (mil.)	$ (bil.)	# (mil.)	$ (bil.)
Reported	Est.	40.44	17,224	25,471	5,179	6,680	12,002	18,667
	(S.E.)	*(3.35)*	*(269)*	*(553)*	*(202)*	*(278)*	*(242)*	*(622)*
One-Stage	Est.	66.73	17,434	25,435	5,291	6,865	12,143	18,570
	(S.E.)	*(2.34)*	*(186)*	*(472)*	*(117)*	*(114)*	*(151)*	*(459)*
Two-Stage	Est.	37.65	17,395	25,457	5,313	6,928	12,082	18,529
	(S.E.)	*(1.82)*	*(172)*	*(492)*	*(171)*	*(232)*	*(204)*	*(542)*

Table 2: Estimates of the proportion of banks that adopted and volumes for Inclearings, Electronic, and Paper checks.

that were reported for each item. The one-stage and two-stage approaches used imputed data generated as described above.

The number of reported figures was different for each item, and therefore there was no guarantee that total volume estimates would add up. For this particular case, the difference is small on a national level: The sum of Electronic and Paper was less than Inclearings by 43 million checks, and a value of $124 billion, small proportions of the total estimates. For logical relationships between other items, however, the difference can be larger. Because adding-up constraints are imposed and because the imputed datasets are rectangular, of course, there are no violations of adding-up constraints for the estimates from the one-stage and two-stage approaches.

The point estimates of the various volumes across the different approaches are similar to each other. The standard errors, however, exhibit interesting differences. Both imputation approaches produce smaller standard errors, which are, of course, computed using the usual techniques of multiple imputation. Compared with the two-stage method and the method that used only the reported data, the one-stage method produced smaller standard errors for Electronic. However, the imputed figures in that approach do not reflect the distribution of Electronic volumes, which are conditional on the adoption decision. Also, the one-stage estimated proportion of banks that had adopted is significantly higher than the estimated proportion from the reported and two-stage approaches. This estimate from the one-stage method is not credible, because it implies that the banks that did not report would display a much greater adoption proportion compared with the reported data.

The two-stage estimated proportion of banks that adopted is much closer to the reported data estimates. This is because the two-stage method directly imputed the adoption indicator, and based the imputation on an estimated logit regression of adoption on several bank population variables that fit the data reasonably well. Taking all of the estimates together, the two-stage method achieves more precision (as measured by the estimated standard errors) than the estimates from reported data, while also obtaining volumes and proportions that are consistent with the observed data.

References

[1] Bauer, P.W. and Gerdes, G. R. (2009), "The Check is Dead! Long Live the Check! A Check 21 Update," *Economic Commentary*, June, Federal Reserve Bank of Cleveland

[2] Gerdes, G. R. (2008), "Recent Payment Trends in the United States," *Federal Reserve Bulletin*, 94, Washington, DC: Board of Governors of the Federal Reserve System.

[3] Gerdes, G. R., Liu, M. X., and Parke, D. W. (2009), "Sample Design and Estimation of Volumes and Trends in the Use of Paper Checks and Electronic Payment Methods in the

United States.” In *JSM Proceedings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 100-115.

[4] Little, R. J. A., and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, (2nd Ed.), Hoboken, NJ: John Wiley & Sons.