DEVELOPING IMPUTATION MODELS FOR THE SERVICES SECTORS PORTION OF THE ECONOMIC CENSUS

Katherine J. Thompson and Quatracia Williams[1]

U.S. Census Bureau
Room 3108-4, Washington, DC 20233/katherine.j.thompson@census.gov and quatracia.williams@census.gov

Introduction

The editing software used by the Economic Census offers a variety of imputation options, many of which employ statistical models. In prior censuses, the services sectors portion of the Economic Census has relied on industry average imputation as its primary statistical imputation model. This ratio imputation method uses weighted least squares estimates for no-intercept simple linear regression models as imputation parameters. The weighted regression method currently used compensates for heteroscedasticity (unequal error variances). The industry average imputation method used by the services sectors portion of the Economic Census assumes the following weighted regression model for each basic data item Y: $Y_{ij} = \beta_i X_{ij} + \xi_{ij}$, $\xi_{ij} \sim (0, X_{ij}\,\sigma^2)$, i.e., $\mathrm{var}(\xi_{ij}) = \sigma_{ij}^2 = X_{ij}\,\sigma^2$, where j indexes the establishments within industry i that satisfy the ratio edit $l_i \le Y_{ij}/X_{ij} \le u_i$. Thompson and Sigman (1996) and Huang (1984) demonstrate the plausibility of this model for the services sectors data. The best linear unbiased estimator (B.L.U.E) of β_i (the industry average) for this model is the weighted least square estimate with weight of $1/X_{ij}$, so that $\sum_j w_{ij} X_{ij} Y_{ij} / \sum_j w_{ij} X_{ij}^2 = \sum_j Y_{ij} / \sum_j X_{ij}$ (Draper and Smith, 1981, p.111).

Although statistically sound, the industry average imputation method is limited. First, it assumes a strictly linear relationship between the independent and dependent variables, when the actual regression line may not go through the origin. Second, it assumes that the variability in the dependent variable can be best "explained" by one variable. A multivariate regression model could have more predictive power.

This paper compares alternative methods of developing multiple regression models using data from the 1997 Economic Census. We consider three different approaches, two designed to obtain optimal estimates under conditions of heteroscedasticity and in the presence of outliers (weighted least squares regression and resistant regression) and one designed to obtain stable estimates in presence of multicollinearity (ridge regression). We discuss issues of model selection and validation in the presence of multicollinearity and heteroscedasticity. Finally, we make recommendations for developing imputation models for the services sectors portion of the Economic Census.

Current Editing and Imputation Practices

The services sectors portion of the Economic Census mails out over 400 different forms to over four million businesses. This portion of the Economic Census encompasses five trade areas: Retail Trade, Wholesale Trade, Services Industries, Transportation, Communication, and Utilities Industries (Utilities), and Finance, Insurance, and Real Estate (FIRE). All trade areas collect a core set of "basic data" items: annual payroll, 1st quarter payroll, employment, and receipts (or sales). In addition, tax-exempt Services industries collect operating expenses, and Wholesale Trade collects operating expenses and costs of purchases.

Basic data items are automatically edited using the ratio edit module of a generalized editing and imputation subsystem, called Plain Vanilla (PV), which was developed for the 1997 Economic Census. A ratio edit compares the quotient of two highly correlated items to upper and lower bounds, known as tolerances. The ratio editing software employed by the Economic Census tests all ratio edits simultaneously and requires that imputed values satisfy all of these edits (Greenberg, 1986). The software determines the minimum number of edit-failing items that must be deleted (replaced with imputed values) so that the edited questionnaire passes all edits. To maximize the efficiency of this editing software, the input items

[1]The authors thank Samson Adeshiyan, Scot Dahl, and Richard Sigman for their useful comments on earlier versions of this manuscript. This paper reports research and analysis undertaken by Census Bureau Staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties and to encourage discussion.

must all be highly correlated. From a regression model-building perspective, this introduces the issue of multicollinearity. Thompson and Adeshiyan (2003) provide more detail on the PV ratio edit methodology.

PV users specify the order in which edit-deleted items are imputed and the sequence of imputation models attempted for each item. For each edit-failing item, the program attempts each listed imputation model sequentially until obtaining a satisfactory imputation. The imputation region for each deleted item is calculated from the values of the remaining non-deleted items and their associated ratio edit bounds. When more than one item is deleted, the choice of imputation model can greatly affect the imputation region for all but the first deleted item.

Many of the available PV imputation models are deterministic, including direct substitution or logical edits (e.g., substituting an associated sum of details for a total). There are, however, several available statistical model-based methods, including regression imputation. Of these statistical methods, industry average imputation is intuitively appealing. The statistical model follows logically from the form of the ratio edit: a no-intercept regression model where the edit's numerator is the dependent variable, and the denominator is the independent variable. Moreover, industry average parameters are easy to interpret and verify, e.g., average annual payroll per employee. In this paper, we use the term "industry average" to denote simple linear no-intercept regression modeling, and "regression imputation" to refer to any other regression-based imputation model.

Research Data

We used 1997 Economic Census data for this evaluation and conducted each of the described analyses separately for each trade area. The five trade areas each classify their data slightly differently for editing and imputation purposes. For example, Wholesale Trade develops edit/imputation parameters for each type-of-operation classification within industry, whereas Services Industries uses industry-by-tax status to define edit/imputation parameters. We used the trade area's edit/imputation cells for our research, but for simplicity refer to each classification cell as an industry. We examined a total of 149 industries: 19 for Wholesale Trade; 30 for FIRE; 30 for Utilities; 30 for Services; and 40 for Retail. For the Wholesale Trade research, we used all six-digit industries within one four digit NAICS industry code for our research. For the other trade areas, we selected a random sample of industries that contained at least 40 observations whose reported data satisfied all ratio edits (ten observations per independent variable). By restricting our model-building data to regions defined by the set of all ratio edits, we reduced the incidence of outliers and guaranteed that our prediction region would correspond to the imputation region used for the 2002 Economic Census data.

Using 5-year old data may be sufficient for researching the viability of alternative regression imputation models, but we have no way of knowing how well the recommended 1997-data-models will predict 2002 data values. We are particularly concerned about production implementation of 1997-data regression models containing employment for two reasons. First, the effect of the between-census economy change could be negligible in models that contained only dollar value covariates (affecting all variables equally), but might not be negligible in models that contain both employment and dollar value covariates. Second, employment data is collected differently in 2002 from 1997: the 1997 Economic Census requested total employment (which included leased and direct employees), whereas the 2002 Economic Census collects payroll and employment data for direct and leased employees separately.

Methodology

Regression Parameters Estimation

As mentioned above, the weighted least squares regression model used to obtain industry averages assumes that the error term variances increase with the independent variable. It is difficult to theoretically extend this simple weighted regression model to multiple regression. Neter, Wasserman, and Kutner (1989, pp. 420-421) suggest grouping the model building data according to the fitted value of $\hat{Y}$, calculating the variance of the residuals for each group, and using this as each group's weight. Chen and Shao (1993) propose constructing weights based solely on each group's residuals. Our dependent variable distributions varied greatly by industry, making these approaches impossible in a generalized way. Instead, we used the following model

$$Y_{ij} = \beta_{1i} X_{1ij} + \beta_{2i} X_{2ij} + ... + \beta_{pi} X_{pij} + \xi_{ij}, \xi_{ij} \sim (0, X_{1ij} \sigma^2) \quad (1)$$

Like the industry average model, the regression weight is given by $1/X_{1ij}$. We used different weighting variables for different combinations of covariates.

Weighted least squares estimation reduces heteroscedasticity by "down-weighting" observations with large **expected** residuals. However, this approach assumes a particular data + error model, which may not be correct for all cases. For example, the assumed model may not adequately explain the presence and magnitude of outliers. Several of the current services sectors surveys as well as the manufacturing, mining, and construction sectors of the Economic Census use resistant regression (Mosteller and Tukey,1977, pp.356-361) to develop imputation parameters. This iterative procedure develops regression weights (biweights) that are smaller for observations with large residuals (generally outliers), assumes a particular data model with an unspecified error structure, and uses weighted least squares estimation to obtain coefficient estimates. When the data model contains no intercept, resistant regression yields the same coefficient estimates as the weighted least squares estimates for the model

$$Y_{ij} = \beta_{1i} X_{1ij} + \beta_{2i} X_{2ij} + ... + \beta_{pi} X_{pij} + \xi_{ij},\ \xi_{ij} \sim (0, w_{ij}^{-1}\sigma^2) \qquad (2)$$

The biweight (w_{ij}) measures the "influence" of each observation on the estimation of model coefficients. The biweight is a function of the observed value, the predicted value from the k^{th} fit, and a measure of spread for all the residuals in the k^{th} fit (S_{pk}); w_{ij} is small for large residuals. Unlike (ordinary) weighted least squares regression, resistant regression down-weights observations with large **observed** residuals. We evaluated resistant regression with two different measures of spread: the **median** of the absolute value of the residuals (proposed by Mosteller and Tukey), and the **mean** of the absolute value of the residuals (used in the current survey's programs).

Regardless of regression weight, the majority of our fitted multiple regression models exhibited heteroscedasticity. Thus, the parameter estimates from these weighted least squares regressions are still unbiased but do not have minimum variance.

Using the different weighted regression methods reduced our models' heteroscedasticity but did not address the problem of multicollinearity. We used the variance inflation factor (VIF) to measure multicollinearity of our full covariate models: a single VIF of 10 or an average VIF greater than the number of covariates indicates severe multicollinearity (Neter, Wasserman, and Kutner, 1989, pp. 408-410). The majority of our full covariate models for **each** data item had at least one VIF greater than 10.

Ridge regression is a biased estimation technique designed to reduce multicollinearity (Neter, Wasserman, and Kutner, 1989, pp. 412-418). This method introduces a biasing constant *c* into the normal regression equations for correlation-transformed variables [The correlation transformation standardizes all covariates, helping control roundoff errors and making the regression coefficients units comparable]. Fitted ridge regression models always include an intercept term (coefficients calculated from correlation-transformed data are translated to the original scale). To choose *c* in our test industries, we tried to find one value of *c* per dependent variable (per trade area) for each unique set of covariates that yielded all $VIF_p \leq 1$ and also yielded similar parameter estimates for all greater values of *c.* Ridge regression has two major drawbacks. First, the choice of the biasing constant *c* is judgmental, and there is no guarantee that the same value of *c* will work "from one application to another" (e.g., 1997 and 2002 data for the same dependent variable and covariates). The second drawback is that ridge regression sums-of-squares estimates' distributional properties are unknown and consequently cannot be used.

Evaluation Statistics

Because our data were highly multicollinear, we could not use estimates based on the weighted regression sums-of-squares to select the "best" method of parameter estimation or model selection. Moreover, there are no sums-of-squares diagnostic statistics for ridge regression. This restriction eliminates most of the "traditional" multiple regression diagnostics, such as adjusted-R^2 or F-tests. It also eliminates any form of automatic model selection procedure. Instead, we used cross-validation (Neter, Wasserman, and Kutner, 1989, pp. 466-468) and delete-a-group jackknife estimation (Rao, 1993) to obtain measures of model performance and t-statistics for individual parameter estimates, respectively. To perform cross-validation, we randomly split the establishments in each industry into two equal-sized sets: a model-building set and a validation set. Then for each multiple regression model, we computed four alternative sets of regression parameters (one per regression method) from data in the model-building set and used these parameters to predict the same characteristic in the validation set. After this, we calculated three different statistics to measure prediction performance:

Mean Absolute Error (MAE) for method *m* $\sum_{i}^{n} \left| \hat{Y}_{im} - Y_i \right| / n$

(DeGroot, 1987, pp. 209-211) where *i* indexes the industry within trade area, $\hat{Y}_{im}$ is the tabulated value obtained from imputing data item Y in the validation data set using regression parameters obtained with method m (m = 1- 4: weighted least squares, resistant regression with S_{pk} = mean of residuals, resistant regression with S_{pk} = median of residuals, and ridge regression), and Y_i is the tabulated value (from reported data in the validation data set) of the data item in industry i.

Mean Absolute Deviation (MAD) for method *m* and industry *i* $\sum_{j}^{n_i} \left|\hat{Y}_{ijm} - Y_{ij}\right| / n_i$

(Nordholt, 1998) where j indexes the establishment within industry i and n_i is the number of establishments in industry i .

MAD score for method *m* # of industries where MAD(m,i) is minimum of all four methods. For each model, this score identifies the method that produces the lowest MAD in the most industries.

Mean Ratio of Predicted to True Value (MRPT) for method *m* $\sum_{i}^{n} \left|\hat{Y}_{im} / Y_i\right| / n$

For delete-a-group jackknife estimation, we randomly split the establishments in each industry into 16 groups. Each delete-a-group jackknife replicate estimate was obtained by dropping one group k at a time, weighting the remaining observations by (16/15) and calculating the parameter estimates from the remaining 15 groups. Delete-a-group standard errors of each parameter estimate β_{pim} are:

$$\hat{\sigma}(\hat{\beta}_{pim}) = \left[(k-1)/k\right]\sum_{k=1}^{K}(\hat{\beta}_{pimk} - \hat{\beta}_{pim})^2 .$$

where p indexes the regression parameter. Within each trade area, we evaluated the average coefficient of variation of each parameter (for each multiple regression model) and the frequency of statistically significant parameters in each model by industry (using t-statistics at the 90% confidence level).

Selecting Regression Methods (Parameter Estimation Methods) for Each Item

Prior to model selection, we had to choose a parameter estimation method for each item, assuming that the "best" regression method within a trade area could differ by item, and the "best" regression method for a given item could differ by trade area. To predict annual payroll, we considered all permutations of available covariates (up to three covariates in Retail Trade, Utilities, FIRE and certain Services Industries; up to four covariates in Wholesale Trade (excluding purchases) and the remaining Services Industries. For the remaining data items' models, we excluded 1st quarter payroll as an independent variable because of its perfect correlation with annual payroll (i.e., we did not include both items as independent variables in the same model).

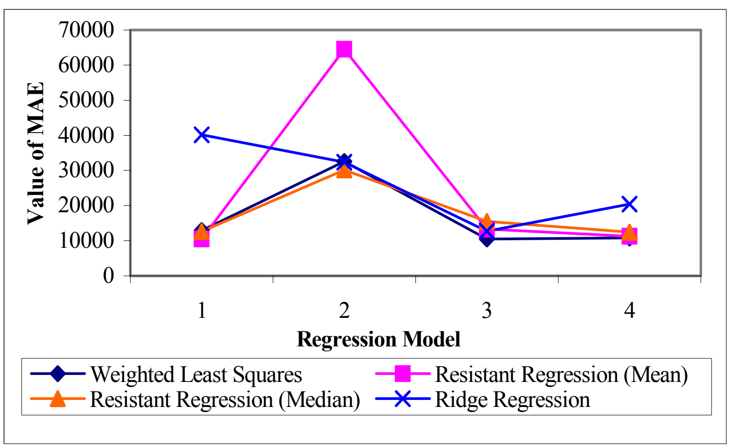

Figure 1: MAE for Multiple Regression Models Estimating Annual Payroll

To assess the performance of the four parameter estimation methods, we looked at all possible multiple regression models for an item within each trade area [Note: we considered only weighted least squares regression for the industry average imputation]. For each item (within trade area), we selected one parameter estimation method from the four after examining their respective MAE and MAD statistics. We believed that the estimation method should **consistently** yield the lowest MAE for a given dependent variable regardless of covariates and should generally yield the lowest MAD within trade area. Ideally, the mean ratio of the predicted to true value (MRPT) should also be close to one.

Figure 1 illustrates this process, plotting MAE for all 3- and 2-covariate models predicting annual payroll in our 30 FIRE industries [Note: we use only weighted least squares estimation – the production method – for the single covariate models]. Each point on the horizontal axis represents a multiple regression model (see Models 1- 4 in Table 1). Except for model 2, weighted least squares regression yields the lowest MAE (for model 2, the MAEs for the three "close" models are all within 150 units of one another). Frequency distributions of MAD/method by industry (MAD scores) show

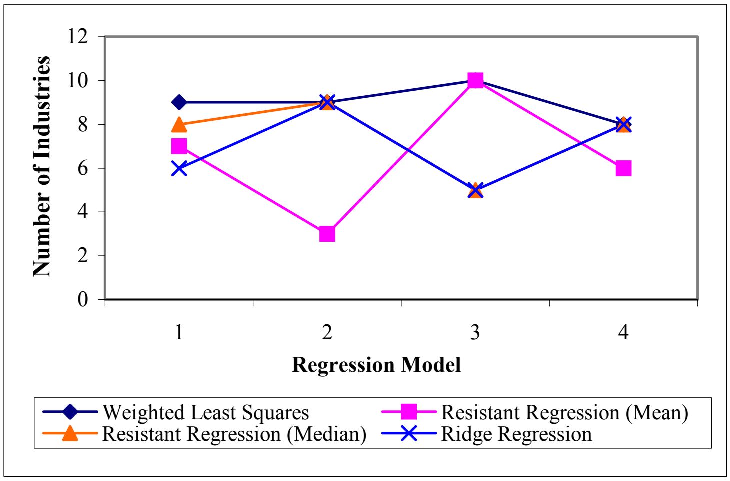

Figure 2: MAD score for Multiple Regression Models Estimating Annual Payroll

the same pattern, as plotted in Figure 2, i.e., methods with low MAEs in Figure 1 are associated with high MAD scores in Figure 2.

In all trade areas, the multiple regression models fit with weighted least squares generally had the lowest MAEs for **all** dollar value items (annual payroll, receipts, operating expenses, and purchases). Except for the Retail Trade and Utilities industries, ridge regression yielded the lowest MAE for the models predicting employment.

Initially, we were hesitant to recommend ridge regression models based on 1997 data for modeling 2002 employment. Choosing a biasing constant for ridge regression is a subjective procedure, even when the choice is based on diagnostic statistics. To assess the sensitivity of our ridge regression models for employment to the choice of biasing constant, within our Wholesale Trade industries, we refit the ridge regression employment models in with c = 0.1, 0.2, 0.3, and 0.5 and found the same patterns of MAE (i.e., lowest from ridge regression of the four parameter estimation methods for all multiple regression models). Since our employment models introduce very small biasing constants, this insensitivity to c is not unreasonable. However, we were still concerned that the values of c obtained from our model-building datasets might not be appropriate for other samples in the same prediction range. We were also concerned that we were simply not comparing enough multiple regression models within industry. Except for tax-exempt Services Industries and Wholesale Trade, we only had one available regression multiple regression model to compare to the industry averages. To provide more models and to assess the sensitivity of c to the model-building dataset, we repeated the cross-validation in many of our test industries. The results were quite consistent.

While ridge regression worked quite well for predicting employment, its performance was inconsistent when modeling dollar value items. We believe that ridge regression works so well for modeling employment because all of the available independent variables are dollar value items. Much of the multicollinearity is alleviated by the correlation-transformation, and the biasing constants used in the employment ridge regressions are quite small (compared to those used to model dollar variables).

We were surprised by the poor performance of resistant regression with most of our data, given the current surveys' excellent resistant regression results. However, the model-building data sets used by the current surveys include outliers, and the current surveys generally use simple-linear regression as imputation models. The census model-building data are restricted to observations that satisfy all of the ratio edits and consequently do not contain many outliers.

Model Selection

The MAE comparisons used for parameter estimation selection provided credible evidence of improved predictions over industry average predictions with many of our multiple regression models. The degree of improvement over industry average varied considerably by trade area and by item, however. Our subject-matter analyst experts were generally satisfied with the industry average imputation results and required proof of "substantial" improvements via multiple regression to justify the additional analyst-time in parameter estimation and data storage space. We developed the following procedure for obtaining reduced sets of multiple regression models for a data item/trade area:

1. Using the statistics obtained with the recommended regression method (weighted least squares for all dollar value items and employment in the Retail and Utilities trade areas, ridge regression for employment in the FIRE, Wholesale and Services trade areas), rank the no-intercept regression models for the same data item by ascending MAE;
2. Drop all models that do not show any improvement over each of the **already implemented** industry average imputations in terms of MAE;
3. Use significance test results to further reduce the candidate set of reduced models (e.g., drop models that contain more than one independent variable whose parameter tests are rarely significant, especially if they do not exhibit marked

improvements in MAE from other models);

4. If the MAE between two or more models was quite close, examine MRPT to further reduce models;
5. Again using MAE and significance test results, determine whether including an intercept term in the recommended beta models **and** industry average models improves their predictive power.

Characteristics of the production software also affected our model selection procedure. The PV ratio module attempts all regression imputations in one consecutive block. Regression models cannot be interspersed with other types of imputation models, e.g., the user could not try a regression model, then a logical edit, then another regression model for the same variable.

This example from the FIRE trade area illustrates our model selection procedure. The FIRE data had three available covariates for modeling annual payroll: employment (EMP), receipts (RCP), and 1st quarter payroll (QPR). FIRE currently uses three different industry average imputation models, one per covariate. Table 1 presents the seven candidate no-intercept regression models, with their associated MAE rankings. Model 5, which uses the industry average from 1st quarter payroll, has the lowest MAE. Models 6 and 7 – the other industry average models – have considerably higher MAE's than many of the multiple regression models. So, any retained multiple regression models must have lower MAE than Model 6. For this reason, we dropped Model 2, leaving three remaining candidate multiple regression models. Of these three, we eliminated Model 1 since it had higher MAE than those from Models 3 and 4 (and more covariates). Finally, we dropped Model 4 because its MAE was very close to the Model 3 MAE, and the MRPT was lower than the corresponding value for Model 3. This left us with four possible models (models 3, 5, 6, and 7). We then refit these models with an intercept and repeated the same analysis. Table 2 presents the MAE and MRPT for each of these models, again ranked by ascending MAE.

Table 1: Candidate Models for Annual Payroll

Model	Model Rank	Model Type	MAE	MRPT
1. $APR = \beta_1 EMP + \beta_2 RCP + \beta_3 QPR$	4	Multiple Regression	12872.77	.9643
2. $APR = \beta_1 EMP + \beta_2 RCP$	6	Multiple Regression	32592.12	1.0266
3. $APR = \beta_1 QPR + \beta_2 EMP$	2	Multiple Regression	10479.41	.9838
4. $APR = \beta_1 QPR + \beta_2 RCP$	3	Multiple Regression	10759.53	.9614
5. $APR = \beta_1 QPR$	1	Industry Average	8838.25	.9839
6. $APR = \beta_1 EMP$	5	Industry Average	24924.56	.9958
7. $APR = \beta_1 RCP$	7	Industry Average	65526.00	1.1275

Of the four new with-intercept models, only Model 3a appears promising, with slightly improved predictions over its corresponding no-intercept model. However, the intercept term was significant in only four of the 30 test industries, so we decided not to recommend this imputation model. In fact, the intercept term was rarely significant in our models. This is not unreasonable: for example, one would expect an establishment with no payroll to have no 1st quarter payroll, employment, or receipts. Ultimately, we recommended adding only multiple regression Model 3 to the suite of available imputation options.

Table 2: Candidate Reduced Models for Annual Payroll With and Without Intercept Terms

Model	Model Rank	Model Type	MAE	MRPT
3. $APR = \beta_1 QPR + \beta_2 EMP$	4	Multiple Regression	10479.41	.9838
3a. $APR = \beta_0 + \beta_1 QPR + \beta_2 EMP$	3	Multiple Regression	10215.30	.9864
5. $APR = \beta_1 QPR$	1	Industry Average	8838.25	.9839
5a. $APR = \beta_0 + \beta_1 QPR$	2	Multiple Regression	8850.87	.9839
6a. $APR = \beta_0 + \beta_1 EMP$	5	Multiple Regression	23053.72	1.0114
6. $APR = \beta_1 EMP$	6	Industry Average	24924.56	.9958
7. $APR = \beta_1 RCP$	8	Industry Average	65526.00	1.1275
7a. $APR = \beta_0 + \beta_1 RCP$	7	Multiple Regression	62371.41	1.1035

Table 3 presents the number of recommended multiple regression models in each trade area for each item. For models that do not predict annual payroll, we can substitute 1st quarter payroll for annual payroll in any recommended model, thereby increasing the number of available imputation options.

Table 3: Number of Recommended Multiple Regression Models

		Payroll	Employment	Sales	Operating Expenses	Purchases
Wholesale		3	5	3	4	17
FIRE		1	1	1	N/A	N/A
Services	Tax exempt	4	3	2	3	N/A
	Non Tax Exempt	3	2	4	N/A	N/A
Retail		2	6	5	N/A	N/A
Utilities		4	6	2	N/A	N/A

At first glance, it may appear that we are recommending adding an excessive number of new imputation models for purchases. However, this item is traditionally very poorly reported and is very difficult to impute from a single covariate. Recall that Wholesale Trade industries collect six basic data items (five excluding 1st quarter payroll), giving up to 41 possible imputation models. For the other items, the number of recommended additional imputation models is quite in line with the number of available covariates [Note that we did **not** include purchases as a covariate in any Wholesale Trade model because of its poor reliability].

Conclusion

This paper presents the results of a comparison of alternative regression parameter estimation methods using data from the 1997 Economic Census. Our model-building data are highly multicollinear and our fitted models exhibit heteroscedasticty. We present methods of model evaluation and model selection that are not dependent on sums-of-squares statistics. Thus, our evaluation methods can be applied to other (similar) situations.

In theory, (weighted) least squares estimates from multicollinear regression models can be used for prediction, as long as the prediction region is in the same range as the model-building data. We attempted to enforce this by using the region defined by the ratio edit parameters as our prediction region. While this may be sufficient for our dollar value variables, we are not sure that it will be true for employment. Furthermore, we are uncertain whether our recommended ridge regression models for employment will perform consistently with the 2002 Economic Census data because of the change in data collection as well as the subjectivity of the modeling procedure.

All of the trade areas represented by the services sectors calculate two sets of imputation parameters for the Economic Census: the first set are developed from the complete prior census data and the second set are developed from the current census data after sufficient cases have been processed. Rather than develop two sets of multiple regression parameters for production, we used the 1997 data for research, developing recommendations for trade-area specific regression models for each basic data item. We plan to use the available 2002 Economic Census data to validate the recommended models and to re-assess the proposed imputation hierarchy. Thus, the production regression parameters will be developed from 2002 data **only,** if in fact, our results hold up.

References

DeGroot, M. (1987). Probability and Statistics. Reading, MA: Addison Wesley.

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis. New York: John Wiley and Sons, Inc.

Huang, E.T. (1984). An Imputation Study for the Monthly Retail Trade Survey. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 610 – 615.

Greenberg, B. (1986). *The Use of Implied Edits and Set Covering in Automated Data Editing.* Washington, D.C.: U.S. Census Bureau (Technical report Census/SRD/RR-86/02, available from the Statistical Research Division).

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression. Reading, MA: Addison Wesley.

Neter, J., Wasserman,W. and Kutner, M.H. (1989). Applied Linear Regression Models. Homewood, IL: Richard D. Irwin, Inc.

Nordholt, E.S. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review,* **66**, pp. 157-180.

Rao, J.N.K. (1993). Resampling Methods for Complex Surveys. *Course notes from Washington Statistical Society Seminar.*

Thompson, K.J. and Adeshiyan, S.A. (2003). Quality Effects of Alternative Edit Parameters. *Journal of Data Science*, **1**, pp. 1-27.

Thompson, K.J. and Sigman, R.S. (1996). *Evaluation of Statistical Methods for Developing Ratio Edit Module Parameters.* Washington, D.C.: U.S. Census Bureau (Technical report #ESM-9601, available from the Economic Statistical Methods and Programming Division).