LMI Training Institute
June 16-20, 2003

Prepared by the Staff of the LEHD Program

June 2, 2003

Chapter 3

QWI Imputations

3.1 Using Imputation to Obtain Characteristics of the Workplace

We developed the QWI system to provide you with detailed workforce indicators that reflect the characteristics of the employees (for example, age, sex and place of residence) and the characteristics of the workplace (for example, industry and geographical location). Obtaining the characteristics of the employee is a direct application of data integration. We use the Census Personal Characteristics File (which is based on the Census Numident) and the Census Place of Residence File (an administrative record data system) to obtain these items. Obtaining the characteristics of the workplace is a more complicated task. We use information collected at both the unemployment insurance account level (SEIN) and the ES-202 workplace level (SEINUNIT) to develop the characteristics of the work place. This section describes the process of obtaining workplace characteristics and the quality assurance system that we designed to support it.

The fundamental unit of micro data in the QWI system is a record that describes an employment relation between an individual (identified by PIK) and an employer (identified by SEIN). This record contains UI wage reports for all the quarters in which the individual received such payments from the indicated employer. All of the QWIs are based upon information that can be derived from this record, which we call an employment history record. To integrate characteristics of the individual we look up the PIK in our data base of individual characteristics, which was constructed from the sources noted above. To integrate characteristics of the workplace, we cannot simply look up the SEIN in our data base of employer characteristics because many employers have multiple locations (called multi-unit employers). We use statistical methods to impute the workplace characteristics for multi-unit employers. The method is called "multiple imputation" because our statistical models are used to generate several different sets of workplace characteristics for each employment history record associated with a multi-unit employer. We calculate the QWIs for each of the scenarios and average them to produce the estimates that we release for your use. (We also use "multiple imputation" to address missing data problems when integrating the individual characateristics.)

In order to apply multiple imputation to the problem of measuring workplace characteristics for employees of multi-unit employers, we developed a statistical model for predicting the establishment associated with a particular employment history record. Our partner state Minnesota provided us with a UI wage record data base in which the ES-202 workplace unit (SEINUNIT) was coded. For Minnesota, therefore, we do not have the problem of missing workplace characteristics for multi-unit employers because we know both the employer unemployment insurance account (SEIN) and the workplace (SEINUNIT). Thus, we compute the Minnesota QWIs based on the actual place of employment, as recorded in their UI wage records. We also use Minnesota data to estimate the statistical model that is the basis for imputing the multi-unit work place in the other states.

The statistical model for imputing multi-unit employer characteristics is based upon three important inputs. First, we developed an entity demographic history for every SEIN and SEINUNIT in your state's UI wage record and ES-202 data. This entity demographic history summarizes births, deaths, consolidations, breakouts, false births, and false deaths for every UI account number and every ES-202 reporting unit number in the underlying data bases. The entity demographic history file is the basis for the successor/predecessor micro data described in other parts of this booklet. It integrates both administrative and statistical information regarding the birth, death, consolidation and break-out events. Second, we developed a history of month-one employment estimates for each ES-202 unit (SEINUNIT) that

was consistent with the information in the entity demography history. Third, we used the Census Place of Residence information and ES-202 address information to develop an estimate of the distance between the residence of each individual employed by a multi-unit employer and the location of all of the feasible employing units. The "distance to work" measure that we computed is based upon geo-coding the latitude and longitude of both the place of residence and the place of work for every indivdiual and establishment in the QWI system.

The most important features of the statistical model that we developed for imputing characteristics of multi-unit employers are listed below:

- for a given employment history, only the feasible units (those which had positive employment for all of the quarters that this individual was employed by the multi-unit employer) have a positive probability of imputation;
- once an individual-employer combination has been processed by the probability model, it is never processed again (thus, the imputation does not create false labor market transistions);
- the probability that an individual is employed by a given establishment increases as the distance between the individual's residence and the potential place of work decreases;
- the probability that an individual is employed by a given establishment increases as the number of employees at that establishment increases;
- the probability that an individual is employed in a given establishment is dynamically consistent–it increases and decreases to account for accessions and separations that occur over the entire period of the QWI data base;
- the imputation model was sampled 10 times, providing 10 independent implicates of the place of work for each individual employed by a multi-unit employer;
- characteristics of each of the 10 imputed places of work were used to compute a complete set of QWIs; and
- the 10 independently generated QWIs were averaged to produce the estimates we released.

Estimation of the statistical model using data for Minnesota gave very reliable results. The estimated probability model showed a strong, reliable relationship between the distance-to-work and the true employing establishment. This relation was different, as expected, for small, medium, and large businesses. The estimated probability model also did a reliable job of tracking the sizes of the underlying units based on the time-series of month-one employment levels for all of the SEINUNITs in a given SEIN, which established that we would be able to use the model for other states. To assess the overall reliability of the imputation system, we calculated the entire QWI statistical system for Minnesota using the true establishment (SEINUNIT) and using the imputation system as it would be applied in the other partner states. For Minnesota, we compared the value of each QWI estimate based on the true establishment with its counterpart based on the 10 imputed establishments for every geography-industry combination in every quarter. These results established that our imputation system worked.

3.2 Implementation

Ignoring the time dimension,

$$p\left(\beta|x,y\right) = \int \left\{\prod_j p\left(\alpha_j\right)\right\} \left\{\prod_j \prod_{i\in j} \prod_{r\in i,j} \left(\frac{e^{\alpha_{jr}+x_{ijr}\beta}}{\sum_{s\in i,j} e^{\alpha_{js}+x_{ijs}\beta}}\right)^{d_{ijr}}\right\} d\boldsymbol{\alpha} \tag{3.1}$$

$$\approx \frac{1}{L}\sum_l \left\{\prod_j \prod_{i\in j} \prod_{r\in i,j} \left(\frac{e^{\alpha^l_{jr}+x_{ijr}\beta}}{\sum_{s\in i,j} e^{\alpha^l_{js}+x_{ijs}\beta}}\right)^{d_{ijr}}\right\} \tag{3.2}$$

where $p(\alpha_j)$ are priors on α_j, which are the logits of Dirichlet random variables. The gradient is

$$\frac{\partial p(\beta|x,y)}{\partial \beta} = \int \left\{ \prod_j p(\alpha_j) \prod_{i \in j} \prod_{r \in i,j} (p_{ijr})^{d_{ijr}} \right\} \left\{ \sum_j \sum_{i \in j} \sum_{r \in i,j} d_{ijr} (x_{ijr} - \bar{x}_{ij}) \right\} d\boldsymbol{\alpha} \tag{3.3}$$

$$\approx \frac{1}{L} \sum_l \left\{ \left[\prod_j \prod_{i \in j} \prod_{r \in i,j} \left(p^l_{ijr}\right)^{d_{ijr}} \right] \left[\sum_j \sum_{i \in j} \sum_{r \in i,j} d_{ijr} \left(x_{ijr} - \bar{x}^l_{ij}\right) \right] \right\} \tag{3.4}$$

where

$$p_{ijr} = \frac{e^{\alpha_{jr} + x_{ijr}\beta}}{\sum_{s \in i,j} e^{\alpha_{js} + x_{ijs}\beta}} \tag{3.5}$$

$$\bar{x}_{ij} = \sum_{r \in i,j} p_{ijr} x_{ijr} \; . \tag{3.6}$$

Including the time dimension,

$$p(\beta|x,y) = \int \left\{ \prod_t \prod_{j \in t} p(\alpha_{j \cdot t}) \right\} \left\{ \prod_t \prod_{j \in t} \prod_{i \in j,t} \prod_{r \in i,j,t} \left(\frac{e^{\alpha_{jrt} + x_{ijrt}\beta}}{\sum_{s \in i,j,t} e^{\alpha_{jst} + x_{ijst}\beta}} \right)^{d_{ijrt}} \right\} d\boldsymbol{\alpha} \tag{3.7}$$

$$\approx \frac{1}{L} \sum_l \left\{ \prod_t \prod_{j \in t} \prod_{i \in j,t} \prod_{r \in i,j,t} \left(\frac{e^{\alpha^l_{jrt} + x_{ijrt}\beta}}{\sum_{s \in i,j,t} e^{\alpha^l_{jst} + x_{ijst}\beta}} \right)^{d_{ijrt}} \right\} \; . \tag{3.8}$$