2/2/2021

SUPPORTING STATEMENT

FLORIDA AGRICULTURAL WORKERS SURVEY

PART B:
COLLECTION OF INFORMATION EMPLOYING STATISTICAL METHODS

Introduction

Part A of the Supporting Statement, the justification, is set forth as a separate document. This document, Part B, sets forth the population being sampled and the statistical methods being employed to guide the collection of data. The Florida Agricultural Worker Survey (FAWS) is to collect data on workers active in Florida citrus, tomato and strawberry production. A two-stage probability sampling procedure is specified where the first stage is employers, and the second stage is the workers. Design-based estimates are developed for totals and means of the population, with primary emphasis on earnings. In addition, model-based estimates are developed for the estimation of legal status domain means for two categories of workers: those authorized for work in the United States and those unauthorized for work in the United States.

1. Description of Universe and Sample

Universe and sample

The universe for the study is the population of field workers active in citrus, tomato, and strawberry production in Florida. The Florida Agricultural Worker Survey will use two-stage sampling to interview approximately 1,624 randomly selected workers in these three commodities. The two-stage sampling procedure (first employers, and then workers) not only reduces sampling error, but is also a viable procedure for sampling workers in specific activities when there is no readily available sampling frame for the universe of workers in those activities.

In the first stage, employers are stratified into five strata to minimize sampling error and to reduce the cost of sampling. Five strata are designated to assure coverage across the three commodities of interest (citrus, tomato and strawberry employers) and important types of employers. There are three significant types of employers in citrus: growers, labor contractors, and grove care firms; each is designated as a separate stratum. Some citrus growers are direct employers of their labor while others hire no labor and use intermediaries such as labor contractors and grove care firms to perform the production activities. Much of the harvesting is done by labor contractors and much of the grove care (horticultural activities) is provided by independent grove care firms. By contrast, most employment in tomatoes and strawberries is direct employment by the grower; a single stratum is designated for each of the latter two commodities. The frame for the sample is the Quarterly Census of Employment and Wages (QCEW) data for Florida (Agency for Workforce Innovation). The universe and the proposed sample for each of the strata are displayed in Table 1.[1] The allocation of the sample across strata is discussed in section 2 below.

[1] Sample design calculations are based on currently available QCEW data. Upon approval, this will be adjusted with the most recently available QCEW data prior to the sample selection. We do not anticipate significant differences with the later data.

Table 1. Universe and sample of employers and workers

Stratum		Employers		Workers		Worker
		Population*	Sample	Population*	Sample	Sample (%)
1	Citrus growers	360	72	10,870	493	4.54
2	Citrus labor contractors	254	54	17,025	347	2.04
3	Citrus grove care	79	18	1,115	65	5.83
4	Tomatoes	60	22	11,295	431	3.82
5	Strawberries	81	18	10,351	288	2.78
	Total	834	184	50,656	1,624	3.21

*Extracted from Florida Agency for Workforce Innovation (AWI) data.

Key Parameter of Study

Since the primary focus of the analysis to be conducted with the data is on worker earnings, the primary parameter of interest to be estimated with the data is the mean of worker earnings. The population mean of earnings is defined as

$$\bar{Y} = \sum_{h=1}^{5} \frac{M_h}{M} \sum_{i=1}^{N_h} \frac{M_{hi}}{M_h} \sum_{j=1}^{M_{hi}} \frac{Y_{hij}}{M_{hi}}$$

where $\bar{Y}$ = mean of worker earnings

Y_{hij} = earnings of the jth worker of the ith employer in the hth stratum

M_{hi} = number of workers for the ith employer in the hth stratum

M_h = number of workers across all employers in the hth stratum

M = number of workers across all employers and strata

N_h = number of employers in the hth stratum.

Response Rate

The sampling design (described in 2 below) involves obtaining a random selection of employers in each of the three commodities. We anticipate a response rate at least as good as the NAWS has obtained given the long history of association between the University of Florida and Florida agriculture. In FY 2003, 75 percent of the NAWS randomly selected employers (or their surrogates) throughout the U.S. who were eligible cooperated. The NAWS response rate for workers contacted has been about 90%; we anticipate a similar response rate since the same contractor will be conducting the interviews with interviewers experienced from the NAWS worker interviews. The sample numbers in Table 3 have been adjusted for the above projected employer and worker non-response rates; they are the expected number of respondents.

2. Statistical Methodology

Sample Design

The sample is designed as a stratified two-stage random sampling procedure. The first stage sampling units are employers which are to be sampled without replacement with probability proportional to the square root of estimated size. Employers are stratified into five strata by commodity and type of employer as explained in section 1.a. For each selected employer a systematic sample of the second stage units, workers, is to be selected.

Selection of Employers

Within each of the five strata, employers are selected with probability proportional to the square root of estimated size and without replacement. There are alternative measures of size for seasonal agricultural employers, including payroll, average annual employment, and monthly employment levels. Since our objective is to represent the workforce in the three selected commodities, we believe the best measure of employer size is relative to the maximum monthly employment for the employer, M_{hi}, for the i[th] employer in stratum h. Since the survey will be conducted during the peak season, this measure best represents the number of workers to be encountered at each employer. This measure is taken from the Florida Agency for Workforce Innovation data. Since there is a lag in the availability of the information, the size measure used, $\sqrt{M'_{hi}}$, will result in an employer sample that is selected with probability proportional to the square root of estimated size, *ppz.*

The initial probability of selection of the i[th] employer in the h[th] stratum is therefore $z_{hi} = \sqrt{M'_{hi}} \Big/ \sum_{i=1}^{N_h} \sqrt{M'_{hi}}$, and clearly $\sum_{i=1}^{N_h} z_{hi} \equiv 1$. In sampling without replacement, the z_{hi}'s are the selection probabilities only for the first draw of the sampling procedure. The probability that the i[th] employer is included in the sample, the inclusion probability, is defined through an iterative procedure as $\pi_{hi} = \min(1, n_h z_{hi})$ where n_h is sample size for the h[th] stratum. In cases where the product of the sample size (n_h) and initial selection probability (z_{hi}) exceeds one, the inclusion probability is set to unity for the employer (it is selected with certainty), n_h is reduced by one, and the z_{hi} are redefined as above over the remaining non-certainty employers. The process is repeated with the remaining non-certainty employers until there are no further instances of the products of n_h and z_{hi} exceeding one. The end result is the set of inclusion probabilities for the employers defining the probability that the i[th] employer will be selected in the sample of size n_h.

The formal (first stage) sample design for employers is a conditional Poisson sampling (CPS) design. Having defined the inclusion probabilities above, the necessary joint inclusion probabilities can be readily calculated for given sample size with reference to the CPS design (Tillé 2006, §5.6). This design for the employer sample is sometimes referred to as utilizing maximum entropy sampling for drawing the sample (Tillé 2006). The algorithm UPmaxentropy available in Tillé and Matei (2008) and based on the CPS design is to be used to draw the sample of employers in accordance with the *a priori* inclusion probabilities.

Selection of Workers

The second stage of the sampling procedure is the selection of workers. Given a selected employer, workers are sampled using systematic sampling. With a planned sampling rate for the i[th] employer's workers of $p'_{hi} = m'_{hi} / M'_{hi}$, m'_{hi} is the planned sample size for the i[th] employer in stratum h. Following Lahiri's (1951) suggestion, set $k'_{hi} = M'_{hi} / m'_{hi} = 1/p'_{hi}$ and round k'_{hi} to the closest integer, k_{hi} (Cochran 1977, p. 206). Thinking of the M'_{hi} workers arranged in a circle, first choose a random number between 1 and M'_{hi}. Select this worker, and then every k_{hi}^{th}

worker until m'_{hi} have been selected; if M'_{hi} is reached before m'_{hi} workers have been selected continue around the circle until m'_{hi} workers have been selected. Combining the employer and worker selection probabilities, for any worker in the universe, the selection probability for any worker is $f_{0hi} = \pi_{hi} / k_{hi}$.

Accuracy

Sample size is constrained by the budget available for the survey. Allocations among strata follow Cochran's (1977, p. 317) suggestion for sample allocation with primary units of unequal size. The optimum probability (f_{0h}) of selecting any subunit (worker) is set such that

$$f_{0h} \propto \frac{1}{\sqrt{c_{2h}}} \sqrt{\sum_i (M_{hi} / M_h) S_{2hi}^2}$$

where $f_{0h} = n_h z_{hi} m_{hi} / M_{hi}$, c_{2h} is the cost of including an additional worker in the sample, assuming a linear cost function, and S_{2hi}^2 is the population variance of earnings for the i^{th} employer in stratum h (equation .[2] Since the selection probabilities are constant within stratum, the sample is self-weighting within stratum. The allocations among strata displayed in Table 3 are based on equation .

The variance of mean earnings is calculated with the QCEW data with the allocations specified in Table 3, based on equation below. The sampling design results in a standard error of the mean of earnings that is 1.8% of the population mean. Using a confidence interval that is ± 2 standard errors from the sample mean implies that we can be 95% confident that the true population mean is no more than 3.6% from the sample mean as calculated with the survey data. This is more than an adequate degree of accuracy for the desired use of the data in establishing the potential effects of changes in the labor market on considerations such as the adoption of mechanization. For example, a recent study of the adoption of mechanization in the Florida sugar cane industry indicated that the threshold value for labor cost to result in an immediate adoption of mechanical harvesting of sugar cane in the early 1970s was in excess of 58% of the existing labor cost (Iwai, et al. 2008). This was a time when there was limited mechanical harvesting of sugar cane, and analysts were indicating that the industry should adopt at that time. While we do not know the corresponding threshold value for citrus, tomatoes, or strawberries, it is clear that there remains a significant threshold value of labor cost given the limited adoption of mechanical harvesting in Florida citrus over the past several years. The precision achieved with the sample will permit the evaluation of changes in subgroups such as authorized and unauthorized workers on the labor market outcome. Existing work indicates that the wage differential between the two groups is between 10 and 20 percent (Isé and Perloff 1995; Iwai, et al., 2006; Walters, et al. 2008).

Design Effect

[2] Since there is no information on within employer wage variation in the QCEW data used for the sample design, external information was used to supplement the QCEW data. See the appendix.

An important consideration for a survey with a complex design is the design effect (deff) (Kish 1965). The design effect contrasts the efficiency of the survey design in estimating a parameter relative to a simple random sample of the same size.

$$deff = \frac{V\left(\hat{\bar{\bar{Y}}}\right)}{V_{SRS}\left(\hat{\bar{\bar{Y}}}\right)}$$

The variance of the estimated mean under the assumption of a simple random sample is

$$V_{SRS}\left(\hat{\bar{\bar{Y}}}\right) = \frac{S^2}{m}\frac{M - m}{M}$$

with

$$S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{\bar{Y}})^2/(M - 1)$$

Although the y_{ij} are not directly observable in the QCEW data, the necessary sums of squares can be recovered from the S^2_{2hi} defined in equation below. The resulting calculations based on the QCEW data yield a *deff* = 1.401, implying that the complex design results in an increase in the variance of 40.1% relative to a simple random sample for the given sample size. This is the necessary tradeoff for using a design that facilitates the location of a random sample of farm workers in comparison to an exorbitantly costly household survey to find a small subgroup of the general population.

Estimation Procedures

The estimation relies on both design-based estimation and model-based estimation. The estimate of average earnings for the population is a design-based estimate. Although design-based estimates are typically preferred for summarizing survey data since they incorporate few assumptions, a model-based approach is set forth for the estimation of average earnings for two important subgroups: authorized and unauthorized workers. As is discussed below, the likely misclassification of unauthorized workers as authorized workers due to errors in self-reporting of legal status is problematic for a design-based estimate of average earnings by work authorization status. The model-based approach accounting for the misclassification is particularly useful for estimating the difference in earnings by work authorization status.

Average earnings. With the two-stage random sample of workers, the worker data can be inflated, where necessary, to generate population estimates. The underlying estimator is the standard Horvitz-Thompson estimator. We illustrate with the estimation of the key variable of interest, average earnings. The general form of the estimate of average earnings is

$$\hat{\bar{\bar{Y}}} = \frac{\hat{Y}}{M}$$

where $\hat{Y}$ is the estimate of total earnings for the population, and M is the total number of workers. Recall that the sampling of employers is *ppz*, probability proportional to square root of *estimated* size. Although at the time of sampling the size of the employers sampled, M_{hi}, will be

known from building the sampling frame, the size of employers not sampled will not be known, and as a result the true population size, M, is unknown. Since the population of workers must also be estimated, the denominator of equation is an estimate based on the observed M_{hi}. The M_{hi} are random variables dependent on which employers are randomly selected for the sample. As a result, the population estimate of average earnings is a ratio estimate of the form

$$\hat{R} = \hat{\bar{Y}} = \frac{\hat{Y}}{\hat{M}}$$

Starting with the earnings of the j^{th} sampled worker of the i^{th} sampled employer in the h^{th} stratum, y_{hij}, estimated total earnings for the h^{th} stratum are

$$\begin{aligned}
\hat{Y}_h &= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{z_{hi} m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \\
&= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi} \bar{y}_{hi}}{z_{hi}} \\
&= \sum_{i=1}^{n_h} \frac{M_{hi} \bar{y}_{hi}}{\pi_{hi}} \\
&= \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{\pi_{hi}}
\end{aligned}$$

The last line of equations is the standard Horvitz-Thompson (1952) estimator for a total (Cochran 1977, p. 259). Similarly, the estimated number of workers in stratum h is

$$\hat{M}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{z_{hi}} = \sum_{i=1}^{n_h} \frac{M_{hi}}{\pi_{hi}}$$

The estimated population totals are the sums of the separate strata estimates:

$$\hat{Y}_{ST} = \sum_{h=1}^{5} \hat{Y}_h$$

$$\hat{M}_{ST} = \sum_{h=1}^{5} \hat{M}_h$$

The resulting population estimate for average earnings is the ratio estimate

$$\hat{\bar{Y}} = \frac{\hat{Y}_{ST}}{\hat{M}_{ST}}$$

Although biased, the ratio estimate is a consistent estimate for the population parameter. It is also asymptotically normally distributed so that inference is not an issue for statistics based on large samples (Cochran 1977, p. 153); with an expected sample size of 1,624 observations, the reliance on asymptotic results is not problematic.

As a ratio estimate, the variance of the estimate is necessarily an approximation. The approximation for the mean square error (MSE) or variance of the mean earnings statistic defined in equation is defined, following Cochran (1977, Ch. 11) as:

$$V\left(\hat{\bar{Y}}\right) = V\left(\hat{R}\right) \approx \frac{1}{M^2}\sum_{h=1}^{5}\sum_{i=1}^{N_h}\left\{\sum_{j>i}^{N_h}\left[\left(\pi_{hi}\pi_{hj} - \pi_{hij}\right)\left(\frac{D_{hi}}{\pi_{hi}} - \frac{D_{hj}}{\pi_{hj}}\right)^2\right] + \frac{M_{hi}(M_{hi} - 1)S_{d2hi}^2}{m_{hi}\pi_{hi}}\right\}$$

where $D_{hi} = Y_{hi} - RM_{hi}$

$$R = Y/M$$

Y_{hi} = *sum* of worker earnings for the i^{th} employer in the h^{th} stratum

π_{hij} = joint inclusion probability for employers *i* and *j* in stratum *h*, and

$$S_{d2hi}^2 = \frac{1}{M_{hi} - 1}\sum_{j=1}^{M_{hi}}\left[\left(y_{hij} - Rm_{hij}\right) - \left(\bar{Y}_{hi} - R\bar{M}_{hi}\right)\right]^2$$

$$= \frac{1}{M_{hi} - 1}\sum_{j=1}^{M_{hi}}\left[\left(y_{hij} - R\right) - \left(\bar{Y}_{hi} - R\right)\right]^2$$

since $m_{hij} \equiv 1$, and $\bar{M}_{hi} = \frac{\sum_{j=1}^{M_{hi}} m_{hij}}{M_{hi}} = \frac{M_{hi}}{M_{hi}} \equiv 1$, and therefore

$$S_{d2hi}^2 = \frac{1}{M_{hi} - 1}\sum_{j=1}^{M_{hi}}(y_{hij} - \bar{Y}_{hi})^2 \equiv S_{2hi}^2$$

Given the sample data, there are a number of alternatives for estimating the MSE of the mean of earnings in equation using the sample data. A standard approach utilizes the Sen-Yates-Grundy estimator for the variance component of the first stage sampling (the first term in brackets of equation) (Cochran, ch. 11, 1977):

$$v\left(\hat{\bar{Y}}\right) = v\left(\hat{R}\right) \approx \frac{1}{\hat{M}^2}\sum_{h=1}^{5}\sum_{i=1}^{n_h}\left\{\sum_{j>i}^{n_h}\left[\frac{\pi_{hi}\pi_{hj} - \pi_{h,ij}}{\pi_{h,ij}}\left(\frac{D'_{hi}}{\pi_{hi}} - \frac{D'_{hj}}{\pi_{hj}}\right)^2\right] + M_{hi}^2\left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}}\right)\frac{s_{d'2hi}^2}{\pi_{hi}}\right\}$$

where $\hat{M} = \hat{M}_{ST} = \sum_{h=1}^{5}\hat{M}_h = \sum_{h=1}^{5}\sum_{i=1}^{n_h}\frac{M_{hi}}{\pi_{hi}}$

$$D'_{hi} = \hat{Y}_{hi} - \hat{R}M_{hi}$$

$$\hat{R} = \frac{\hat{Y}_{ST}}{\hat{M}_{ST}}$$

$$d'_{hij} = y_{hij} - \hat{R}m_{hij}$$

$$s_{d'2hi}^2 = \frac{1}{m_{hi} - 1}\sum_{j=1}^{m_{hi}}\left[\left(y_{hij} - \hat{R}m_{hij}\right) - \left(\bar{y}_{hi} - \hat{R}\bar{m}_{hi}\right)\right]^2$$

$$= \frac{1}{m_{hi} - 1}\sum_{j=1}^{m_{hi}}\left(y_{hij} - \bar{y}_{hi}\right)^2$$

$$= s_{2hi}^2$$

$$\hat{Y}_{hi} = M_{hi}\bar{y}_{hi}$$

$\pi_{h,ij} > 0$.

The variance for the second stage of sampling (the final term in the brackets of equation) is based on the variance for the systematic sample of workers within employer under the assumption that they are drawn from a population in random order (Cochran 1977, pp. 224, 302-309).

Inference regarding the population estimates may be conducted with reference to the normal approximation since the ratio estimate is asymptotically normally distributed. For example, a confidence interval for average earnings may be specified as

$$\hat{\bar{Y}} \pm 2v\left(\hat{\bar{Y}}\right)$$

providing an approximate 95% confidence interval for the mean of earnings.

Weeks of work. A second variable of primary interest is the weeks of farm work for the respondents. One measure is the weeks of farm work for the previous 12 months; a second measure of interest is the average annual weeks of farm work through the worker's career. The estimation procedure for each of these two variables is identical to the procedures presented above for earnings. Wherever the various forms of the *y* and *Y* variables appear in the above equations, they are redefined as 1) weeks of work for the previous 12 months for the worker, and 2) average annual weeks of farm work over the worker's work history since first entering the U.S. for work. All other variables in the equations remain the same, and all estimators remain the same.

Legal status and earnings. In addition to estimating the overall mean of earnings, we are interested in estimating average earnings by the worker's immigration status, referred to as domain means in the sampling literature. The two domains of interest are workers who are either citizens or have an immigration status legally authorizing them to work in the United States as compared to foreign workers who do not have the appropriate authorization for work in the United States. The same series of questions regarding legal status is being asked in the FAWS questionnaire as have been asked in the NAWS questionnaire over the past several years. The series of questions permits the construction of a categorical variable discriminating between authorized workers and unauthorized workers.

The domain means of earnings for the population are $\bar{Y}_k = \frac{Y_k}{M_k}$ for the k^{th} domain where

$$Y_k = \sum_{h=1}^{5} M_{hk}\bar{Y}_{hk}$$

$$M_k = \sum_{h=1}^{5}\sum_{i=1}^{N_h} M_{hik}$$

$$\bar{Y}_{hk} = \sum_{i=1}^{N_{hi}}\sum_{j=1}^{M_{hik}} \frac{y_{hijk}}{M_{hik}}$$

$$M_{hk} = \sum_{i=1}^{N_h} M_{hik}$$

y_{hijk} = earnings of the j^{th} worker of the i^{th} employer in the k^{th} domain in stratum h, and
M_{hik} = number of workers in the k^{th} domain in the i^{th} employer of stratum h.
Corresponding population estimates, assuming a simple random sample for the moment, could be constructed as $\hat{\bar{Y}}_k = \frac{\hat{Y}_k}{M_k}$ where $\hat{Y}_k$ is the estimate of total earnings for all workers in the k^{th} domain, and M_k is the number of workers in the k^{th} domain. There are two major problems with this estimation. The first is the relatively standard problem that the number of workers in the k^{th} domain of the population is unknown, so it would have to be estimated resulting in a ratio estimate for the mean of earnings for the k^{th} domain. More significant, however, is that legal status is self-reported. Recent NAWS findings indicate that over 50% of the workers self-report to be unauthorized for work in the United States (Aguirre 2006, p. 116). Although obtaining information regarding a worker's legal status is a sensitive issue, the interviewers are clearly establishing the workers' confidence to reveal an estimate that is in excess of 50%.[3] Nevertheless, there is good reason to believe that legal status is measured with error. Not only is legal status a sensitive question, but knowledgeable observers in the industry believe the true proportion of unauthorized workers to be much higher than is currently reported in the NAWS. As a categorical variable, this results in misclassification of workers into domains and presents serious problems for the domain ratio estimator for earnings.

Misclassification errors have been considered in the sampling literature, notably by Hansen, Hurwitz and Bershad (1961), Cochran (1968), and summarized in Cochran (1977). Koch (1973) specifically addressed ratio estimators for domain estimation when the domain indicator was subject to error, resulting in misclassification. The result is that with misclassification, there is no way of knowing whether the bias in the *difference* between two ratio estimates is positive or negative. Since it is the difference in the ratio estimates of average earnings for the two legal status domains that is of interest, there is no way of knowing if the estimated difference in earnings between authorized and unauthorized workers is under- or over-estimated, a particularly problematic result.

An alternative approach to characterize the difference between earnings of authorized and unauthorized workers is through a regression model. This offers the advantage that estimation procedures can utilize additional information to statistically correct for the misclassification error in legal status. The standard economic model for earnings (typically attributed to Mincer (1974)) is:

$$Y_i = \delta_0 + \delta_1 s_i + \delta_2 exper_i + \delta_3 exper_i^2 + z_i \delta_x + \varepsilon_i \qquad E[\varepsilon_i] = 0,\ E[\varepsilon_i \varepsilon_j] = \begin{cases} \sigma^2 & for\ i = j \\ 0 & for\ i \neq j \end{cases}$$

where Y_i = natural log of earnings for the i^{th} worker
s_i = years of schooling
$exper_i$ = years of experience
z_i = a set of socio-economic variables to control for variations among workers
ε_i = a random disturbance.

[3] This a primary reason for the decision to employ the same contractor to conduct the FAWS as does the NAWS.

The δ's are parameters to be estimated. Earnings are typically specified in logs since we are most interested in proportionate effects on earnings from changes in explanatory variables. Moreover, there is considerable empirical evidence that the log specification of earnings is most consistent with the data generating process (Heckman and Polachek 1974). Although the original specification of the earnings equation in was to evaluate the returns to education, this general form of the earnings equation has been used for a wide variety of phenomena regarding labor markets, including migration, health, retirement, and legal status among other things.

Our interest is in evaluating differences in earnings for workers authorized to work in the U.S. relative to those who do not have such authorization. For clarity of discussion, we rewrite equation in matrix form:

$$Y = [Z \vdots X] \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \varepsilon$$

where *Z* represents all variables (including the column of ones for the constant term) except legal status, *X* is a dummy variable representing legal status, and *Y* represents the log of earnings; the parameter vector is partitioned correspondingly. We focus primarily on the measurement error associated with the legal status categorical variable (*X*) in this document; we have no reason to believe that the remaining variables (*Z*) will have unusual errors of measurement.

Measurement error of the legal status variable can be characterized as

$$x_i = X_i + u_i$$

with the lower case x_i representing the observed indicator and the upper case X_i representing the true legal status; u_i represents the measurement error. Substituting the measured legal status variable (equation) for the true legal status in equation , we have:

$$Y = [Z \vdots x] \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + (\varepsilon - u\gamma)$$

The new error term for the equation to be estimated becomes $\varepsilon - u\gamma$. For ordinary least squares estimates of β and γ to be unbiased and consistent, the covariance between the explanatory variables (*Z* and *x*) and the error term $\varepsilon - u\gamma$ must be zero.[4] The covariance between the *Z* variables and the equation disturbance (ε) is typically assumed to be zero, and the covariance between true legal status (*X*) and the disturbance (ε) will also be assumed to be zero for the moment. However, it follows directly from the error model in equation that the covariance between the observed legal status, *x*, and the measurement error, *u*, is necessarily non-zero. As a result, the ordinary least squares estimates of β and γ would be biased and inconsistent.

A widely used procedure to obtain consistent parameter estimates in the presence of classical measurement error is instrumental variables. The procedure requires obtaining one or more

---

[4] The ordinary least squares estimator is $\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = ([Z : x]'[Z : x])^{-1}[Z : x]' y$. For additional details regarding ordinary least squares estimates, see any econometrics text such as Greene (2003) or Cameron and Trivedi (2005).

variables that are highly correlated with the true variable (X), but not correlated with the disturbance term, $\varepsilon - u\gamma$. A critical assumption of the classical measurement error model is that the measurement error, u, is uncorrelated with the true value of the variable, X. However, when the problematic variable is binary, the measurement error is necessarily correlated with the true value of the variable. For example, if the true value of X is one and the observation is misclassified, the measured variable must be zero. As a result, u = -1. Correspondingly, if the true value of X is zero and the observation is misclassified, x must be one, and u =1. The only other possible value for u is zero, i.e. when there is no misclassification. Consequently, $Cov(X,u) < 0$. The standard instrumental variables procedure makes use of additional variables W which are correlated with the true variable, X, but uncorrelated with the disturbance $\varepsilon - u\gamma$. When X is binary this is not possible since $Cov(X,u) < 0$. As a result, the standard instrumental variable estimator[5] yields inconsistent parameter estimates and does not resolve the estimation problem.

Aigner (1973) demonstrated an approach utilizing external information when available to correct the parameter estimate for the binary variable when there is misclassification. Another approach to the problem in the recent literature is to utilize available multiple indicators for the true variable (legal status) (Black, et al. (2000), Bound, et al. (2003), and Kane, et al. (1998)). Recognizing that it is rather unusual to have more than one indicator for the same variable in survey data, Frazis and Loewenstein (2003) (hereafter referred to as FL) developed a GMM (generalized method of moments) estimator that is consistent as long as the binary variable is exogenous, i.e. $Cov(X,\varepsilon) = 0$. Although the parameter on the binary variable is not identified when the binary variable is endogenous ($Cov(X,\varepsilon) \neq 0$), they established a bounding procedure for the parameter in that case.

The approach with the most promise for our problem follows the FL GMM estimator. For the moment, suppose there are no other explanatory variables (Z) in equation ; the binary variable X is exogenous and is the only explanatory variable, other than a constant term:

$$Y_i = \gamma_0 + \gamma_1 X_i + e_i$$

With $E[e] = E[X'e] = 0$,

$$E[Y_i \mid X_i = 0] = \gamma_0$$

$$E[Y_i \mid X_i = 1] = \gamma_0 + \gamma_1$$

The estimated value of γ_0 is the estimated mean of log earnings for unauthorized workers (X_i = 0), and the estimated value of $\gamma_0 + \gamma_1$ is the estimated mean of log earnings for authorized workers (X_i = 1).

To examine the effects of measurement error, consider the probabilities of misclassification:

$$\alpha_0 = \Pr[x_i = 1 \mid X_i = 0] = \Pr[u_i = 1 \mid X_i = 0]$$

$$\alpha_1 = \Pr[x_i = 0 \mid X_i = 1] = \Pr[u_i = -1 \mid X_i = 1]$$

---

[5] The standard instrumental variable estimator with a single instrument for the variable measured with error is $\begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} = (W'[Z : x])^{-1} W'Y$ where W=[Z:w] and w is the instrumental variable for x. See any econometrics text such as Greene (2003) or Cameron and Trivedi (2005) for additional details on instrumental variable estimation.

where it is typically assumed that .[6] Then

$$\begin{aligned} E[Y_i \mid x_i = 0] &= E[\gamma_0 + e_i - \gamma_1 u_i \mid x_i = 0] \\ &= \gamma_0 - \gamma_1\{\Pr[X_i = 0]\Pr[x_i = 0 \mid X_i = 0] + \Pr[X_i = 1]\Pr[x_i = 0 \mid X_i = 1]\} \\ &= \gamma_0 - \gamma_1\{(1 - p)(1 - \alpha_0) + p\alpha_1\} \\ &= \gamma_0 - \gamma_1\{(1 - \alpha_0) - p(1 - \alpha_0 - \alpha_1)\} \end{aligned}$$

where $p = \Pr[X_i = 1]$. Similarly,

$$E[Y_i \mid x_i = 1] = \gamma_0 + \gamma_1\{(1 - \alpha_0) - p(1 - \alpha_0 - \alpha_1)\}$$

With $\gamma_1 > 0$, the mean of earnings for unauthorized workers is biased downward by the term $\gamma_1\{(1 - \alpha_0) - p(1 - \alpha_0 - \alpha_1)\}$, and the mean of earnings for authorized workers is biased down by the term $\gamma_1\{\alpha_0 + p(1 - \alpha_0 - \alpha_1)\}$. Most importantly, since the primary interest is in the *difference* in expected earnings, rather than estimating

$$E[Y_i \mid X_i = 1] - E[Y_i \mid X_i = 0] = \gamma_1$$

in the presence of misclassification, we would be estimating

$$\begin{aligned} E[Y_i \mid x_i = 1] - E[Y_i \mid x_i = 0] &= 2\gamma_1[(1 - \alpha_0)(1 - p) + p\alpha_1] \\ &\geq \gamma_1 \quad as\ [(1 - \alpha_0)(1 - p) + p\alpha_1] \geq \tfrac{1}{2} \\ &\leq \gamma_1 \quad as\ [(1 - \alpha_0)(1 - p) + p\alpha_1] \leq \tfrac{1}{2} \end{aligned}$$

In the absence of knowledge of the parameters p, α_0 and α_1, there is no way of knowing if γ_1 would be over or under estimated. To properly estimate γ_1 requires a procedure which can identify the true probability of being authorized for work (p) and the misclassification probabilities (α_0 and α_1).

The FL procedure augments an instrumental variable estimator for the parameters with additional moment equations to consistently estimate γ_1, and to identify the p, α_0 and α_1 parameters. In the context of the data to be collected pertaining to legal status, there are two potential candidates as instruments for self-reported legal status. Given the stigma attached to being unauthorized for work, it is anticipated that the legal status variable is subject to error, particularly for persons whose true status is unauthorized for work. Potential instrumental variables are based on the first year of entry to the U.S. Given the successive changes in immigration laws and regulations, early entrants are much more likely to have attained legal status than are recent entrants who face increased barriers to attaining legal status. A significant tightening of immigration regulations occurred with IRCA in 1986, and another significant shift occurred following the 2001 tragedy. These two events can be defined as two dummy variables: one for entry after 1987, and a second for entry after 2001. With these two instruments, the FL procedure can generate consistent estimates of the legal status effect, or in the worst case scenario, provide upper and lower bounds to the legal status parameter while formally accounting for the misclassification errors.

Equation was a simplified version of equation for the purpose of clarifying the misclassification implications. Equation is the appropriate specification; in the absence of the *Z* variables, the necessary requirement that $E[X'e] = 0$ is unlikely to be satisfied. The *X* variable representing legal status is likely to be correlated with the omitted *Z* variables such as education, experience

[6] The probability of an authorized worker being reported as an unauthorized worker, α_1, is likely to be small, if not zero. We expect $\alpha_0 > \alpha_1$.

and other socio-economic variables. Since the latter variables are important and theoretically relevant determinants of earnings, their omission would imply correlation between the error term and the legal status variable, resulting in least squares bias due to omitted variables. Including the *Z* variables results in no additional estimation complications as long as the measurement errors are not correlated with the *Z* variables, and there is little reason to think that they would be correlated. The only change in interpretation of the estimation of mean earnings for the authorized and unauthorized workers is that it is now conditional on *Z*. Given the parameter estimates, values are specified for the Z variables (such as the means for the authorized or unauthorized group) and the conditional means of earnings for authorized and unauthorized workers are directly calculated.

Under the assumption that the categorical variable, *X*, and the *Z* variables are exogenous, FL show that a GMM estimator using appropriate instruments will provide consistent estimates of the γ parameter in equation , the effect of legal status on earnings conditionally on the *Z* variables and under a misclassification model such as equations . Following FL, the method of moments estimator is developed based on moment conditions which must be satisfied in estimation.[7] First, the instrumental variable estimation is based on the moment conditions following from equation :

$$Cov(Z,Y) = Var(Z)\beta + Cov(Z,X)\gamma$$
$$Cov(W,Y) = Cov(W,Z)\beta + Cov(W,X)\gamma$$

where *W* is the set of instrumental variables,[8] in our case the two dummy variables based on year of first entry to the U.S.

$$W_1 = \begin{cases} 1 \; if \;\; entry \; was \; after \; 1986 \; and \; before \; 2002 \\ 0 \; otherwise \end{cases}$$

$$W_2 = \begin{cases} 1 \; if \;\; entry \; was \; after \; 2001 \\ 0 \; otherwise \end{cases}$$

Since true legal status, *X*, is not observable, *Cov*(*Z*,*x*) and *Cov*(*W*,*x*) need to be substituted in equations for their unobservable counterparts. It can be shown (FL) that

$$\frac{Cov(Z,X)}{Cov(Z,x)} = \frac{Cov(W,X)}{Cov(W,x)} = \frac{1}{1 - \alpha_0 - \alpha_1} \equiv k_1$$

Substituting into equation results in

$$Cov(Z,Y) = Var(Z)\beta + Cov(Z,x)k_1\gamma$$
$$Cov(W,Y) = Cov(W,Z)\beta + Cov(W,x)k_1\gamma$$

which are observable, but yield a parameter $\gamma_1 \equiv k_1\gamma$ rather than γ, hence the inconsistency of the standard instrumental variable estimator. The first two moment equations in terms of the sample data are therefore

---

[7] The interested reader is referred to FL (pp. 159-161) for further details of the estimator and proofs of consistency. A general treatment of methods of moments estimators may be found in any recent econometrics text such as Greene (2003) or Cameron and Trivedi (2005).

[8] The assumptions for the instruments *W* are that $Cov(W,\varepsilon) = 0$ and $Cov(W,u \mid X) = 0$, although with *X* binary, $Cov(W,u) \neq 0$ since $Cov(W,X) \neq 0$ (FL).

2/2/2021

$$m_1(\beta,\gamma_1) = z'y - z'z\beta - z'x\gamma_1 = 0$$
$$m_2(\beta,\gamma_1) = w'y - w'z\beta - w'x\gamma_1 = 0$$

where the lower case letters represent deviations from the means for the variables.

FL then introduce two additional moment equations based on *Cov*(*x*,*Y*) and *Cov*(*T*,*Y*) where $T = (W - \bar{W}) \cdot x$. These are based on the observation by Black, et al. (2000) that the observations where the observed variable *x* and the alternative indicators in their approach give conflicting results are informative and can lead to identification of the error parameters. The covariances of the legal status variable and the instruments with *Y* are the counterpart in the FL approach with instrumental variables. This establishes the following two equations based on equation :

$$Cov(x,Y) = Cov(x,Z)\beta + Cov(x,X)\gamma$$
$$Cov(T,Y) = Cov(T,Z)\beta + Cov(T,X)\gamma$$

Again, the substitution from

$$\frac{Cov(X,x)}{Var(x)} \equiv k_2$$
$$\frac{Cov(T,X)}{Cov(T,x)} \equiv k_3$$

into equations results in the covariances for the observables:

$$Cov(x,Y) = Cov(x,Z)\beta + Var(x)k_2\gamma$$
$$Cov(T,Y) = Cov(T,Z)\beta + Cov(T,x)k_3\gamma$$

The corresponding moment equations for the sample data in deviations from the mean are:

$$m_3(\beta,\gamma_2) = x'y - x'z\beta - x'x\gamma_2 = 0$$
$$m_4(\beta,\gamma_3) = t'y - t'z\beta - t'x\gamma_3 = 0$$

where $\gamma_2 \equiv k_2\gamma$ and $\gamma_3 \equiv k_3\gamma$. The system is closed with $E[x] = \tilde{p}$, resulting in

$$m_5(p) = \bar{x} - p = 0$$

Following FL, the GMM estimator can be specified after defining the following matrices:

$$H \equiv \begin{bmatrix} w & z & 0 & 0 & 0 \\ 0 & 0 & x & 0 & 0 \\ 0 & 0 & 0 & t & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad G \equiv \begin{bmatrix} y \\ y \\ y \\ x \end{bmatrix} \quad F \equiv \begin{bmatrix} x & 0 & 0 & z & 0 \\ 0 & x & 0 & z & 0 \\ 0 & 0 & x & z & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \Pi \equiv \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \beta \\ p \end{bmatrix}$$

where *all* lower case variables (*y*, *x*, *z*, *w*, *t*) are in deviations from the means. The original moment equations in , , and can then be written in matrix form as:

$$\bar{m} \equiv (1/n)\, H'[G - F\Pi]$$

The GMM estimator is the Π that minimizes the quadratic form $\bar{m}'R\bar{m}$ (Hansen 1982). The matrix *R* is a weighting matrix, and is typically taken to be the identity matrix for the first round of estimation. The resulting parameter estimates, $\hat{\Pi}$, are consistent estimates of Π.

As FL show, k_2 and k_3 are functions of *p*, α_0, and α_1:

2/2/2021

$$k_2 = \frac{(p - \alpha_0)(1 - p - \alpha_1)}{p(1 - p)(1 - \alpha_0 - \alpha_1)}$$

$$k_3 = \frac{(1 - p - \alpha_1) + \alpha_0}{(1 - p)(1 - \alpha_0 - \alpha_1)}$$

Solutions can then be obtained for γ, α_0, and α_1 as

$$\hat{\gamma} = \sqrt{4\hat{p}(1 - \hat{p})\hat{\gamma}_1\hat{\gamma}_2 + ((1 - \hat{p})\hat{\gamma}_3 - \hat{p}\hat{\gamma}_1)^2}$$

$$\hat{\alpha}_0 = \frac{\hat{p}\hat{\gamma}_1 + (1 - \hat{p})\hat{\gamma}_3 - \hat{\gamma}}{2\hat{\gamma}_1}$$

$$\hat{\alpha}_1 = \frac{(2 - \hat{p})\hat{\gamma}_1 - (1 - \hat{p})\hat{\gamma}_3 - \hat{\gamma}}{2\hat{\gamma}_1}$$

Since $\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\beta}, \text{and } \hat{p}$ are consistent estimates as the GMM estimates, it follows that the solutions for $\hat{\gamma}, \hat{\alpha}_0, \text{and } \hat{\alpha}_1$ are also consistent estimates. FL then specify the optimal GMM estimator while accounting for heteroskedasticity.

The preceding development of the GMM estimator assumes the data are from a simple random sample rather than a complex sample as in our case. There are two primary considerations: weighting of the observations and the correlation of responses within employer. Whether or not to weight the observations is typically dictated by the purpose of the analysis. In analyses such as ours where the purpose is to characterize the underlying population, weighting by the square root of the inverse of the sampling rate for each observation is the usual case (Cameron and Trivedi (2005), Deaton (1997)): $\upsilon_{hij} = 1/\sqrt{f_{0hij}}$ and $f_{0hij} = \pi_{hi} m_{hi} / M_{hi}$.

The complex sampling structure can be introduced into the basic model reflected in equation by adding cluster effects in the following way:

$$Y_{hij} = \mu_{hi} + Z_{hij}'\beta + x_{hij}\gamma + \zeta_{hij}$$

where $\zeta_{hij} = \varepsilon_{hij} - \gamma u_{hij}$. The new parameter, μ_{hi}, represents the employer (cluster) effect. A fixed effects model is assumed since the effects are likely to be correlated with the disturbance ζ_{hij}. A convenient way to estimate the fixed effects model is to convert the original variables to deviations from the cluster means, implementing a within-clusters estimator. This eliminates the employer effects (μ_{hi}) from the model for estimation, although they can be recovered after estimation if they are needed.

Given a complex sample, we first weight the variables using υ_{hij}:

$$Y^0_{hij} = \upsilon_{hij} Y_{hij}$$

$$Z^0_{hij} = \upsilon_{hij} Z_{hij}$$

$$x^0_{hij} = \upsilon_{hij} x_{hij}$$

Converting to deviations from cluster means, equation becomes

$$Y^0_{hij} - \bar{Y}^0_{hi} = (Z^0_{hij} - \bar{Z}^0_{hi})'\beta + (x^0_{hij} - \bar{x}^0_{hi})\gamma + \zeta^0_{hij} - \bar{\zeta}^0_{hi}$$

2/2/2021

Representing the deviations from the cluster means as $\tilde{Y}_{hij} = Y^0_{hij} - \bar{Y}^0_{hi}$, for example, the equation is:

$$\tilde{Y}_{hij} = \tilde{Z}_{hij}'\beta + \tilde{x}_{hij}\gamma + \tilde{\xi}_{hij}$$

The variables are now in a form suitable for implementation in the above GMM procedure. After transforming the above variables again to deviations from the overall means, substitute into the matrix definitions . For clarity, the variables are defined as

$$\ddot{y}_{hij} = \tilde{Y}_{hij} - \bar{Y}^0_{...}$$
$$\ddot{z}_{khij} = \tilde{Z}_{khij} - \bar{Z}^0_{k...}$$
$$\ddot{x}_{hij} = \tilde{x}_{hij} - \bar{x}^0_{...}$$
$$\ddot{w}_{hij} = \tilde{W}_{hij} - \bar{W}^0_{...}$$
$$\ddot{t}_{hij} = \tilde{T}_{hij} - \bar{T}^0_{...}$$

The sample moment equations are as before, but with the above transformations of the variables:

$$m_1(\gamma_1, \beta) = \ddot{w}'\ddot{y} - \ddot{w}'\ddot{z}\beta - \ddot{w}'\ddot{x}\gamma_1 = 0$$
$$m_2(\gamma_1, \beta) = \ddot{z}'\ddot{y} - \ddot{z}'\ddot{z}\beta - \ddot{z}'\ddot{x}\gamma_1 = 0$$
$$m_3(\gamma_2, \beta) = \ddot{x}'\ddot{y} - \ddot{x}'\ddot{z}\beta - \ddot{x}'\ddot{x}\gamma_2 = 0$$
$$m_4(\gamma_3, \beta) = \ddot{t}'\ddot{y} - \ddot{t}'\ddot{z}\beta - \ddot{t}'\ddot{x}\gamma_3 = 0$$
$$m_5(p) = \bar{\ddot{x}} - p = 0$$

The GMM estimator is $\tilde{m}'R\tilde{m}$ where $\tilde{m}$ is the vector of moment equations , and *R* is the weighting matrix. The optimal GMM estimator sets $R = V^{-1}$, where V^{-1} is the inverse of the covariance matrix of the moment equations (Hansen 1982). The procedure is to first set $R = I$, the identity matrix, to obtain a set of consistent estimates of the parameters, $\hat{\Pi}$, using these to form a consistent estimate of *V* and again solve the moment equations with $R = \hat{V}^{-1}$.

The variance estimate used by FL assumes the data are from a simple random sample. With clustering as in our complex sample design, the variance estimator must be modified slightly. Following Cameron and Trivedi (2005), the covariance matrix *V* can be estimated consistently in the following way. The residuals, $\hat{\xi}_{hij}$, are defined as: $\hat{\xi} = \ddot{G} - \ddot{F}\hat{\Pi}$ where $\hat{\Pi}$ is the set of consistent estimates resulting from the first round of GMM estimation based on moment equations .[9] The residuals are pre-multiplied by the $\ddot{H}'$ matrix as in FL and Wooldridge (1996) to correct for the presence of different instruments in each of the moment equations:

$$\tilde{\xi}_{hij} = \ddot{H}_{hij}'(\ddot{G}_{hij} - \ddot{F}_{hij}\hat{\Pi}) = \ddot{H}_{hij}'\hat{\xi}_{hij}$$

The covariance matrix $\hat{V} = \tilde{\xi}\tilde{\xi}'$ is an L×L matrix where L is the sum of the number of variables in the *w*, *z*, *x*, and *t* matrices. The typical element of the $\hat{V}$ matrix is (adapting from Cameron and Trivedi (2005), pp. 854-856):

$$\hat{v}_{ab} = \sum_{h=1}^{5} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\bar{\tilde{\xi}}_{a,hi} - \bar{\tilde{\xi}}_{a,h})(\bar{\tilde{\xi}}_{b,hi} - \bar{\tilde{\xi}}_{b,h}) \qquad \text{for } a, b = 1, ..., L$$

[9] Matrices such as $\ddot{G}$ are defined in the same way as in equation except that the variables have been redefined to be deviations from the means of the cluster means of the weighted variables.

2/2/2021

where

$$\left.\begin{aligned}\tilde{\xi}_{l,hi} &= \sum_{j=1}^{m_{hi}} \tilde{\xi}_{l,hij} \\ \bar{\tilde{\xi}}_{l,h} &= \frac{1}{n_h}\sum_{i=1}^{n_h} \tilde{\xi}_{l,hi}\end{aligned}\right\} \text{for } l = 1,\ldots,L$$

This assumes that the residuals are independent across strata and iid between employers, but permits correlation between employees within employer as well as heteroskedasticity. Minimizing $\tilde{m}'\hat{V}^{-1}\tilde{m}$ with respect to the parameters, Π, the solution is

$$\tilde{\Pi} = \left[\ddot{F}'\ddot{H}\hat{V}^{-1}\ddot{H}'\ddot{F}\right]^{-1}\ddot{F}'\ddot{H}\hat{V}^{-1}\ddot{H}'\ddot{G}$$

Under fairly general conditions, the optimal GMM estimator has an asymptotic normal distribution with mean Π and covariance matrix $(\ddot{F}'\ddot{H}V^{-1}\ddot{H}'\ddot{F})^{-1}$ (Hansen 1982). The estimated covariance matrix is $(\ddot{F}'\ddot{H}\tilde{V}^{-1}\ddot{H}'\ddot{F})^{-1}$ where $\tilde{V}^{-1}$ is based on the optimal GMM estimates $\tilde{\Pi}$.

Recall that the underlying parameter of most interest is γ, the parameter on the legal status variable. The estimate of γ is obtained indirectly from functions of $\gamma_1, \gamma_2, \gamma_3, \text{and } p$ as indicated in equation . Using the delta method, the estimated covariance matrix for $\tilde{\theta} = (\tilde{\gamma}, \tilde{\beta}, \tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{p})$ is

$$\tilde{V}(\tilde{\theta}) = \tilde{D}'(\ddot{F}'\ddot{H}\tilde{V}^{-1}\ddot{H}'\ddot{F})^{-1}\tilde{D}$$

where

$$\tilde{D} = \left.\frac{\partial\theta}{\partial\Pi'}\right|_{\tilde{\Pi}}$$

where the functional relationships between θ and Π are specified in equations for γ, α_0, and α_1, and β and p are estimated directly as part of Π.

Reflecting back to equation , $\tilde{\gamma}$ is the estimate of the percentage increase in earnings for authorized workers relative to unauthorized workers, holding all other characteristics constant as specified in the earnings equation. Since the parameter estimate is asymptotically normally distributed, standard hypothesis tests and confidence intervals can be applied based on the asymptotic normal distribution. The test of most interest is based on the statistic:

$$\frac{\tilde{\gamma}}{\sqrt{\tilde{V}_{11}}} \sim N(0,1) \qquad \begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma \neq 0 \end{cases}$$

where $\tilde{V}_{11}$ is the estimated variance of $\tilde{\gamma}$, i.e. the 1,1 element of $\tilde{V}(\tilde{\theta})$ in equation .

The consistency of the above GMM estimates requires zero correlation between the true legal status variable and the disturbance term ε. If legal status is endogenous to the model (jointly determined with earnings), then it is correlated with ε. The result will be inconsistent estimates of the γ parameter, although the β parameter estimates remain consistent (FL). This is a possibility that at least needs to be considered. Previous work using the NAWS data (Isé and Perloff (1995) and Iwai, et al. (2006)) has modeled the legal status variable as endogenous in the context of the earnings equation, although misclassification of legal status has not been considered. A test for consistency of the model with the data generating process under the

2/2/2021

assumption that the model has been properly specified, including both measurement error and the exogeneity of legal status can be conducted based on the above GMM estimates (FL). The test statistic is based on $L \geq 0$ where $L \equiv [A_0 - \alpha_0, A_1 - \alpha_1, \alpha_0, \alpha_1]$ and where $A_i = p\lim(\hat{A}^i(\hat{\delta}, \hat{\kappa}_q))$. The $\hat{A}^0$ and $\hat{A}^1$ are empirically estimated upper bounds for α_0 and α_1.[10] Define a restricted estimate of L, $\hat{L}^R$, as the boundary values of zero for any of the four unrestricted parameter estimates that are less than zero, and the remaining parameters are averages of random draws from the distribution of unconstrained estimates following Geweke's (1986) Bayesian procedure. Then set $\eta = (\hat{L}^R - \hat{L})$, and η_A as the subvector of η with any binding constraints. FL show that the test statistic

$$\tau = \eta_A' V_A(\hat{L})^{-1} \eta_A \sim \chi^2(2)$$

where $V_A(\hat{L})$ is the submatrix of the covariance matrix of $\hat{L}$ corresponding to the elements in η_A. With a non-statistically significant τ, we could proceed with inference regarding the difference between earnings of authorized and unauthorized workers as measured by the estimate $\tilde{\gamma}$ with assurance that it is a statistically consistent estimate.

Alternatively, with sufficiently large values of τ relative to the $\chi^2(2)$, the underlying model of measurement error with exogeneity of legal status would be rejected. An alternative is a specification permitting the legal status variable to be endogenous. One option is to adopt the typical switching regression model with legal status controlling the regime to which an observation belongs: authorized or unauthorized. Existing specifications of the labor market using NAWS data take this approach in the absence of controlling for misclassification of legal status (Isé and Perloff (1995) and Iwai, et al. (2006)). Dye and McMillen's (2007) recent work on urban renewal employs an endogenous switching model with misclassification of the switching variable resulting in consistent estimation of all parameters in the presence of misclassification in addition to the classification variable being endogenous. However, a disadvantage of the switching regression approach is that most implementations use a parametric model assuming a multivariate normal distribution for the switching equation and the regression equation disturbances.

An alternative in the presence of endogeneity of the classification variable (legal status) suggested by FL is based on their parameter bounding procedure. Although the procedure does not permit consistent estimation of the γ parameter, it does provide upper and lower bounds to the unknown parameter. The advantage is that it requires much weaker distributional assumptions, i.e. normality is not required. Maximum values are obtained for α_0 and α_1 from the $\hat{A}^0$ and $\hat{A}^1$ specified above. They demonstrate that the bounds for γ are

$$p\lim \hat{\gamma}_{IV}(1 - \alpha_0^{\max} - \alpha_1^{\max}) \leq \gamma \leq p\lim \hat{\gamma}_{IV}$$

where $\hat{\gamma}_{IV}$ is the instrumental variable estimate of γ and that the remaining instrumental variable parameter estimates are consistent.

[10] The parameter estimates $\hat{\delta}$ are from a quasi-maximum-likelihood estimation of $G(Z,\delta)$ where $E[x|Z] = G(Z,\delta)$ and $\hat{\kappa}_q$ is the estimated q-quantile for the cdf of $G(Z,\delta)$. The interested reader is referred to FL for details.

2/2/2021

Unusual Problems

There are no unusual problems requiring specialized sampling procedures.

Frequency

Each respondent will be interviewed once.

3. Statistical Reliability

Anticipated Response

As indicated in section B.1.c, we anticipate an employer (stage 1) response rate of no less than 75%, and a worker response rate (stage 2) of 90%. Our basis for assuming these response rates is from the NAWS experience conducting a similar survey. We have deliberately chosen to employ the same firm to conduct the survey for their expertise and experience in surveying this population. Moreover, the questionnaire is very similar to the NAWS questionnaire. As indicated in section A.1, the survey is designed to obtain detailed, accurate information specific to the research questions for our project.

We have good reason to believe that the employer response rate will be higher for the current survey than for the NAWS. As the name suggests, the NAWS is a national survey with minimal local identification to the organization conducting the survey. The proposed survey is to be conducted only in Florida, and is under the direction of faculty from the University of Florida. Given the long history of cooperation between the Florida agricultural industry and the University of Florida as the state's land grant institution operating programs for the benefit of the agricultural and rural population through the agricultural experiment station and the cooperative extension service, we anticipate an employer response rate no less than has been achieved at the national level, and most likely considerably higher. The worker response rate of 90% achieved at the national level is a very respectable response rate. With the experienced interview staff, we anticipate a very similar rate in Florida.

Methods to maximize response rate

Groves and Couper (1998) and Dillman et al. (2002) identify two sets of influences on survey response under the control of the survey design: "(a) survey protocols … and (b) the selection and training of interviewers." (Dillman et al. 2002, p. 8) Primary attention is addressed to maximizing employer response as the first point of contact to locate workers for interview, and since this has had the lower response rate with the NAWS. Dillman et al. (2002) identify a number of factors regarding the survey design that can influence the response rate and are applicable to the employer contact for this survey.

- Agency of data collection. Dillman et al. (2002) note that "Sample persons' knowledge and attitudes concerning the sponsor can affect whether they grant an interview …" (p. 11) As noted above, the survey is under the direction of the University of Florida which

2/2/2021

has a long history of cooperation with the Florida agricultural industry. Advance material provided to the industry and to employers selected to be sampled will be on University of Florida stationery and will emphasize that the University of Florida is the organization operating the survey in conjunction with the U.S. Department of Agriculture. Both are familiar organizations to the industry.

- Advance warning of the survey request. Dillman et al. (2002) highlight a number of pertinent features that will be followed for employers selected for the sample. As noted above, each will be mailed a letter on official University of Florida stationery indicating that the survey is sanctioned by the University of Florida, and granting authenticity to the survey. The letter will highlight benefits to the industry following from the survey and the research to be conducted with the data. Confidentiality assurances for any information collected from their workers will also be highlighted. Most importantly, the letter will emphasize that the only information being requested from the employer is the number and listing of employees for the purpose of constructing the sampling frame. The listing is to be done in the presence of the employer and will not be taken from the employment site or kept by the interviewers. Once the frame is constructed and the worker sample selected, the only remaining role of the employer is to permit the interviewer to talk briefly with the selected workers to arrange a time and location for the formal interview. The letter will provide contact information for the UF Principal Investigator for the Florida Agricultural Workers Survey.
- Call scheduling. Subsequent to the advance mailing, interviewers are to call the employer in advance to arrange an interview time convenient for the employer. They will be instructed to follow-up with telephone requests if the initial request does not result in a scheduled interview.
- Locating the appropriate interviewee. Interviewers must be attuned to locating the person in the firm who can provide the requested information. Trained and experienced interviewers are key to this process.
- Respondent incentives. The incentives to employer respondents are only of an indirect nature. However, advance material will emphasize the importance of the research to the industry and labor market participants, and the importance of their role in cooperating to assist in generating the best possible information on which to base the research.
- Follow-up procedures. Dillman et al. (2002) suggest standard follow-up procedures that will be followed in the survey. If the initial contact indicates they may not be willing to participate, a second more persuasive letter with more detailed information will be sent to the prospective respondent. Should there be problems with a particular interviewer, an alternate interviewer will be sent to try to salvage the interview.
- Interviewer training. We have made the deliberate choice to work with the same group that conducts the NAWS survey so that the interviewers are familiar with the survey, and that they have extensive previous experience with contacting agricultural employers and interviewing farm workers. Training will be provided regarding the minor variations between our survey protocol and the NAWS protocol.
- Interviewer workload. Every effort will be taken to maintain reasonable workloads for the interviewers. Overburdened interviewers are unable to take the necessary time to gain the cooperation of the employer and successfully complete the interview.

2/2/2021

Although we expect the worker response rate to be higher than the employer response rate, every effort will be extended to assure that the worker response rate is as high as possible. Many of the same principles apply to the worker interviews as to the employer interviews. The following procedures are highlighted specifically in the context of maximizing the worker response rate.

- Advance information. Information regarding the worker survey will be distributed in advance to organizations working with farm workers. The purpose is to familiarize the farm worker community with the survey, its purpose, and procedures so that they in turn can encourage the farm worker community to participate. They can also reassure any inquiring workers that it is a legitimate survey and that any information they provide is held in confidence, used only for research purposes. Contact information will be provided for the UF Principal Investigator for the Florida Agricultural Workers Survey.
- Trained interviewers. The group conducting the interviews is to be the same group that conducts the NAWS, and is thus experienced with interviewing farm workers, and is experienced with the questionnaire. The experience and training is essential in gaining the cooperation and trust of the farm worker. The firm provides training to address any subtle differences between the NAWS and the FAWS.
- Language. Most farm workers either are not fluent in English, or do not speak it at all. Interviews are to be conducted in the worker's native language, primarily Spanish.
- Scheduled time for interview. Workers are randomly selected for interview at the time of the employer interview. Once the workers have been selected, interviewers are to meet briefly with the selected workers at the employment site to gain the worker's cooperation to participate and to arrange an amenable time and place for the worker interview.
- Interview location. Interviews are to be conducted in a neutral location away from the employment site. The intent is to allay any worker concerns that the employer may be listening-in on the interview. Most are expected to be conducted at the worker's residence, but again in a location offering privacy from others who may be nearby.
- Incentives. Workers are to be offered a $15 honorarium to compensate them for the time spent for the interview. The intention is to encourage their participation while at the same time communicating to them in this way that it is a serious effort to obtain valid information from them. They will also be informed of the potential contribution to the general well-being of farm workers resulting from research based on the data.
- Confidentiality pledge. Workers are to be given a statement of the confidentiality pledge regarding any information collected from them.

Days per Week Weight Adjustment

Before addressing the nonresponse adjustment, worker weights must be adjusted for their likelihood of being available for sampling on any given day of the week. The probability sampling is designed to give every farm worker a known probability of being sampled. Although sampling rates differ among strata, the sample is designed so that each farm worker has an equal probability of being sampled. Part time work is not uncommon in agriculture; the NAWS reports from 12-17% of workers reported fewer than five days of work per week for the years 1997-2002 (Aguirre p. 60). Any part time workers not working on the day of the survey will not be identified in the sample. By contrast, full time workers are always identified regardless of the day of interview. Consequently, a weighting adjustment is necessary to

2/2/2021

properly represent workers who work less than a full week. The weight adjustment is to be done as with the NAWS (U.S. DOL). Once the data are collected, the days of work reported by the worker are to be used to adjust the weight. Workers reporting six or seven days of work per week receive a weight of one; workers with fewer than six days receive a larger weight. The week weight is

$$w_{hij}^{d} = \frac{6}{\min(6, days\ of\ work\ per\ week)}$$

Information for each worker is multiplied by w_{hij}^{d} to adjust for the probability of being included in the sample given the days worked per week.

Nonresponse Adjustment

Consider the estimation of the population mean earnings in the presence of non-response as specified in equation earlier. As a ratio estimate, we focus on the numerator and denominator separately. The numerator is estimated total earnings specified in equation along with the denominator, the estimated population of workers. Since each of these is the sum of the individual stratum estimates, the procedure is illustrated for an individual stratum.

Estimated total earnings for stratum *h* are

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{w_{hij}^{d} y_{hij}}{\pi_{hi} p_{hi}}$$

where p_{hi} represents the sampling rate for workers in the i^{th} employer, π_{hi} is the inclusion probability for the i^{th} employer, and w_{hij}^{d} is the days of the week adjustment from equation . Define the weights as $w_{hi} = w_{hij}^{d} / (\pi_{hi} p_{hi})$ with

$$w_{hij}^{w} = \frac{w_{hij}^{d}}{p_{hi}}$$

$$w_{hi}^{e} = \frac{1}{\pi_{hi}}$$

for the worker and employer weights, respectively, so that $w_{hij} = w_{hij}^{w} w_{hi}^{e}$. With complete response, the estimate of total earnings for stratum *h* may be represented as

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij}^{w} w_{hi}^{e} y_{hij} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij} y_{hij}$$

To focus on employer response, rewrite equation as

$$\hat{Y}_h = \sum_{i=1}^{n_h} w_{hi}^{e} \hat{Y}_{hi}$$

$$\hat{Y}_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}^{w} y_{hij}$$

With employer non-response, the weights w_{hi}^{e} are too small and require adjustment. The adjustment approach to be used is a calibration approach with what Särndal and Lundström

2/2/2021

(2005 p. 59) refer to as *standard weighting*. The approach is to find a set of alternative weights defining

$$\hat{Y}_h^c = \sum_{i=1}^{n_h^r} \omega_{hi}^e \hat{Y}_{hi}$$

where n_h^r represents the number of employer respondents in stratum *h*. The basic calibration equation is

$$\sum_{i=1}^{n_h^r} \omega_{hi}^e x_{hi} = X$$

The x_{hi} are the individual employer values of externally available information, and the X is the sum of all such information over the universe. Available data to use from the QCEW to characterize the employers and their employees are mean wage for each employer, employment seasonality (percentage that the minimum monthly employment is of the maximum monthly employment over the past year), and employer size as measured by annual payroll.

A linear specification is to determine weight adjustments, v_{hi}^e, such that $\omega_{hi}^e = w_{hi}^e \times v_{hi}^e$. Särndal and Lundström (2005 pp. 57-59) suggest $v_{hi}^e = 1 + \lambda' x_{hi}$ where λ is an unknown j-element vector, the same dimension as the x_{hi} vector, i.e. corresponding to the number of information items defining x_{hi}; three items were suggested above for the current problem. After substitution of ω_{hi}^e and v_{hi}^e in equation , λ is found as

$$\lambda' = (X - \sum_{i=1}^{n_h^r} w_{hi}^e x_{hi})' (\sum_{i=1}^{n_h^r} w_{hi}^e x_{hi} x_{hi}')^{-1}$$

The resulting non-response adjusted estimator for total earnings in stratum *h* may be written as

$$\begin{aligned}
\hat{Y}_h^c &= \sum_{i=1}^{n_h^r} w_{hi}^e (1 + (X - \sum_{i=1}^{n_h^r} w_{hi}^e x_{hi})' (\sum_{i=1}^{n_h^r} w_{hi}^e x_{hi} x_{hi}')^{-1} x_{hi}) \hat{Y}_{hi} \\
&= \sum_{i=1}^{n_h^r} w_{hi}^e (1 + \lambda' x_{hi}) \hat{Y}_{hi} \\
&= \sum_{i=1}^{n_h^r} w_{hi}^e v_{hi}^e \hat{Y}_{hi} \\
&= \sum_{i=1}^{n_h^r} \omega_{hi}^e \hat{Y}_{hi}
\end{aligned}$$

The estimated population total earnings across all strata is obtained in the usual way as

$$\hat{Y}_{ST}^c = \sum_{h=1}^{5} \hat{Y}_h^c$$

A simpler adjustment is made for worker nonresponse since it is not expected to be as significant a problem. As with employer nonresponse, the weights w_{hij}^w are too small in the presence of nonresponse. We assume completely ignorable nonresponse and inflate the weights at the employer level to adjust for nonresponse. With an intended employer sample of m_{hi} workers, let the number of responding workers be m_{hi}^r. Then the adjusted weights are

2/2/2021

$$\omega_{hij}^{w} = w_{hij}^{w} \frac{m_{hi}}{m_{hi}^{r}}$$

Substituting this into the second line of equations , estimated worker earnings for the i^{th} employer are

$$\hat{Y}_{hi}^{c} = \sum_{j=1}^{m_{hi}^{r}} \omega_{hij}^{w} y_{hij}$$

Combining both the worker and employer adjustments, the estimated total earnings for stratum *h* are

$$\hat{Y}_{h}^{c} = \sum_{i=1}^{n_{h}^{r}} \sum_{j=1}^{m_{hi}^{r}} \omega_{hi}^{e} \omega_{hij}^{w} y_{hij}$$

The denominator for estimating mean earnings in equation requires similar adjustment. Recall that the estimate for stratum *h* with full response as specified in equation is

$$\hat{M}_{h} = \sum_{i=1}^{n_{h}} \frac{M_{hi}}{\pi_{hi}}$$

The days of week adjustment is also necessary for the variable M_{hi}. The procedure is to weight the i^{th} employer's employment by the sum of the days of week weights divided by the sample size for the i^{th} employer so the initial weight becomes:

$$w_{hi}^{m} = \frac{1}{\pi_{hi}} \sum_{j=1}^{m_{hi}^{r}} \frac{w_{hij}^{d}}{m_{hi}^{r}}$$

The revised estimate of M_h is then

$$\hat{M}_{h} = \sum_{i=1}^{n_{h}} w_{hi}^{m} M_{hi}$$

Applying the same methodology as in equations - , a new set of weights is obtained yielding a corrected estimate of total employment

$$\hat{M}_{h}^{c} = \sum_{i=1}^{n_{h}^{r}} \omega_{hi}^{m} M_{hi}$$

$$\hat{M}_{ST}^{c} = \sum_{h=1}^{5} \hat{M}_{h}^{c}$$

The most reasonable external information (*X* in equation) to use to define the new weights is the maximum monthly employment data available from the QCEW data.[11] With the revised estimates for the number of workers, the corrected population estimate for mean earnings is

$$\hat{\bar{Y}}^{c} = \frac{\hat{Y}_{ST}^{c}}{\hat{M}_{ST}^{c}}$$

4. Tests

[11] This differs from the collected data for M_{hi} due to the time lag in availability of the QCEW data.

The questionnaires to be used in the survey are fundamentally the NAWS questionnaires developed by the Department of Labor. The modification is simply extending the work history (already a part of the NAWS questionnaire (see A. 1. above)) for a longer period. The modifications have been prepared by UF in consultation with USDA/RMA and the Department of Labor. Selected previously OMB approved NAWS questionnaires are included in Appendix B. Pilot tests of the extended work history will be conducted to test the procedures for this modified segment of the data collection. The pilot test will consist of nine or fewer workers interviewed.

5. Statistical Consultation

Malay Ghosh, Distinguished Professor of Statistics, University of Florida, (352) 273-2992 has been consulted extensively on the statistical aspects of the survey design. The adopted questionnaire closely follows the NAWS survey conducted annually by the Department of Labor. Daniel Carroll, Employment and Training Administration, U.S. Department of Labor, (202) 693-2795, has been consulted extensively on the survey design. The following individuals were consulted by the Department of Labor on the statistical aspects of the NAWS survey design:; Stephen Reder and Robert Fountain, Statisticians, Portand State University, (503) 725-3999 and (503) 725-5204; Philip Martin, Professor, University of California, Davis, (916) 752-1530; Jeff Perloff, Professor, University of California, Berkeley, (510) 642-9574; and John Eltinge, the Bureau of Labor Statistics (BLS) Office of Survey Methods, (202) 691-7404.

The data will be collected under the auspices of the Partnership Agreement between UF and USDA/RMA through a contract with Aguirre International, (650) 373-4900. Aguirre International is the entity conducting the NAWS survey for the Department of Labor. Use of the same survey organization is a deliberate effort to maintain consistency with the NAWS data. Analysis of the data will be conducted at the University of Florida under the direction of Robert Emerson, UF, (352) 392-1881 x300.

References

Agency for Workforce Innovation. Quarterly Census of Employment and Wages. Tallahassee, Florida. http://www.labormarketinfo.com/qcew/index.htm

Aguirre International. 2006. *National Agricultural Workers Study: Frequencies for Public Access Data, 1989-2002.* Burlingame, CA. June. Available at: http://aguirreinternational.com/naws/downloads/National_report_2002.pdf

Aigner, D. J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1:49-60.

Black, D. A., M. C. Berger, F. A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95:739-748.

Bound, J., C. Brown, and N. Mathiowetz. 2001. "Measurement Error in Survey Data." *Handbook of Econometrics, V. 5*. Ed. by J. J. Heckman and E. Leamer, pp. 3705-3843. New York: Elsevier Science.

Cameron, A. C. and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Cochran, W. G. 1968. "Errors of Measurement in Statistics." *Technometrics* 10:637-666.

Cochran, W. G. 1977. *Sampling Techniques. 3rd Ed.* New York: John Wiley and Sons.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: Johns Hopkins University Press.

Dillman, D. A., J. L. Eltinge, R. M. Groves, and R. J. A. Little. 2002. "Survey Response in Design, Data Collection, and Analysis," in *Survey Nonresponse*, ed. by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 3-26. New York: John Wiley & Sons.

Dye, R. F. and D. P. McMillen. 2007. "Teardowns and Land Values in the Chicago Metropolitan Area." *Journal of Urban Economics* 61:45-63.

Frazis, H. and M. A. Loewenstein. 2003. "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables." *Journal of Econometrics* 117:151-178.

Geweke, J. 1986. "Exact Inference in the Inequality Constrined Normal Linear Regression Model," *Journal of Applied Econometrics* 1:127-141.

Greene, W. H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Groves, R. M., and M. P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

2/2/2021

Hansen, L. P. 1982. "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica* 50:1029-1054.

Hansen, M. H., W. N. Hurwitz, and M. Bershad. 1961. "Measurement Errors in Censuses and Surveys." *Bulletin of the International Statistical Institute* 38:359-374.

Heckman, J. J. and S. Polachek. 1974. "Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship." *Journal of the American Statistical Association* 69:350-354.

Horvitz, D. G., and D. J. Thompson 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47:663-685.

Isé, S., and J. M. Perloff 1995. "Legal Status and Earnings of Agricultural Workers." *American Journal of Agricultural Economics* 77:375-386.

Iwai, N., R. D. Emerson, and L. M. Walters 2006. "Legal Status and U.S. Farm Wages." Selected Paper for presentation at the Southern Agricultural Economics Association Annual Meeting, Orlando, FL. http://agecon.lib.umn.edu/cgi-bin/pdf_view.pl?paperid=19740&ftype=.pdf

Iwai, N., R. D. Emerson, and L. M.Walters. 2008. "Labor Cost and Technology Adoption: Least Squares Monte Carlo Method for the Case of Sugarcane Mechanization in Florida." Selected Paper for presentation at the American Agricultural Economics Association Annual Meeting, Orlando, FL, July. Available at http://purl.umn.edu/6479

Kane, T. J., C. E. Rouse, and D. Staiger. 1999. Estimating Returns to Schooling when Schooling is Misreported. Natl. Bur. Econ. Res. Working Paper 7235. National Bureau of Economic Research, Cambridge, Mass. July. Available at: http://www.nber.org/papers/w7235

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Koch, G.G. 1973. "An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means." *Journal of the American Statistical Association* 68:906-913.

Lahiri, D. B. 1951. "A Method for Sample Selection Providing Unbiased Ratio Estimates." *Bulletin of the International Statistical Institute* 33:133-40.

Mincer, J. 1974. *Schooling, Experience and Earnings.* New York: Columbia University Press.

Särndal, C-E, and S. Lundström. 2005. *Estimation in Surveys with Nonresponse.* Chichester: John Wiley & Sons.

Tillé, Y. 2006. *Sampling Algorithms.* New York: Springer.

2/2/2021

Tillé, Y., and A. Matei. 2008. "The Sampling Package." Available at: http://cran.r-project.org/web/packages/sampling/sampling.pdf

U.S. Department of Labor. The National Agricultural Workers Survey. Washington, D.C. http://www.doleta.gov/agworker/naws.cfm

Walters, L. M., R. D. Emerson, and N. Iwai. 2008. "Proposed Immigration Reform and Farm Labor Market Outcomes." Selected Paper for presentation at the American Agricultural Economics Association Annual Meeting, Orlando, FL, July. Available at: http://purl.umn.edu/6285

2/2/2021

APPENDIX A

Within Employer Variance

A key piece of information required for both the strata allocations and the calculation of the variance is the parameter S_{2hi}^2, the within employer variance of earnings since there are no individual worker earnings data in the QCEW data. We follow Cochran's (1977) suggestion to use an estimate of the variance from available alternative data sources. Previous surveys of citrus and tomato employers and workers conducted at the University of Florida include earnings data for multiple workers per employer. These permit the direct estimation of a mean and variance for earnings for three of the five strata: citrus growers, citrus labor contractors, and tomato growers. The coefficient of variation of earnings (CV) is calculated for each of these groups as a more stable estimate to be transferred across surveys (Cochran 1977). The CV for the mean of earnings, $\bar{y}_h$, is defined as

$$CV_h = \frac{\sqrt{S_{2h}^2}}{\bar{y}_h}$$

where

$$S_{2h}^2 = \sum_{i=1}^{N_h} \frac{M_{hi}}{M_h} S_{2hi}^2$$

$$S_{2hi}^2 = \frac{1}{M_{hi} - 1} \sum_{j=1}^{M_{hi}} \left(y_{hij} - \bar{y}_{hi} \right)^2$$

There are no available external data for the citrus grove care stratum to calculate a *CV*. Since employment in grove care work tends to be less seasonal than for much agricultural employment, the lowest of the three available *CV*'s was used for that stratum. Similarly, there are no available external data for strawberry growers. The *CV* was taken to be the average of the *CV* for citrus and tomato growers. Although the *CV*'s are clearly constant across employers when transferred to the QCEW data set, the variances are not. The within employer variances in the QCEW data set are then calculated as

$$S_{2hi}^2 = CV_h^2 \times \bar{y}_{hi}^2$$