Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitudinal Employer-Employee Data for the State of Washington

John M. Abowd, Hampton Finer and Francis Kramarz
Cornell University, CREST and NBER; Cornell University; and INSEE-CREST

Revised: January 1999

Prepared for the May 21-22, 1998 International Symposium on Linked Employer-Employee Data sponsored by the United States Bureau of the Census. We thank David Margolis for comments and suggestions. We are grateful to the Washington State Division of Unemployment Insurance, and Wayne McMahon in particular, for providing access to the State's UI data files and to INSEE-CREST for providing access to the French data we have previously analyzed and which we use for comparative purposes in this paper. Abowd and Finer acknowledge support from the National Science Foundation (SBER 96-18111 and SBR 93-21053). Laurence Allain labored long and hard to create the research version of the Washington State UI data and we are grateful for her efforts. The data used in this paper are confidential but the authors' access is not exclusive. For further information regarding the Washington State UI Data, contact Wayne McMahon, State of Washington, Employment Security Department, Unemployment Insurance Division, Box 9046, Olympia, WA 98507-9839.

Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitudinal Employer-Employee Data for the State of Washington

Abstract

We study a statistical decomposition of full time earnings that includes correlated components related to observable individual characteristics, unobservable individual characteristics and unobservable employer characteristics. Using data from the State of Washington Unemployment Insurance System, we estimate the components of this model and the associated correlations among those components. Unobservable individual and firm heterogeneity are the most important of the components of hourly wages earnings, each accounting for 24% of the variance. The two components are not correlated ($\rho = -0.005$). We decompose the industry effects in hourly wage rates into the components due to the average unobservable individual heterogeneity and the average unobservable employer heterogeneity. Unobservable individual heterogeneity and unobservable employer heterogeneity are both important. The results for the State of Washington are compared with existing results for France where similar data have been analyzed.

John M. Abowd
Dept. of Labor Economics
Cornell University
Ithaca, NY 14583-3901

john.abowd@cornell.edu

Hampton Finer
Federal Trade Commission
6th and Pennsylvania, NW
Washington, DC 20580

hfiner@ftc.gov

Francis Kramarz
INSEE-CREST
15, bd Gabriel Péri
92245 Malakoff Cedex
France

kramarz@ensae.fr

I. Introduction

Although labor economists have seen the benefits of microeconomic data relating characteristics of individuals to the characteristics of their employers for many years (see, for example, Rosen (1986) and Willis (1986)), recent progress using research data created from comprehensive administrative reports of earnings has made such analyses feasible. Most work has been done on European data, French in particular (see Abowd and Kramarz 1999 for a comprehensive review). Because such data permit researchers to begin to disentangle the effects of employer decisions from the effects of choices made by workers, interest has focused on models in which individual and employer effects are separately identifiable. This means that the data must have a longitudinal dimension for both the individuals and employers. In one of the first analyses of earnings based upon such data, Abowd, Kramarz and Margolis (1999) presented an extensive statistical study of simultaneous individual- and employer-level heterogeneity in the determination of compensation for French private-sector employees.[1] Until recently, comparable American data were not available. In the present paper we continue that line of research by

[1] Other data with which one can identify correlated person and individual effects for either wages or employment spells now exist for Belgium (Leonard and Van Audenrode, 1995, 1996), Denmark (Belzil, 1997, and Bingley and Westergård-Nielsen, 1996), France (Entorf and Kramarz, 1997 and Entorf, Gollac and Kramarz, 1999, and Margolis, 1996, Goux and Maurin forthcoming, in addition to Abowd, Kramarz and Margolis, 1999). See the review in Abowd and Kramarz (1999).

applying some of their methods to the study of American data taken from the Washington State Unemployment Insurance system.[2]

We are concerned with two main issues in this paper. First, to what extent does individual and employer unobservable heterogeneity relate to earnings outcomes? In addressing this question we use statistical techniques that permit a decomposition of full time earnings into correlated components related to observable individual characteristics, unobservable individual characteristics and unobservable employer characteristics. In the Washington State data, observable personal characteristics, which are limited in number but which represent all the usual human capital and demographic variables, play a very limited role in explaining full-time earnings heterogeneity. Unobserved personal and employer heterogeneity are equally important factors—each explaining about 24% of the variation in log real full time hourly wage rates. The two unobserved heterogeneity factors are not correlated (although the estimation procedure permitted such correlation) and, together with observed personal heterogeneity, explain 89% of the variation in full-time hourly wages. An important feature of our model, which is different from other component of variance approaches, is that we allow the unobserved heterogeneity (individual and employer) to be correlated with the observed personal heterogeneity. These results are similar to French analyses that have been performed on models with comparable statistical structure except that the employer heterogeneity is much more important in the Washington State data.

[2] Other research teams have used state unemployment insurance data for similar purposes. See Lane, Burgess and Theeuwes (1997), and Lane et al. (1998), for examples.

The second question we consider is the extent to which inter-industry wage differentials can be explained by personal or employer unobserved heterogeneity. Our analysis is in the spirit of Krueger and Summers (1988) but we use the formal relations among the various statistical effects derived in Abowd, Kramarz and Margolis (1999). Our statistical model is the formal decomposition of the industry (firm-size) effect, conditional on observable personal heterogeneity, into the weighted average unobserved personal heterogeneity plus the weighted average unobserved employer heterogeneity (conditional on the same observables). We estimate these decompositions using the effects generated by our correlated unobserved heterogeneity model. Unobserved personal heterogeneity is slightly more important than unobserved employer heterogeneity in explaining the Washington State industry effects, explaining as much of the inter-industry wage differences as was found for France. In the Washington State data, however, unobserved employer heterogeneity plays a significant role in explaining the industry effects, in marked contrast to the French case where only unobserved personal heterogeneity was important.

The paper is organized as follows. Section II reviews our statistical techniques, relying heavily on Abowd, Kramarz and Margolis (1999). Section III discusses the Washington State UI data. Section IV discusses our results and Section V concludes.

II. Statistical Theory for Matched Employer-Employee Data

The underlying statistical model is:

$$y_{it} = x_{it}\beta + \alpha_i + \psi_{J(i,t)} + \varepsilon_{it} \qquad \text{(SM1)}$$

where y_{it} is the dependent outcome for individual *i* in period *t*, stated as a deviation from its grand mean μ_y, α_i is the person effect—the effect of unmeasured personal characteristics, $\psi_{J(i,t)}$

is the firm effect—the effect of unmeasured employer characteristics,[3] x_{it} is a $(1 \times P)$ vector of time-varying personal characteristics stated in deviations from their grand means m_x, b the effect of measured time-varying characteristics of the individual, e_{it} is a statistical error with properties given below, and the function $J(i,t)$ associates an employer, indexed by *j*, with individual *i* at time *t*. The subscripts run $i = 1,\ldots,N$, $t = B_i,\ldots,L_i$, $j = 1,\ldots,J$, where B_i is the first period an individual appears in the sample, L_i is the last period the individual appears in the sample and, hence, $T_i \equiv L_i - B_i + 1$ is the number of periods available for the individual.[4] For convenience, we restate equation (SM1) in matrix notation as:

$$y = Xb + Dq + Fy + e \qquad \text{(SM2)}$$

y is the $N^* \times 1$ vector of dependent outcomes, *D* is the individual effect design matrix, $N^* \times N$, for the person effects, *F* is the firm effect design matrix, $N^* \times J$, for the firm effects, *X* is the $N^* \times P$ design matrix for the time-varying personal characteristics, e is the $N^* \times 1$ vector of statistical errors further specified below, q is the $N \times 1$ vector of person effects, y is the $J \times 1$ vector of firm effects, b is the $P \times 1$ vector of time-varying personal characteristic effects, and

$$N^* \equiv \sum_{i=1}^{N} T_i \ .$$

[3] We deliberately abstract from measured, non time-varying personal and firm characteristics because they are easy to incorporate in the statistical analysis. A time-varying firm effect is also easy to incorporate, see Abowd, Kramarz and Margolis (1999).

[4] The methods discussed here can easily accommodate unbalanced data with holes but the notation becomes more cumbersome, so we have omitted that possibility from the theoretical discussion but not from the actual analysis formulas.

When we require only the moments of e they are specified as

$$\mathrm{E}\left[e \mid X, D, F\right] = 0 \text{ and } \mathrm{V}\left[e \mid X, D, F\right] = s_e^2 I_{N^*}.^5 \qquad \text{(SM3)}$$

When a distribution is required, we assume *i.i.d.* normal errors, but for most methods this is not essential.

II.a. Estimation Biases in Models that Ignore Correlated Person and Firm Heterogeneity

The most common forms of statistical analyses of models like (SM2) involve aggregation of the person effects into occupational indicators, the firm effects into industry indicators, or both.[6] Since the statistical properties of these aggregations are all identical, we illustrate the problems using an aggregation of the firm effects into industry effects. Let the matrix A, $J \times K$, define an aggregation of J firms into K industries so that the element a_{jk} is equal to 1 if firm j is

[5] The homoscedastic and serially uncorrelated error assumptions are made for convenience only. The consistent method in Abowd, Kramarz and Margolis (1999), and discussed below, requires neither. Random effects and semi-parametric methods are somewhat complicated by the relaxation of this assumption.

[6] For analyses of wage rates using models like (SM2) with aggregation of firm effects into industry effects see Dickens and Katz (1987), Krueger and Summers (1987, 1988), Murphy and Topel (1987), and Gibbons and Katz (1992). For analyses of wage rates using models like (SM2) with aggregation of firm effects into firm-size effects, see Brown and Medoff (1989) and the references therein. For analyses of wage rates using these models with aggregation of person effects into occupational indicators see Groshen (1991, 1996) and especially the references in the latter.

in industry *k*. Define a pure industry effect, k_k as the employment-duration weighted average of the firm effects within industry *k*:

$$k_k \equiv \sum_{i=1}^{N} \sum_{t=B_i}^{L_i} \left[\frac{1\left[\mathrm{K}(\mathrm{J}(i,t)) = k\right] y_{\mathrm{J}(i,t)}}{N_k} \right] \qquad \text{(EB1)}$$

where

$$N_k \equiv \sum_{j=1}^{J} 1\left[\mathrm{K}(j) = k\right] N_j \ , \ N_j \equiv \sum_{i=1}^{N} \sum_{t=B_i}^{L_i} 1\left[\mathrm{J}(i,t) = j\right],$$

the function 1[*C*] is the indicator function for the condition *C*, and the function K(*j*) denotes the industry classification of firm *j*.[7] Equation (SM2) can now be restated as

$$y = Xb + Dq + FAk + M_{FA} F y + e \qquad \text{(EB2)}$$

where k is the $K \times 1$ vector of pure industry effects and the matrix $M_Z \equiv I - Z(Z'Z)^{-1} Z'$ for any matrix *Z*. Direct manipulation of the definition of a pure industry effect yields the identity $k \equiv (A'F'FA)^{-1} A'F'Fy$. If the equation (EB2) is estimated omitting the firm effects $M_{FA} F y$, then the resulting estimator for the industry effects, say k^*, is biased as shown in the expression:

$$k^* = k + \left(A'F'M_{[D \ \ X]} FA \right)^{-1} A'F'M_{[D \ \ X]} M_{FA} F y \ , \qquad \text{(EB3)}$$

---

[7] Although it appears awkward, the definition of the pure industry effect in equation (EB1) is exact for a simple random sample of *N* individuals and requires only sample weights for other sampling schemes. All the authors cited in the note above, except for Groshen, use this definition of a pure industry effect when they estimate models using representative samples of individuals. Groshen's models are normally fit on samples that are firm-based and, hence, not representative of individuals. Assuming that one wants an estimate that is representative of individuals, all other definitions of the pure industry effect have an aggregation bias.

which simplifies to $k^{*} = k$ if, and only if, the industry effect design, *FA*, is orthogonal to $M_{FA}F$, given *D* and *X*. This condition is generally not true even though *FA* and $M_{FA}F$ are orthogonal by construction.

Using French data, Abowd, Kramarz and Margolis (1999) found that the bias shown in equation (EB3) is serious and that the French data support the interpretation that the pure industry effects are much less important than aggregated person effects in explaining inter-industry wage differentials in France. The simplest way to understand this result is to consider the estimation of equation (EB2) omitting both Dq and $M_{FA}Fy$. The resulting parameter for the raw industry effects conditional on *X*, say k^{**}, can be expressed as

$$k^{**} = k + (A'F'M_X FA)^{-1} A'F'M_X (M_{FA}Fy + Dq), \qquad \text{(EB4)}$$

which simplifies to

$$k^{**} = (A'F'M_X FA)^{-1} A'F'M_X Fy + (A'F'M_X FA)^{-1} A'F'M_X Dq. \qquad \text{(EB5)}$$

Equation (EB5) says that raw industry effects k^{**}, estimated conditional on observable personal characteristics *X*, can be decomposed into the sum of the (properly-weighted, conditional) average firm effect and average person effect within the industry. The French result means that the component related to the average person effect explains 92% of the variability in raw industry effects by itself, the component related to the average firm effect explains 24% of the variability in raw industry effects by itself, and together they explain 95% of the variability in raw industry effects.[8] The data analyses we perform in this paper will allow examination of these issues using American data from the State of Washington, which has not previously been possible.

[8] If the properly-weighted, conditional average person and firm effects within the industry had been estimated without error, the two effects would have explained 100% of the variability in

II.b. Identification and Estimation by Consistent, Fixed-Effect Methods

The most familiar statistical method that may be used to estimate the coefficients of all time-varying effects (including time-varying firm effects when they are part of the model) is a particular form of fixed-effect estimation.[9] After a redefinition of the non-time-varying firm effect, this method can also be used to recover estimates of the fixed firm and person effects. Assumption (SM3) is not required for this method and is replaced by

$$\mathrm{E}\left[e \mid X, D, F\right] = 0, \quad \mathrm{E}\left[e_{it} e_{ns} \mid X, D, F\right] = \begin{cases} \text{bounded, if } i = n \\ 0, \text{ if } i \neq n \end{cases}$$

$$\text{and } \mathrm{E}\left[e_{it} e_{ns} e_{mu} e_{pv} \mid X, D, F\right] = \begin{cases} \text{bounded, if } i = n = m = p \\ 0, \text{ otherwise.} \end{cases}.$$

The estimating formulas for this method can be found in Abowd, Kramarz and Margolis (1999).

II.c. Identification and Estimation by Conditional Methods

The technique we have used in other work on French data produces consistent estimates of the time-varying personal characteristics, firm effects and functions of person effects when a set of Q variables, called Z, can be found such that the conditional covariance between X, D, and F is zero, given Z. The complete results can found in Abowd, Kramarz and Margolis (1999) but we summarize them here. The model is restated as

$$y = Xb + Dq + Z l + M_Z F y + e \qquad \text{(ZM1)}$$

---

the raw industry effect, as shown in equation (EB5). The correlation between the two industry averages explains why the sum of the two percentages explained individually is not 95%.

[9] The technique described in the section has been used by Abowd and Kramarz (1996) and by Allain (1996) on the French and Washington State UI data. Allain studied layoff rates and not compensation.

with the auxiliary $Q \times 1$ parameter vector $l \equiv (Z'Z)^{-1} Z'Fy$. The statistical properties of e are unchanged. Estimation is based on the maintained hypothesis

$$X'M_Z F = 0 \text{ and } D'M_Z F = 0 \qquad \text{(ZM2)}$$

and the least squares estimators are

$$\begin{bmatrix} \hat{b} \\ \hat{q} \\ \hat{l} \end{bmatrix} = \begin{bmatrix} X'X & X'D & X'Z \\ D'X & D'D & D'Z \\ Z'X & Z'D & Z'Z \end{bmatrix}^{-} \begin{bmatrix} X'y \\ D'y \\ Z'y \end{bmatrix} \qquad \text{(ZM3)}$$
$$\hat{y} = \left(F'M_Z F\right)^{-} F'M_Z y$$

where $[.]^{-}$ is any generalized inverse. This method is called conditional: persons first in Abowd, Kramarz and Margolis (1999)

II.d. Identification and Estimation by other Fixed-Effect Least Squares Methods

We expand on the Abowd, Kramarz, Margolis (1999) conditional methods by describing a technique, which we apply in this paper, that allows for the estimation of a large number of firm effects along with all of the person effects. The conditioning effects, *Z*, apply only to the remaining, firm effects. Those remaining firm effects are estimated using the conditional method described in section II.c. The model is restated as

$$y = Dq + Xb + F_1 y_1 + Zl + (F_2 y_2 - Zl) + e \qquad \text{(ZM6)}$$

where the partitioning of F and y is based on keeping the firm effects associated with the firms in F_1 in the model along with D and relegating only the effects in F_2 to the conditional estimation step with the auxiliary $Q \times 1$ parameter vector $l \equiv (Z'Z)^{-1} Z'F_2 y_2$. The statistical properties of e are unchanged. Estimation is based on the maintained hypothesis

$$X'M_Z F_2 = 0 \text{ and } D'M_Z F_2 = 0 \qquad \text{(ZM7)}$$

and the least squares estimators are

$$\begin{bmatrix} \hat{b} \\ y_1 \\ \hat{l} \end{bmatrix} = \begin{bmatrix} X'M_D X & X'M_D F_1 & X'M_D Z \\ F_1'M_D X & F_1'M_D F_1 & F_1'M_D Z \\ Z'M_D X & Z'M_D F_1 & Z'M_D Z \end{bmatrix}^{-1} \begin{bmatrix} X'M_D y \\ F_1'M_D y \\ Z'M_D y \end{bmatrix} \qquad \text{(ZM8)}$$

$$\hat{q} = (D'D)^{-} D'\left(y - X\hat{b} - F_1\hat{y}_1 - Z\hat{l}\right) \qquad \text{(ZM9)}$$

and

$$\hat{y}_{2\{j\}} = (F_{2\{j\}}'F_{2\{j\}})^{-} F_{2\{j\}}'\left(y_{\{j\}} - X_{\{j\}}\hat{b} - D_{\{j\}}\hat{q}_{\{j\}}\right) \qquad \text{(ZM10)}$$

where the notation $\{j\}$ means all the observations that occur in firm j for all i and t. We call this method "least squares persons and large firms," below.

III. Washington State Unemployment Insurance System Data

The State of Washington maintains a very complete data for all unemployment insurance recipients and a random sample of 10% of the unemployment insurance eligible work force. These data have been used by Anderson and Meyer (1994) and by Allain (1996) to study characteristics of the unemployment insurance system. The data used in this paper come from two types of administrative records, collected in the context of the Continuous Wage and Benefit History (CWBH) project over the 1984-1993 period. The first type are quarterly wages records for a 10% sample of Washington State's UI-eligible workers. Since coverage of workers is nearly universal except for the self-employed, our sample is close to a representative random sample of all employees in Washington State. Our wage file covers the years 1984 through 1993. In addition to the quarterly wage data, the data also include a firm identifier, as well as the firm's 4-digit Standard Industrial Classification code, the firm's average monthly employment, total wages, taxable wages and tax rate. The second type of data are UI claims records for any worker who filed for UI over the 1984-1993 period. This data set contains, for each claim filed, the worker's

identifier, the date the claim was filed, the first pay date and the exhaustion date, the total amount of benefits paid, the reason for work separation, as well as the usual personal characteristics (age, sex, race, schooling). Our analysis sample is restricted to quarters in which respondents worked more than 400 hours. There are roughly 400,000 quarterly observations for each data year in the wage files, corresponding to 296,801 unique individuals and 89,397 unique firms in our analysis sample over the 10 year period. A little over 300,000 valid UI claims were filed in Washington State over our sampling period, originating from approximately 150,000 individuals. We are able to match these two types of record, in order to form quarterly job-match histories. Table 1 shows summary statistics from these data.

The Washington State data contain both employer and employee identifiers (the latter in scrambled form), thus permitting direct estimation of models with correlated person and firm heterogeneity. Allain (1996) has used these data to decompose UI-eligible layoffs into components related to person effects, firm effects and time-varying personal characteristics as shown in equation (SM1) with y_{it} defined as the quarterly individual layoff rate.

One important limitation of the Washington UI data is that, while the employer-reported earnings, the employer ID, and the employer's industry are observed for all persons in the 10% sample, other personal characteristics, such as sex, race, age, schooling and initial seniority, are observed only when a person has filed for unemployment insurance benefits. Allain (1996) imputed the missing personal characteristics using a two-sample multiple imputation algorithm (Rubin 1976, 1987, 1996). We followed a similar procedure in this paper.

The Expectation Step

Our ancillary data consisted of a 1.8 million observation sample of employed persons from the outgoing rotation group files of the Current Population Survey for the months January 1984 to December 1993. For the "expectation" step of our imputation strategy, we computed, for each individual, the expected value of our four missing variables (sex, race, schooling and potential labor market experience), conditioning on the set of variables which we completely observe in both samples, namely industry and wage. In order to compute the conditional expectation of our two categorical variables (sex and race), we estimated a logistic regression model that predicts four sex-race categories (white male, non-white male, white female, non-white female) on the basis of an indicator for Washington State, indicators for the five wage quintiles (based on the distribution of the wage data in the CPS sample), indicators for 10 1-digit industry classifications, and the interaction of the wages and industry effects. We chose to estimate the sex and race variables simultaneously in order to preserve the possible correlation between these two variables in samples of employed persons. Our potential labor force experience and schooling variables were predicted in a multivariate least-square regression framework, using the same independent variables.

Let x_{it} represent the vector of individual characteristics, including time-invariant observable characteristics in this section—namely, sex, race, schooling and potential labor market experience), ℓ_{it} the unemployment insurance status of individual *i* at period *t*. In the Washington State data, we observe:

$$x_{it} = \begin{cases} x_{it} \text{ if } \ell_{is} = 1 \text{ for some } s = B_i, \ldots, L_i \\ \text{missing, otherwise} \end{cases} \qquad \text{(IM1)}$$

Let z_{it} be the vector of variables that is always observed (namely wage and industry). Using our ancillary sample, we estimated the following regression by ordinary least squares (for the continuous *x* variables schooling and experience) or by logistic regression (for the sex/race pairs):

$$x_{it}^{*} = x^{*} z_{it}^{*} + n_{it}^{*} \qquad \text{(IM2)}$$

where variables indexed with a star denote that only the individuals in our ancillary sample who were never laid off are used in the regression (i.e. $\ell_{is} = 0, \forall s$) and n is the error from the appropriate statistical model. Equation (IM2) provides us with a consistent estimator of $x^{*}, \hat{x^{*}}$, which in turn enables us to compute the conditional expectation of x_{it}. We assume that sex, race and schooling do not vary for a given individual within our sample period, and that the value of the potential experience variable in the first period that an individual is observed in the sample is sufficient to provide values for every additional quarter an individual is observed in the sample. Therefore, we only need to compute the conditional expectation of our missing variables for the first period an individual is observed in the sample. A consistent estimator of this conditional expectation is:

$$E[x_{iF_i} | z_{iF_i}, \ell_{iB_i} = \ldots = \ell_{iL_i} = 0] = x^{*} z_{iB_i} \qquad \text{(IM3)}$$

The Imputation Step

The second step of the imputation procedure is add to our predicted values random noise drawn from the appropriate distribution (i.e. the empirical distribution of the CPS data), in order to reflect sampling variability, and to repeat this procedure *m* times (five, in our case) in order to further account for the uncertainty inherent to process generating the missing values. To impute missing sex and race information, we computed the predicted probabilities from the logistic

regression model given an individual's information in the Washington UI data for the first year in which that individual appears. We then drew a random number from the uniform distribution over the [0,1] interval, and imputed each individual's sex/race category based on a comparison between this random number and the four predicted conditional probabilities. The procedure was performed five times (with independent draws of the random component in each case) for each individual.

In a given imputation file, a person's imputed sex/race cell is a constant over all observations; however, there is, generally, variability in this imputed value across imputation files. We adopted a slightly different procedure to impute an individual's schooling/potential experience pair. For the first quarter observed for an individual who never filed for unemployment, we imputed the predicted mean of the schooling and potential labor market experience variables from the multivariate regression (given the individual's wage/industry information) plus a randomly chosen residual pair drawn from the same wage decile/2-digit industry/half cell as the individual. The imputed schooling is not changed for subsequent quarterly observations on the individual. The imputed potential experience is updated for each additional quarter that the individual appears in the sample. This imputation procedure is performed, with independent imputation of the residual component, across all five imputation files. Thus,

$$x_{iF_i}{}^{m} | z_{iF_i}, \ell_{iF_i} = \ldots = \ell_{iL_i} = 0 = \hat{x}^{*} z_{iF_i} + \hat{F}^{-1}(t^{(m)} | z_{iF_i}, \ell_{iF_i} = \ldots = \ell_{iL_i} = 0))$$

where $\hat{F}$ (.) is the empirical cumulative distribution function of the residuals in the CPS data, and t is a uniform random variable defined on [0,1]. The superscript *m*, *m*=1,...,5, indexes the multiple imputation, since there is independent draw of the residual component for every imputation.

Given these imputations, we imputed initial seniority for the employment spell in progress when an individual entered the sample using six CPS samples from 1983-1994 that included information on seniority in the current job as the ancillary estimation sample. The result is an analysis-ready data set consisting of five samples with independently imputed missing values. The standard multiple-imputation formulas can be applied to all statistical analyses of these samples.

After repeating the imputation procedure five times, we have created five distinct imputed data sets. Rubin has shown that valid statistical inferences can now be made, using all the information available to us. Consistent measures of central tendency (moments, conditional moments) are obtained by averaging these statistics across the five imputation samples. Measures of variability (standard deviations, standard errors) are computed using the classic decomposition of the unconditional variance into the sum of the expected conditional variance and the variance of the conditional expectation. In our case, the statistics of interest are mostly first and second moments, explicitly computed using information from all five imputed data sets based on the following formulas:

$$\hat{b} = \frac{1}{5}\sum_{m=1}^{5} \hat{b}^{(m)}$$

and

$$\hat{\mathrm{V}}[\hat{b}] = \frac{1}{5}\sum_{m=1}^{5} \hat{\mathrm{V}}[\hat{b}^{(m)}] + \frac{1}{5}\sum_{m=1}^{5} (\hat{b}^{(m)} - \hat{b})(\hat{b}^{(m)} - \hat{b})'$$

for a regression coefficient and similar formulas for the other fixed parameters in our analysis. We use only the first imputed sample in this paper because experimentation with the multiple imputation formulas revealed that neither the means nor the standard errors were affected by the multiple imputations. Table 1 shows summary data for men and women, separately.

The imputation procedure that we used has been widely studied in statistics (see Rubin 1976, 1987, 1996 and the references therein). Since the missing data occur for individuals who never experienced a spell of insured unemployment over the sample period, it is clear that any procedure for imputing the missing data must use an ancillary sample—one that contains information for individuals with infrequent spells of unemployment. Nevertheless, there is some concern that such extensive imputation of missing data might lead to econometrically invalid results. While there is no close substitute for having sample information on sex, race, schooling and age for most of the individuals, we must stress that the critical data items—wage rates, person IDs and employer IDs—are never missing in our data. The use of the imputed data is limited to the decomposition of the personal heterogeneity into a part explained by sex, schooling and race and an unexplained part. All of the results that we present are substantively unaffected by whether we consider the person effect inclusive of measured non-time-varying characteristics (q), or the person effect excluding the part due to measured non-time-varying characteristics (a). Results based on q do not use any imputed data, except for the role of schooling in computing the initial value of labor force experience. Subsequent values of labor force experience are accumulated from the observed employment spells. In specifications that include person effects, an error in the imputation of the initial period value of labor force experience could affect the decomposition of log wage rates into parts due to personal heterogeneity and observed time-varying personal characteristics.

IV. Results

Table 2 shows the results of estimating the coefficients of the time-varying variables by he consistent method, the conditional method with persons first and the least squares method with

persons and large firms.[10] The estimates are quite similar across techniques except for the time trend implied by the consistent method as compared to the other two methods. This result is probably due to the heavy reliance of the consistent method on the designation of experience and seniority as time varying effects.

Table 3 gives the standard deviations and correlations of a decomposition of the dependent variable (log hourly wage rates, y) into the components related to observable time-varying personal characteristics (xb), personal heterogeneity that is not time-varying (q), observed non time-varying heterogeneity (h), unobserved non time-varying heterogeneity (a), unobserved employer heterogeneity (f), unobserved employer seniority effects (g). The results are based on the column labeled "Least Squares Persons and Large Firms" in Table 2 using the formulas in section II.d, above. The unobserved firm heterogeneity was estimated using the firm effects based upon the methods discussed above and equations (ZM8) and (ZM10). The decomposition into the unobserved component f and the firm-specific seniority component, gs follows Abowd, Kramarz and Margolis (1999). The non time-varying personal heterogeneity (q) was was decomposed into an observable part (based on sex, race and schooling) using least squares applied to the equation

[10] Standard errors presented in Table 2 have not been corrected for the multiple imputation of missing personal characteristics. In our experience this correction rarely affects any of the significant digits in the standard errors because none of the critical data, log wage rates, person identifiers and firm identifiers, is missing.

(ZM9) and the non time-varying unobserved heterogeneity (a) is the residual from that equation.[11]

Table 3 demonstrates that unobservable individual and firm heterogeneity are the most important components of log real hourly wage rates, correlation with y = 0.475 for the person effect (a) and 0.494 for the firm effect (f). The two components are not highly correlated (correlation of a and f = -0.005). Observed individual heterogeneity is the least important component (correlation with y=0.323).

These results are similar to French results in Abowd, Kramarz and Margolis (1999) except that the employer heterogeneity is much more important in the Washington State than in France. This is a surprising result given the differences in the labor markets between the two labor markets, particularly the French use of industry level bargaining and nearly universal coverage by collective bargaining agreements. These results are also similar to Danish results in Bingley and Westergård-Nielsen (1996); however, those authors did not allow correlation between the unobserved personal heterogeneity and the observable personal characteristics. The small amount of correlation between individual and firm effects is a surprising, but consistent, finding in these models as well.

We next consider the ability of our individual and employer effects to explain inter-industry wage differentials. We computed a raw industry effect for 2-digit industries, conditional on the variables shown in Table 2 and the individual characteristics shown in Table 1, which corresponds to the effect k^{**} in equation (EB4). We next computed the two components of the

[11] See Abowd, Kramarz and Margolis (1999) for details of the decomposition of the person and firm effects.

exact decomposition shown in equation (EB5) by substituting our estimated firm effects for y and our estimated individual effects (a) for q.[12] The R^2 of the decomposition is 0.99, indicating that the formula, which is exact when the parameters are not estimated, is also quite good given our parameter estimates. Figure 1 shows that the industry effect in the State of Washington is very closely related to the industry average person effects. For comparison Figure 2 shows the same relation in France. In both figures, we have inserted the predicted industry effect using only the industry average person effect as an explanatory variable. This line provides a reference against which to judge the degree of explanatory power of the variable that is equivalent to using the R^2 from the single variable regression.

Figure 3 shows that there is a less tight, but still quite important, relation between the industry average firm effect and the raw industry effect in the State of Washington. This stands in marked contrast to the results one gets for France, shown in Figure 4, where there is almost no relation between the industry average firm effect and the raw industry effect.

V. Conclusion

We study a statistical decomposition of full time earnings that includes correlated components related to observable individual characteristics, unobservable individual characteristics and unobservable employer characteristics. Using data from the State of Washington Unemployment Insurance System, we estimate the components of this model and the associated correlations among those components. Unobservable individual and firm heterogeneity

[12] Because the difference between a and q is just the part predicted by the non time-varying effects shown in Table 2, which are also included in the regressors used to compute the industry-average person effect, given X, we get exactly the same results whether we use a or q.

are the most important of the components of wages, each accounting for about 24% of the variance in log real hourly wage rates. The two components are not correlated. We also decompose the industry effects in log hourly wage rates into the components due to the average unobservable individual heterogeneity and the average unobservable employer heterogeneity. Both the unobservable individual heterogeneity and the unobservable employer heterogeneity are important in explaining inter-industry wage differentials, with individual heterogeneity being somewhat more important. The Washington State results for unobserved personal heterogeneity are similar to results we have found for France; however, the results on unobservable firm heterogeneity imply that this factor is much more important in the State of Washington than in France.

References

Abowd, John M. and Francis Kramarz, "Les Politiques Salariales: Individus et Entreprises," *Revue Economique*, 47, (1996): 611-622.

Abowd, John M. and Francis Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card Volume 3, Chapter 26 (Amsterdam: North Holland, 1999), forthcoming.

Abowd, John M., Francis Kramarz and David Margolis, "High Wage Workers and High Wage Firms," (1999), *Econometrica* available at http://old-instruct1.cit.cornell.edu:8000/abowd-john/high-wage-workers.pdf.

Allain, Laurence, *Essays in Compensation and Unemployment Insurance*, Cornell University Ph.D. Thesis, August 1996.

Anderson, Patricia and Bruce D. Meyer "The Extent and Consequences of Job Turnover," *Brookings Papers: Microeconomics*, (1994): 177-248.

Belzil, Christian, "Job Creation and Destruction, Worker Reallocation and Wages," Concordia University working paper, 1997.

Bingley, Paul and Niels Westergård-Nielsen, "Individual Wages within and between Establishments," University of Århus working paper, 1996.

Brown, Charles and James L. Medoff, "The Employer Size-Wage Effect," *Journal of Political Economy*, 97 (1989): 1027-1059.

Dickens, William T. and Lawrence Katz, "Inter-Industry Wage Differences and Industry Characteristics," in *Unemployment and the Structure of Labor Markets*, Kevin Lang and Jonathan S. Leonard, eds. (Oxford: Basil Blackwell, 1987).

Entorf, Horst and Francis Kramarz, "Does Unmeasured Ability Explain the Higher Wages of New Technology Workers?" *European Economic Review* 41 (1997): 1489-1510.

Entorf, Horst, Michel Gollac, and Francis Kramarz, "New Technologies, Wages, and Worker Selection," *Journal of Labor Economics*, 1999.

Gibbons, Robert and Lawrence Katz, "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies*, 59, (1992): 515-535.

Goux, Dominique and Eric Maurin, "Persistence of Inter-Industry Wage Differentials: A Re-examination on Matched Worker-Firm Panel Data," *Journal of Labor Economics*, forthcoming.

Groshen, Erica, "Sources of Intra-Industry Wage Dispersion: How Much Do Employers Matter?" *Quarterly Journal of Economics*, 106, (1991): 869-884.

Groshen, Erica, "American Employer Salary Surveys and Labor Economics Research: Issues and Contributions," *Annales d'économie et de statistique*, 41/42 (1996): 413-442.

Krueger, Alan and Lawrence H. Summers, "Reflections on the Inter-Industry Wage Structure," in *Unemployment and the Structure of Labor Markets*, ed. By Kevin Lang and Jonathan S. Leonard (Oxford: Basil Blackwell, 1987).

Krueger, Alan and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, 56, (1988): 259-293.

Lane, Julia, Simon Burgess and Jules Theeuwes, "The Uses of Longitudinal Matched Worker/Employer Data in Labor Market Analysis," *American Statitical Association Papers and Proceedings*, 1997.

Lane, Julia, Javier Miranda, James Spletzer and Simon Burgess, "The Effect of Firm Changes on Earnings Inequality: Longitudinal Evidence from Linked Data," American University Working Paper, 1998.

Leonard, Jonathan S. and Marc Van Audenrode, "Persistence of Firm and Individual Wage Components," Paper presented at the January 1996 AEA meetings (December, 1995).

Leonard, Jonathan S. and Marc Van Audenrode, "Worker's Limited Liability, Turnover and Employment Contracts," *Annales d'économie et de statistique*, 41/42, (January/June 1996): 41-78.

Margolis, David N., "Cohort Effects and Returns to Seniority in France," Annales d'économie et de statistique, 41/42, (January/June 1996): 443-464.

Murphy, Kevin M. and Robert H. Topel, "Unemployment, Risk and Earnings: Testing for Equalizing Wage Differences in the Labor Market," in *Unemployment and the Structure of Labor Markets*, ed. by Kevin Lang and Jonathan S. Leonard (Oxford: Basil Blackwell, 1987).

Rosen, Sherwin, "The Theory of Equalizing Differences" in *Handbook of Labor Economics*, ed. by Orley Ashenfelter and Richard Layard. (Amsterdam: North Holland, 1986), pp. 641-692.

Rubin, Donald, "Inference and Missing Data," *Biometrika* 63 (1976): 581-592.

Rubin, Donald, *Multiple Imputation for Nonresponse in Surveys*, (New York: Wiley, 1987).

Rubin, Donald B., "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association*, 91, No. 434 (1996): 473-489.

Willis, Robert (1986): "Wage Determinants: A Survey" in *Handbook of Labor Economics*, ed. by Orley Ashenfelter and Richard Layard. Amsterdam: North Holland, pp. 525-602.

Male										
Year	N	Wage/Hr.	Log(Wage/Hr.)	Hrs./Qtr.	Schooling	Seniority	Experience	White	Firm Size	Imputed
1984	196,898	12.62	2.41	513	13.13	6.40	18.93	0.89	9981	0.53
		(6.93)	(0.51)	(75)	(2.94)	(8.31)	(11.95)	(0.32)	(26444)	(0.50)
1985	201,739	12.59	2.40	515	13.14	6.15	19.32	0.88	11260	0.53
		(7.10)	(0.52)	(77)	(2.91)	(8.19)	(11.95)	(0.32)	(28094)	(0.50)
1986	206,573	12.84	2.42	517	13.11	6.09	19.72	0.88	11688	0.53
		(7.46)	(0.52)	(73)	(2.88)	(8.10)	(12.00)	(0.32)	(28743)	(0.50)
1987	224,184	12.54	2.39	522	13.14	5.50	19.94	0.88	11694	0.52
		(7.58)	(0.53)	(86)	(2.91)	(7.73)	(12.03)	(0.33)	(28308)	(0.50)
1988	228,447	12.38	2.37	523	13.03	5.68	20.28	0.87	12034	0.51
		(7.67)	(0.53)	(82)	(2.90)	(7.71)	(12.12)	(0.33)	(29244)	(0.50)
1989	252,131	12.31	2.37	523	13.04	5.30	20.48	0.87	11226	0.51
		(7.68)	(0.52)	(82)	(3.00)	(7.47)	(12.18)	(0.34)	(27907)	(0.50)
1990	268,374	12.20	2.36	520	12.97	5.31	20.75	0.86	11738	0.51
		(7.66)	(0.52)	(78)	(3.04)	(7.33)	(12.23)	(0.34)	(28441)	(0.50)
1991	271,807	12.24	2.36	519	12.96	5.42	21.19	0.86	11481	0.51
		(7.79)	(0.52)	(78)	(3.03)	(7.21)	(12.29)	(0.34)	(28136)	(0.50)
1992	276,685	12.31	2.36	523	12.95	5.61	21.48	0.86	11009	0.52
		(8.03)	(0.54)	(80)	(2.99)	(7.19)	(12.35)	(0.35)	(27496)	(0.50)
1993	274,712	12.14	2.35	520	12.93	5.81	21.84	0.86	10233	0.53
		(7.88)	(0.54)	(83)	(2.97)	(7.21)	(12.40)	(0.35)	(26708)	(0.50)
Female										
Year	N	Wage/Hr.	Log(Wage/Hr.)	Hrs./Qtr.	Schooling	Seniority	Experience	White	Firm Size	Imputed
1984	120,531	8.64	2.04	498	13.02	4.51	17.60	0.88	3408	0.54
		(4.91)	(0.46)	(60)	(2.79)	(6.14)	(12.00)	(0.33)	(13318)	(0.50)
1985	125,981	8.78	2.05	499	13.00	4.33	18.10	0.87	4262	0.54
		(5.09)	(0.47)	(63)	(2.74)	(5.95)	(12.06)	(0.33)	(15776)	(0.50)
1986	132,731	9.01	2.08	501	12.99	4.31	18.54	0.87	4816	0.54
		(5.23)	(0.48)	(62)	(2.69)	(5.81)	(12.04)	(0.33)	(17249)	(0.50)
1987	149,580	9.07	2.08	503	13.07	4.00	18.96	0.87	5306	0.55
		(5.27)	(0.48)	(67)	(2.67)	(5.58)	(12.07)	(0.33)	(17739)	(0.50)
1988	153,077	9.02	2.07	505	12.96	4.21	19.35	0.87	5600	0.55
		(5.44)	(0.49)	(71)	(2.64)	(5.59)	(12.15)	(0.34)	(19077)	(0.50)
1989	172,740	9.18	2.09	505	13.00	3.98	19.69	0.87	5797	0.57
		(5.48)	(0.48)	(68)	(2.72)	(5.40)	(12.18)	(0.34)	(18793)	(0.50)
1990	187,072	9.23	2.10	504	12.96	4.05	20.06	0.87	6129	0.57
		(5.56)	(0.47)	(66)	(2.73)	(5.29)	(12.22)	(0.34)	(19365)	(0.50)
1991	192,882	9.40	2.12	504	12.95	4.16	20.50	0.86	6071	0.58
		(5.58)	(0.48)	(65)	(2.72)	(5.17)	(12.25)	(0.34)	(19313)	(0.49)
1992	198,696	9.61	2.13	507	12.95	4.40	20.95	0.86	5930	0.59
		(5.93)	(0.49)	(66)	(2.65)	(5.19)	(12.31)	(0.35)	(18948)	(0.49)
1993	201,331	9.64	2.13	507	12.95	4.61	21.42	0.86	5295	0.60
		(6.02)	(0.50)	(71)	(2.59)	(5.23)	(12.40)	(0.35)	(17829)	(0.49)

Note: Wages are real hourly wages in 1982-84 dollars. There are 4,036,171 total observations.

Table 1
Summary Statistics for the Washington State Unemployment Insurance Data

Method:	Consistent		Conditional Persons First		Least Squares Persons and Large Firms	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
Male						
Experience	0.1076	0.0027	0.1126	0.0005	0.1043	0.0005
$(\text{Experience}/100)^2$	-0.3737	0.0174	-0.4286	0.0025	-0.3853	0.0024
$(\text{Experience}/1{,}000)^3$	0.0725	0.0050	0.0819	0.0007	0.0741	0.0007
$(\text{Experience}/10{,}000)^4$	-0.0050	0.0005	-0.0055	0.0001	-0.0050	0.0001
Year=1985	-0.0056	0.0008	-0.0280	0.0007	-0.0256	0.0007
Year=1986	-0.0079	0.0012	-0.0332	0.0008	-0.0307	0.0008
Year=1987	-0.0345	0.0014	-0.0646	0.0011	-0.0610	0.0010
Year=1988	-0.0482	0.0017	-0.0888	0.0013	-0.0859	0.0013
Year=1989	-0.0636	0.0018	-0.0966	0.0015	-0.0940	0.0015
Year=1990	-0.0901	0.0020	-0.1094	0.0018	-0.1076	0.0018
Year=1991	-0.1136	0.0022	-0.1232	0.0021	-0.1225	0.0020
Year=1992	-0.1353	0.0023	-0.1263	0.0024	-0.1264	0.0023
Year=1993	-0.1789	0.0025	-0.1429	0.0026	-0.1448	0.0026
Female						
Experience	0.0912	0.0029	0.0808	0.0005	0.0731	0.0005
$(\text{Experience}/100)^2$	-0.2753	0.0204	-0.2708	0.0030	-0.2354	0.0029
$(\text{Experience}/1{,}000)^3$	0.0591	0.0060	0.0534	0.0009	0.0465	0.0008
$(\text{Experience}/10{,}000)^4$	-0.0045	0.0006	-0.0037	0.0001	-0.0032	0.0001
Year=1985	-0.0028	0.0010	-0.0121	0.0009	-0.0111	0.0008
Year=1986	-0.0081	0.0015	-0.0091	0.0011	-0.0086	0.0010
Year=1987	-0.0262	0.0018	-0.0245	0.0013	-0.0233	0.0013
Year=1988	-0.0376	0.0021	-0.0400	0.0016	-0.0382	0.0016
Year=1989	-0.0492	0.0023	-0.0419	0.0019	-0.0405	0.0018
Year=1990	-0.0719	0.0025	-0.0474	0.0022	-0.0462	0.0022
Year=1991	-0.0870	0.0027	-0.0459	0.0025	-0.0459	0.0025
Year=1992	-0.1029	0.0029	-0.0404	0.0029	-0.0408	0.0028
Year=1993	-0.1366	0.0031	-0.0461	0.0032	-0.0471	0.0031
R^2	--		0.881		0.887	
N	4,036,171		4,036,171		4,036,171	
Individual Effects	296,801		296,801		296,801	
Firm Effects	36,914		36,914		501	

Table 2
Selected Coefficients and Standard Errors from Statistical Models of Individual and Firm Heterogeneity applied to the Washington State Unemployment Insurance Data

Least Squares with Persons and Large Firms									
	Log Real Wage	xβ	θ	α	uη	ψ	ϕ	γs	γ
Log Real Wage	1.000	0.323	0.583	0.475	0.361	0.513	0.494	0.089	-0.106
ind. char. xβ	0.323	1.000	-0.481	-0.489	-0.089	0.240	0.211	0.080	-0.107
person effect θ	0.583	-0.481	1.000	0.909	0.417	0.070	0.109	-0.070	-0.008
person α	0.475	-0.489	0.909	1.000	0.000	-0.005	0.046	-0.100	-0.009
person uη	0.361	-0.089	0.417	0.000	1.000	0.178	0.162	0.050	-0.001
firm effect ψ	0.513	0.240	0.070	-0.005	0.178	1.000	0.878	0.338	-0.087
firm ϕ	0.494	0.211	0.109	0.046	0.162	0.878	1.000	-0.154	-0.325
seniority eff. γs	0.089	0.080	-0.070	-0.100	0.050	0.338	-0.154	1.000	0.459
seniority slope γ	-0.106	-0.107	-0.008	-0.009	-0.001	-0.087	-0.325	0.459	1.000

Table 3
Correlation Coefficients for Components of Log Real Wage in the Washington State Unemployment Insurance Data

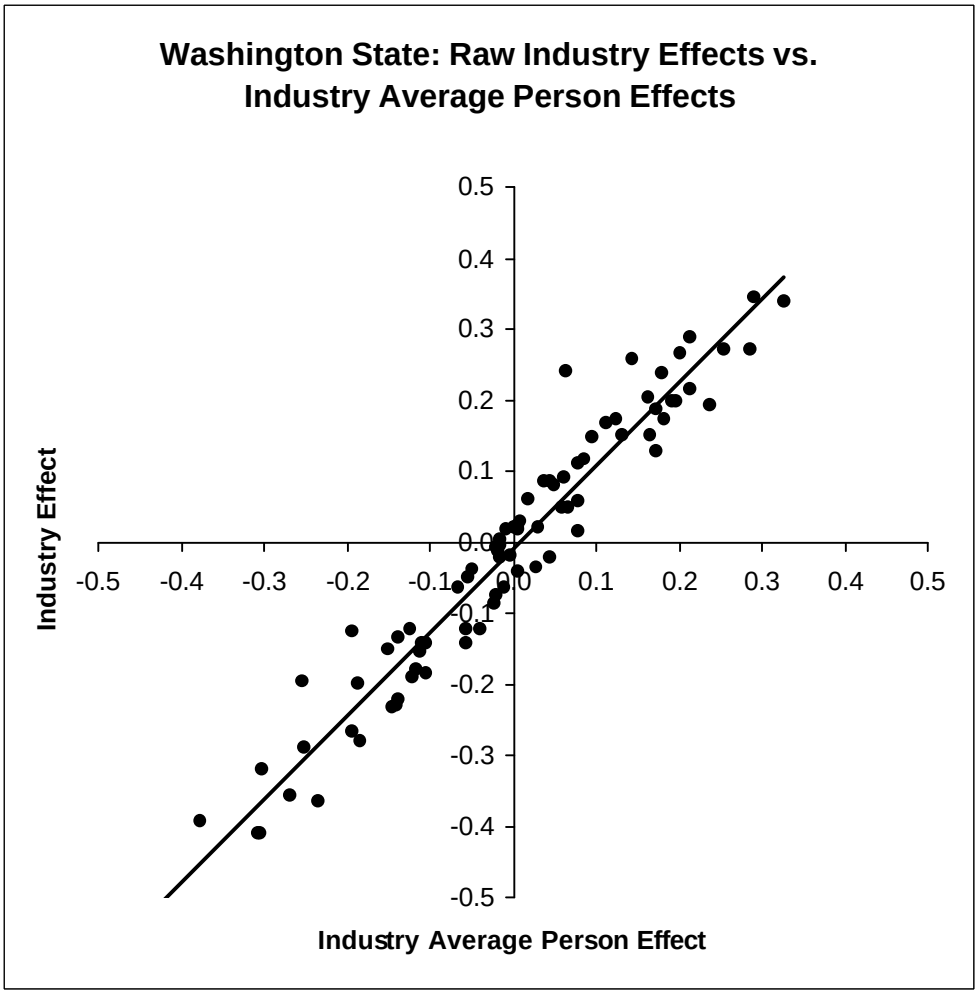
Washington State: Raw Industry Effects vs.
Industry Average Person Effects
Industry Effect
Industry Average Person Effect
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5

Figure 1

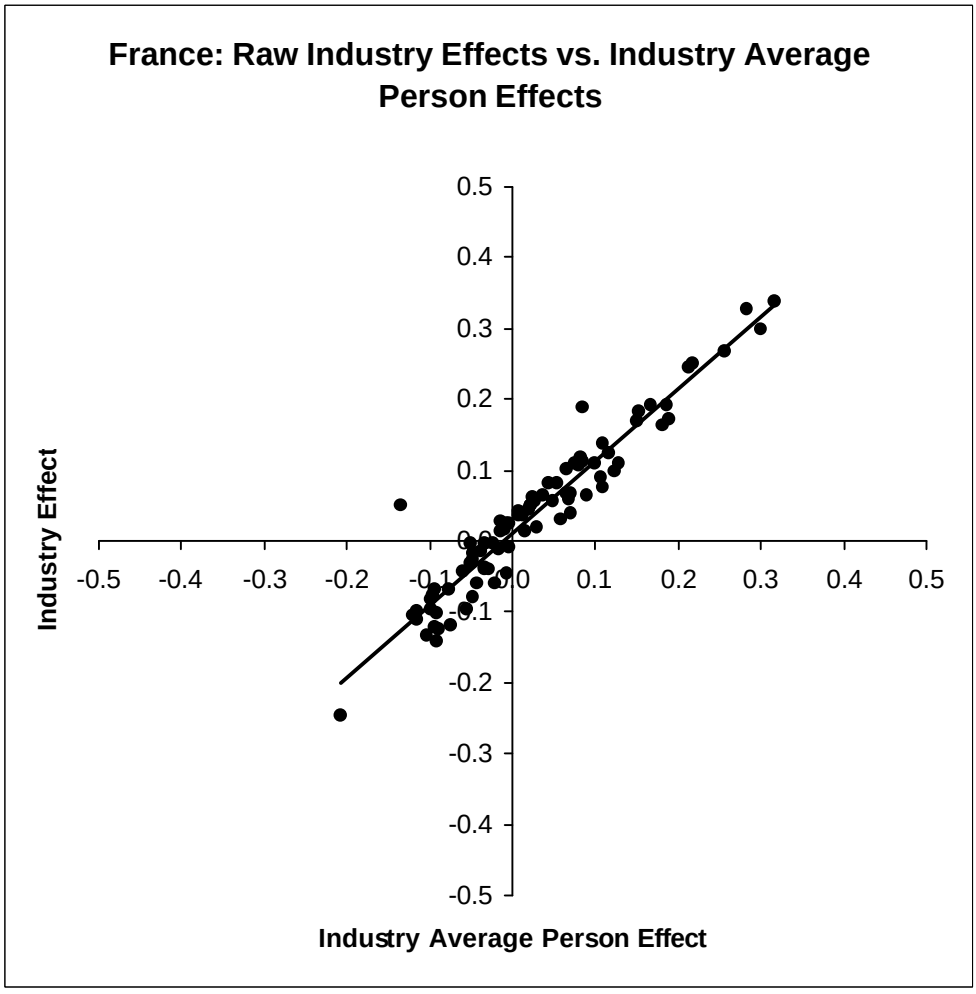
France: Raw Industry Effects vs. Industry Average Person Effects
Industry Effect
Industry Average Person Effect
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.5

Figure 2

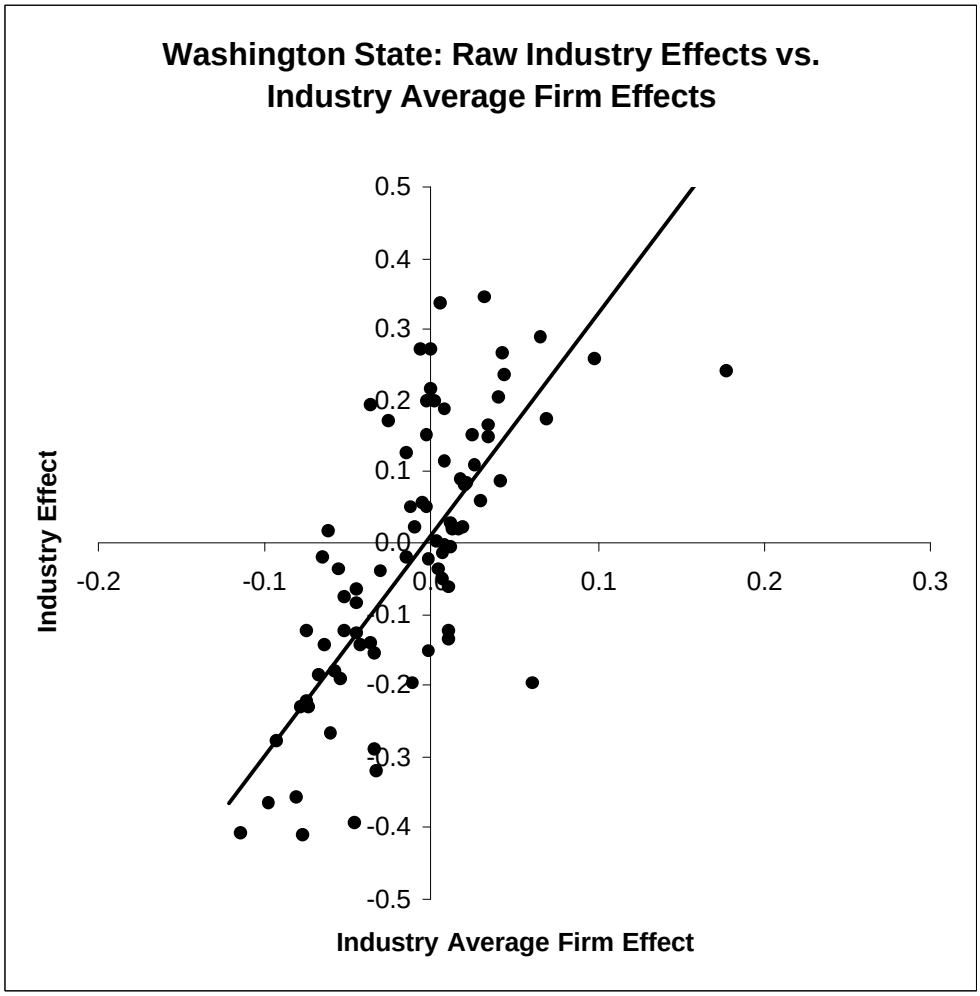
Washington State: Raw Industry Effects vs.
Industry Average Firm Effects
Industry Effect
Industry Average Firm Effect
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.2
-0.1
0.1
0.2
0.3

Figure 3

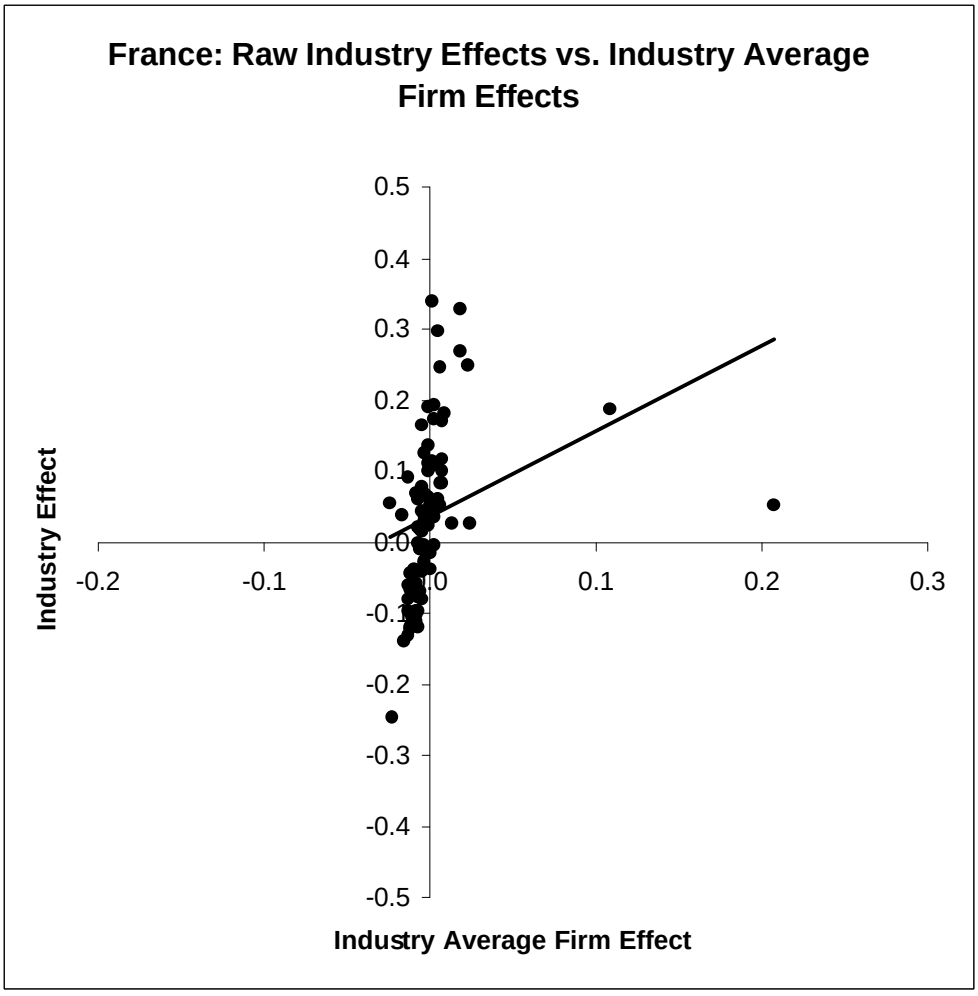
France: Raw Industry Effects vs. Industry Average Firm Effects
Industry Effect
Industry Average Firm Effect
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.2
-0.1
0.1
0.2
0.3

Figure 4