New Jersey
Education to Earnings
Data System

Strategies to Reduce Error in Annualized Unemployment Insurance Wages

Sean Simone, Ph.D.

March 2022

New Jersey Office of the Secretary of Higher Education
New Jersey Department of Education
New Jersey Department of Labor and Workforce Development
New Jersey Higher Education Student Assistance Authority

RUTGERS
Edward J. Bloustein School
of Planning and Public Policy
JOHN J. HELDRICH CENTER
FOR WORKFORCE DEVELOPMENT

Contents

Introduction

The proliferation of statewide integrated data systems linking data across state agencies provides opportunities to examine the efficacy of state programs. Since all states run Unemployment Insurance (UI) compensation programs, wage data collected from employers can be a valuable outcome measure. However, research using a state's UI wage records can introduce biased results if an analyst is not careful. The primary purpose that states have in collecting wage records data is to administer UI, so any use outside of that purpose may bias estimates from populations of interest. While UI records include the majority of wages earned within a state, there are significant limitations. including residents working in another state, federal employees, those who are self-employed, and others with jobs that do not participate in the UI system (Gosa et al., 2016; Stephens, 2007). The purpose of this paper is to document the best method for using and computing wage estimates using data-integrated longitudinal data systems with UI wage records.

Background and Problem

Since the late 1990s, federal and state governments have tried to leverage existing administrative data for use in research. The Longitudinal Household-Employer Dynamics program at the U.S. Census Bureau started around 2000 and has robust national data on wages and employers representing all 50 states and the District of Columbia (Jones-Ruiz, 2020). It is a partnership between states and the federal government where states provide data from their unemployment compensation programs to the U.S. Census Bureau for research purposes. The data are used to provide states and researchers with labor market information, track changes in employment by industry and over time, and examine commuting patterns in the workforce. From a national perspective, there is a high overlap between those who participate in the labor market and what is reported in state UI systems collectively, but there are some exclusions. According to the Bureau of Labor Statistics (2021), the data represent nearly 95% of civilian jobs nationally. At the state level, however, coverage can decrease. In Northeastern and Mid-Atlantic states, for example, up to 20% of the workforce live in one state but work out of state. This can be challenging for states that only have access to their own data.

Table 1 illustrates the issues with coverage when examining UI wages within state borders. It illustrates two challenges with coverage:

1. A state's UI system excludes job classifications such as federal employees, those that are self-employed, farm workers, and others; and
2. State residents working out of state are missing, but out-of-state residents working in the state may not be of interest for many policy questions and can be difficult to exclude without additional data sources.

If an analyst would like to analyze their state's residents using UI wage records, they would uncover coverage errors when analyzing the data. Unemployment compensation data can include out-of-state workers and some state residents are missing if their employer is exempt from UI (e.g., federal employees, farm workers, self-employed, etc.)

Table 1: Comparison of In Wages versus UI Wages, by Workplace Location

	Where people live	Where people work
All wages	State residents' wages, including non-covered UI jobs (federal employees, agriculture, self-employed, etc.)	In-state and out-of-state residents'wages, including non-covered UI jobs (federal employees, agriculture, self-employed, etc.)
UI wages	State residents' UI wages	In-state and out-of-state residents' UI wages

As a result, there are several challenges in using the data at the state level because of missing data and the inclusion of non-residents that are out-of-scope for the analysis:

- Missing generation processes are difficult to identify because states cannot easily share data across state lines. When a wage record is missing, it could indicate that the claimant did not work, they did work but were not in a job covered by the UI system, they had seasonal work (where there is no wage for one quarter), or they worked out of state. As such, imputing a value for these missing records could introduce biased estimates.
- Wages are reported quarterly and many state systems cannot disaggregate to weekly, monthly, or annual wages, which is a more common metric.
- Similarly, converting quarterly wages to other measures can also yield inaccurate information.

These issues are especially problematic when deriving measures, especially when converting a quarterly wage to an annual measure. Although an annual measure is more understandable to the general public, missing data and coverage issues can bias annual wage calculations downward (if zeros are imputed for missing values) or upward (if cases are dropped).

An issue brief from the National Center for Education Statistics (NCES) (Gosa et al., 2016) documents methods for annualizing wages. However, aside from brief warnings cautioning researchers on using some calculations, there is not sufficient research on the validity of such measures. The four methods outlined in the brief are summarized below:

In-state Annual Wages. This is calculated by summing all of the wages for graduates and dividing them by the number of graduates. A median can also be generated from these data. This method yields the lowest estimate of in-state wages because it includes all part-time workers and graduates with zero wages because of work out of state or in jobs not covered by the UI system (Gosa et al., 2016).

Annualized Wages (consecutive quarters). The measure ignores all graduates without four consecutive quarters of wages. For those with four consecutive quarters of wages, the sum of all wages is divided by the number of graduates with consecutive quarters. A median can also be computed (Gosa et al., 2016). In New Jersey, between 20% and 30% of the graduates from 2008 to 2013 are dropped from the analysis. This level of dropped cases concerns New Jersey Education to Earnings Data System (NJEEDS) analysts. It is not known what the impact of these dropped cases has, and whether there is any bias created by reporting on a smaller set of records. This metric does not account for part-time work or out-of-state work.

Multiply Quarterly Wages by Four. This method does not require consecutive quarters of work and simply multiplies quarterly wages by four. Analysts can then obtain a mean or median for this group. Gosa et al. (2016) note that the "… calculation can be performed quickly, but it may not accurately reflect annual earnings for short-term workers or for industries where wages vary seasonally, such as retail and construction."

Full-time Equivalent Wages. Another method documented in the NCES issue brief uses the minimum wage to identify staff likely to be full time. Some use the state minimum wage and others use the federal minimum wage. For example, in New Jersey, which is the source data used for the analysis in this paper, the full-time equivalent annual wages must be greater than or equal to the following values in order for means or medians to be computed:

- For 2009, a minimum wage of $7.19 per hour yields a presumed annual full-time salary of $14,955.
- In 2012, a minimum wage of $7.25 per hour yields a presumed annual full-time salary of $15,080.
- The current minimum of $10 per hour yields an annual full-time salary of $20,800.
- The federal minimum wage method using $6.55 per hour, which can be comparable across states, yields $13,624 annually.

In all cases, the face validity of these thresholds for full-time wages is weak given the higher cost of living in New Jersey. Some states that collect the number of hours worked in their UI data collection (New Jersey does not) are able to calculate actual full-time wages.

Using data from New Jersey's state longitudinal data system, Table 2 provides a percentage distribution of graduates with no wage records, some quarters of wage records, and four consecutive quarters of wage records in order to document the loss of records when computing a wage. There were 322,302 graduates between the 2008 and 2013 academic years with wage matches in the UI records up to five years after graduation. Of those, approximately 40% of records would be dropped if data are annualized using the **Annualized Wages** method. If those with no wage records for a year are excluded, approximately 20% to 30% of cases are dropped. As a general guideline, anytime the number of missing records exceeds 20%, researchers must be concerned about bias and should conduct a missing data and bias analysis to test if the findings are valid (Office of Management and Budget, 2006, p. 8).

Table 2: Count and Percentage Distribution of Graduates by Number of Quarters Employed in Each Year for Five Years after Graduation of New Jersey Graduates from Two- and Four-year Public Institutions, 2008–13

	Count	Percentage of All Graduates	Percentage of Graduates with Wage Records
Total	322,302	100.0	–
Year 1			
Wage records with no data during year	42,755	13.3	–
Wage records without four consecutive quarters of data	87,524	27.2	31.3
Wage records with four consecutive quarters of data	192,023	59.6	68.7
Year 2			
Wage records with no data during year	55,604	17.3	–
Wage records without four consecutive quarters of data	67,111	20.8	25.2
Wage records with four consecutive quarters of data	199,587	61.9	74.8
Year 3			
Wage records with no data during year	63,640	19.7	–
Wage records without four consecutive quarters of data	59,305	18.4	22.9
Wage records with four consecutive quarters of data	199,357	61.9	77.1
Year 4			
Wage records with no data during year	70,777	22.0	–
Wage records without four consecutive quarters of data	52,299	16.2	20.8
Wage records with four consecutive quarters of data	199,226	61.8	79.2
Year 5			
Wage records with no data during year	77,347	24.0	–
Wage records without four consecutive quarters of data	47,139	14.6	19.2
Wage records with four consecutive quarters of data	197,816	61.4	80.8

Source: NJEEDS, Higher Education and Labor and Workforce Development tables, 2019.

Records dropped for these computations are missing not at random (MNAR). Median quarterly wages for those graduates employed full year are higher than for those employed fewer than four quarters. For example, four-year college graduates employed full year earn between $3,400 to $11,000 more per quarter when compared to earnings of four-year college graduates employed for two quarters in the five years after completion.

When comparing annualized wages with national estimates, it is difficult to discern which computation yields the most accurate estimate. Table 3 displays estimates that compare a nationally representative sample from a study of Bachelor's degree graduates produced by NCES (2008/12 Baccalaureate and Beyond Study) and an analysis of NJEEDS data restricted to graduates of four-year institutions who completed an undergraduate degree in 2008. Staff computed annualized wages employing the **Full-time Equivalent Wage** method using the New Jersey state minimum wage from 2008 to 2018 using two different assumptions. In one version, all records that were below the minimum wage threshold for each quarter were excluded (minimum wage multiplied by 40 hours per week times 13 weeks). Using the minimum wage threshold by quarter allowed retention of some records that would likely be dropped if the annual minimum wage threshold was used as these records were part year. In the second version, only full-year graduates (four quarters) were

included. As shown in Table 3, one computation overestimated wage and one underestimated wage. The overestimation of wage is plausible since New Jersey as a state has higher wages, but more research is required to verify which method yields the best estimate of an annual wage.

Table 3: Median Wage Comparison of 2008 Four-year College Graduates (undergraduate degree program only) One Year After Completion

Source/Methodology	Wage in 2009 Dollars
Baccalaureate and Beyond data (national data set)	$36,000
NJEEDS using state minimum wage threshold for all graduates (full year and part year)	$27,500
NJEEDS using state minimum wage threshold for full-year employed graduates	$42,450

Source: 2008–12 Baccalaureate and Beyond Survey (B&B: 2008/2012), NCES, and NJEEDS, Higher Education and Labor and Workforce Development tables, 2019
Notes: The margin of error for the B&B: 2008/2012 annual wage estimate is ±$450

The following research questions guide the analysis in this paper:

1. Which method used to compute annualized wages yields the least amount of bias?
2. Which method yields an estimate that is less than 10% away from the population value?

Methodology

This study uses Monte Carlo simulations to generate samples to estimate differences between computed and actual wages in a fabricated population data set. To simulate how each method performs, simulations introduce data generation processes that are MNAR. After the data sets are created, simulations generate multiple random samples comparing each of the computational methods outlined in the NCES paper:

- In-state annual wages,
- Annualized wages using records with four consecutive quarters,
- Multiple wages from one quarter by four, and
- Full-time equivalent wages (Gosa et al., 2016).

It is presumed that the analysis focuses on higher education graduates from a cohort of Associate degree graduates and Bachelor's degree graduates. The following assumptions are used in deriving the fabricated data sets. These values are used to introduce population values that induce MNAR estimates when missing data are introduced:[1]

Population Annual Wage Values. The annual wages in the population are set at the latest average values from national surveys from NCES (2010, 2011). They include:

- Full-time Bachelor's degree graduates: $58,212 per year
- Part-time Bachelor's degree graduates: $24,781 per year
- Full-time Associate degree graduates: $41,007 per year
- Part-time Associate degree graduates: $31,490 per year

[1] The percentages of self-employed, seasonal, and out-of-state workers are higher than are found in states with the highest rates. These high percentages were selected so it is easier to identify methods that yield the least bias. Differences between the estimate and the population go down as these percentages are reduced.

Self-employment. For workers that are self-employed, it is assumed that the salary is approximately $10,000 below the average rate for the graduate's degree level. Twelve percent of the population are assumed to be in self-employed work.

Seasonal Work. It is assumed that 20% of Bachelor's degree graduates and 15% of Associate degree graduates participate in seasonal work. The percentage of those employed will be variable by quarter to account for seasonal work with higher rates of non-work during the summer and winter quarters, similar to what is found in the economy.

Out-of-the-State Workforce. It is set that 20% of the population are not working, and therefore unlikely to show up in state administrative data. It is also assumed that 30% of the population will be working outside of state lines.

Full Time/Part Time. It is assumed that 10% of workers with Bachelor's degrees and 18% of workers with Associate degrees will be employed part time after graduation.

Using these assumptions, a data set with the actual values for the population is created to determine the actual value for comparison purposes. Then, for each computation, 1,000 random simulation data sets of 30,000 observations each are created with average and median wage estimates across the simulated data sets. These estimates are compared with the population values data set.

Nonresponse bias computations are adapted to measure differences between population and estimated values. Nonresponse bias is when a statistic from a data set overestimates or underestimates a population value. Nonresponse bias, for example, is measured as follows:

$$\hat{B}(\overline{y}_R) = (\overline{y}_R - \hat{\pi})$$

where bias ($\hat{B}$) of the respondent means ($\overline{y}_R$) is the function of the respondent mean minus an **estimate** of the population parameter ($\hat{\pi}$), or the full sample mean.

Relative bias is a measure of bias relative to the population value (in percentage units):

$$\widehat{RB}(\overline{y}_A) = \frac{100\ x\ (\overline{y}_A - \pi)}{\pi}$$

where relative bias ($\widehat{RB}$) of the administrative data estimate means ($\overline{y}_A$) is the function of the respondent mean minus the **actual** population parameter (π), or the observed measure from a census divided by the population parameter multiplied by 100. Relative bias is interpreted as a percentage. A relative bias of zero means that the estimate and the population values are the same. A relative bias of 25 means that the estimate is 25% higher than the estimate and a relative bias of -25 means that the estimate is 25% less than the population estimate. This study seeks to understand which computations represent the population value the best, and if so, which estimates are within 10% of the population value.

Results

The results include eight sets of Monte Carlo simulation estimates for each computational method. They include the:

- Percentage of Associate graduates employed,
- Percentage of Bachelor's graduates employed,
- Mean/median wages excluding missing records for Associate graduates,
- Mean/median wages excluding missing records for Bachelor's graduates,
- Mean/median wages imputing those with missing records with zero wages for Associate graduates, and
- Mean/median wages imputing those with missing records with zero wages for Bachelor's degree graduates.

The estimates of the percentage of graduates employed are presented in Table 4. These estimates include missing records in the denominator but not in the numerator. The true percentage of graduates who are employed are 81% for both Associate and Bachelor's degree graduates. The average estimates from the simulations are 64% for Associate graduates and 58% for Bachelor's graduates. The relative bias ranges indicate that these estimates underreport the actual values by 22% to 28%.

Table 4: Monte Carlo Simulation Percentage Employed Estimates and Relative Bias for Associate and Bachelor's Degree Graduates

	Mean	Relative Bias
Associate Degrees		
Actual	81.3	
Estimate	63.5	-21.9
Bachelor's Degrees		
Actual	80.6	
Estimate	58.4	-27.5

Table 5 displays the estimates for mean and median annualized wage estimates for each computational method. The first set of methods excludes missing cases but examining annual wage computation produces the closest estimates despite these missing cases. The second set imputes zero for missing cases. For both Associate and Bachelor's degree cohorts, imputing zero yielded higher relative biases compared to excluding missing records. No estimates that impute zero were below the 10% goal set for the study. For estimates where missing cases are excluded, multiple estimates had a relative bias below 10% for both mean and medians. Multiplying quarterly wages by four and the full-time equivalent wage method performed better than the other methods yielding the lowest relative bias. For Associate degree graduates, the full-time equivalent computation produced estimates within five percentage points with multiplying quarterly wage by four, yielding estimates within eight percentage points. For Bachelor's degree holders, the relative biases were further from the estimate. For mean wages, the relative bias was 27.5% below the population value. For medians, the relative bias was more than 10% less than the population estimate. Both multiplying quarterly wages by four and the full-time equivalent wages were comparable in terms of the relative bias.

Table 5: Monte Carlo Simulation Mean, Median Annual Wages, and Relative Bias for Associate and Bachelor's Degree Graduates by Estimation Method

	Mean	Relative Bias (mean)		Median	Relative Bias (median)	
Associate Degree (missing cases excluded)						
Actual	38,598			32,598		
Annual wage (sum of quarters)	34,547	-10.4		29,254	-10.2	
Annualized wage (four consecutive quarters)	34,563	-10.5		29,391	-9.8	*
Quarterly wage x 4	35,801	-7.2	*	30,117	-7.8	*
Full-time equivalent wage	37,122	-3.8	*†	31,147	-4.4	*†
Bachelor's Degree (missing cases excluded)						
Actual	54,598			45,736		
Annual wage (sum of quarters)	49,706	-10.6		38,847	-15.1	
Annualized wage (four consecutive quarters)	49,161	-10.0		39,113	-14.5	
Quarterly wage x 4	50,580	-7.4	*†	40,115	-12.3	
Full-time equivalent wage	50,507	-7.5	*	40,151	-12.2	†
Associate Degree (missing cases imputed with zero)						
Actual	31,375			27,543		
Annual wage (sum of quarters)	21,666	-30.9	†	18,388	-33.2	
Annualized wage (four consecutive quarters)	16,403	-47.7		0	-100.0	
Quarterly wage x 4	18,894	-40.0		11,934	-56.7	
Full-time equivalent wage	21,475	-31.6		18,903	-31.4	†
Bachelor's Degree (missing cases imputed with zero)						
Actual	44,004			38,047		
Annual wage (sum of quarters)	28,386	-35.5	†	20,169	-47.0	
Annualized wage (four consecutive quarters)	21,513	-51.1		0	-100.0	
Quarterly wage x 4	24,330	-44.7		0	-100.0	
Full-time equivalent wage	28,186	-35.9		20,639	-45.8	†

* Estimate is within 10% of the actual value

† Lowest relative bias for the measure

Across all computations, the full-time equivalent wage method had the lowest relative bias more than the other methods with five of the eight estimates. Computing an annual wage by summing quarters produced the lowest bias for two of the eight estimates. Multiplying wage by four produced the lowest wage one time.

Conclusion

Overall, most methods used to compute an annualized wage produced poor estimates with high relative bias. The results show that the poorest performing computations are ones where missing values are imputed with zeros. Mean and median wage estimates ranged from 30% to 100% below the value for the population. Similarly, percentage estimates also yielded biased estimates ranging from 20% to 30% below the population value. Producing means generally performed better than producing medians, which can be skewed if there is a high level of MNAR data.

Six computations fell below the 10% threshold reflected in this study's second research question. All six computations excluded missing values (rather than imputing missing values with zero). Both quarterly wages multiplied by four and the full-time equivalent computations yielded the lowest bias estimates. Mean and median estimates for Associate degree holders and means for Bachelor's degree holders produced the lowest relative bias. Mean and median estimates were within 4.5% for both the mean and the median for Associate degree graduates using the full-time equivalent method. For Bachelor's degrees, both methods were within 7.5%.

Given these results, three recommendations are provided to researchers using state administrative data using unemployment compensation wage files:

1. Imputing missing values with zero should be avoided. Similarly, computing percentages, including those with missing values, can also produce biased estimates. Interpretation of percentages should be made with caution or with clarification. If using a percentage estimate, it should be clear that it only measures those that obtain a job within state borders.
2. While medians may be a better representative of income/wage data when complete data sets are available due to a positive skew in the distribution, they do not perform well when data are missing in this context. The preferred measure is to use a mean or average.
3. Generally, using a full-time equivalent wage is the best representation of an annual wage. It was robust for mean estimates and performed better compared to others. It yielded the lowest bias estimate more than half of the time.

References

Bureau of Labor Statistics. (2021). *Handbook of methods.* https://www.bls.gov/opub/hom/home.htm

Gosa, K., McGrew, C., & Sellers, J. (2016). *SLDS issue brief: Best practices for calculating employment and earnings metrics.* https://slds.grads360.org/services/PDCService.svc/GetPDCDocumentFile?fileId=24166

Jones-Ruiz, K. (2020). What began as a small startup is now a successful partnership between states and U.S. Census Bureau. *America counts: Stories behind the numbers.* https://www.census.gov/library/stories/2020/01/census-bureau-innovative-lehd-program-local-workforce-dynamics-turns-20.html

National Center for Education Statistics. (2010). *2004–09 Beginning Postsecondary Students Longitudinal Study* (BPS:04/09). https://nces.ed.gov/surveys/bps/

National Center for Education Statistics. (2011). *2008–18 Baccalaureate and Beyond (B&B).* https://nces.ed.gov/surveys/b&b/

Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys.* https://unstats.un.org/unsd/dnss/docs-nqaf/USA_standards_stat_surveys.pdf

Stephens, D. W. (2007). *Employment that is not covered by state unemployment insurance laws.* https://www2.census.gov/ces/tp/tp-2007-04.pdf

About NJEEDS

The **New Jersey Education to Earnings Data System** (NJEEDS) is the State of New Jersey's centralized longitudinal data system for education and workforce data. Its mission is to safely use the state's existing administrative data for evidence-based policymaking. Developed in 2012 through a grant from the U.S. Department of Education, NJEEDS creates a single place where state education, postsecondary education, employment, and workforce longitudinal data are securely stored to help stakeholders make data-informed decisions to improve student learning and labor market outcomes. The data system is owned by the State of New Jersey and operated by the John J. Heldrich Center for Workforce Development at Rutgers, The State University of New Jersey.