The BLS wage query system: a new tool to access wage data

A new Internet tool available on the BLS website makes it easier than ever to access wage data by area, occupation, and work level

Maury Gittleman
and
William J. Wiatrowski

The search for wage data can be daunting. Data are available for different job characteristics, such as occupation, industry, or geographic area; by demographics of the wage earner, such as race, sex, education, or age; and in a variety of forms, such as hourly wages, annual salaries, total employer payrolls, gross pay, or net pay. Beyond these variations, users of wage data may ask how *wage* is defined in the measure. Is it straight time or does it include overtime? Are other cash payments, such as commissions or year-end bonuses, included? Finally, how reliable are the data? Have they been subjected to the scrutiny of statistical methods? Do they include sufficient observations to support generalizations about wages in the marketplace?

There is no panacea to simplify these complexities or to ensure appropriate application of available data. Users of wage data are advised to learn as much about their data source as possible—in particular, whether the data use the appropriate definition and meet the standard of quality required for their purpose. The Bureau of Labor Statistics publishes a number of different wage measures. To enable data users to find hourly wage data more easily, BLS recently added a new feature to its Internet site—the wage query system. This interactive application allows users to request wage data from the National Compensation Survey (NCS) by certain characteristics. Once they have targeted the specific data, the results are returned almost instantly.

This article provides information on the data behind the new query system, a section on navigating the system, and a discussion on the new regression estimates that recently were added to the query system. Regression estimates help to provide more complete data on area wages by occupation and level of work, an important component of the wage query system. The article concludes with a look at enhancements planned for the future.

Data drive the wage query system

The data behind the new query system come from the National Compensation Survey, which is a BLS survey of wages and benefits throughout the United States. Although the NCS database includes employer costs for wages and benefits, rates of change in those costs, and detailed information on benefit plans, this discussion is limited to the query of hourly wage rates.[1] BLS also publishes other wage measures, each with its own unique characteristics. The feature that sets the NCS wage estimates apart from other wage data currently available is information on "work level." Not only can data users search for the average wage of, for example, accountants in Los Angeles, they also can select by work level to view average wages of entry level or senior accountants in that locality.

The wage query system presents data tabulated in the same manner as all NCS publications. The user is asked to select among

Maury Gittleman is an economist in the Compensation Research and Program Development Group (call 202–691–6318, or email at gittleman_m@bls.gov). William J. Wiatrowski is a supervisory economist in the Division of Compensation Data Analysis and Planning (call 202–691–6255, or email at wiatrowski_w@bls.gov).

choices of area, occupation, and work level. To understand the selections the query system offers, it is helpful to review basic features of the NCS survey and its publications.

Area. The NCS is an area-based survey, meaning data are collected only in selected areas of the country, which are designed to represent all areas of the country. The current NCS sample of areas is made up of 154 areas—81 metropolitan areas and 73 nonmetropolitan counties. Wage data are published for about 90 areas annually, including most of the metropolitan areas and a small number of nonmetropolitan counties. The areas are also designed to represent nine broad geographic regions and to represent the United States as a whole. Wage data are published annually for the nine regions and for the United States.

Occupation. Data are collected from a sample of employers within the 154 areas. BLS economists visit these employers and obtain wage and benefit information from a sample of occupations within the establishments. These occupations are classified into one of 480 occupations, based on the duties and responsibilities of the job.[2] Occupations are narrowly defined—there are 13 different categories for engineers, for example, ranging from civil and industrial to petroleum and aerospace engineers. Data are published for as many occupations as possible, given that data exist and that they meet confidentiality and reliability standards.[3] In many large areas, data are published for 150 to 175 occupations; for the United States, data are published for about 450 occupations.

The occupational classification system used to define each job is hierarchical, which means that each detailed occupation is part of larger and larger groupings. The civil engineer occupation, for example, is part of the larger group engineers, architects, and surveyors, which in turn is part of the still larger group professional specialty occupations. That group is part of the composite group professional specialty and technical occupations, which is part of white-collar occupations. Finally, this last category is part of the much larger "all workers" group. If data are not available for a specific detailed occupation, the user may be able to find data for a larger grouping that incorporates that occupation.

Work level. In addition to classifying each occupation on the basis of duties and responsibilities, BLS economists also determine the work level of the occupation. This is intended to differentiate between workers within the same occupation. The level of work is determined by assessing the following nine key job characteristics:

- Knowledge
- Supervisory controls
- Complexity
- Guidelines
- Scope and effect
- Personal contacts
- Purpose of contacts
- Physical demands
- Work environment

For example, there are several possible levels of knowledge, ranging from the knowledge of simple, routine, or repetitive tasks to mastery of a professional field to generate and develop new hypotheses and theories. Points are associated with each level of each job characteristic; the sum of the points for all characteristics determines the overall work level of the occupation. (See exhibit 1 for a complete description of the work level system.)

Presently, wage data are published by occupation and work level, using work levels that correspond to the Federal General Schedule pay system of 15 grades, numbered 1 to 15.[4] Research is underway to determine alternate groupings for publishing data by work level, in an effort to make the distinction between grades more meaningful. For example, several of the lower grades may be combined into an "entry level" category, while upper grades may be combined into a "senior level" category.

Navigating the wage query system

The wage query system is an interface on the BLS Internet website that prompts the user to enter an area, an occupation, and a work level to retrieve an estimate of the average hourly wages derived from NCS data. The query system is located in the NCS section of the BLS website (**www.bls.gov**) at **http://data.bls.gov/labjava/outside.jsp?survey=nc**. On the entry screen, the user first selects an area and an occupation. The query system displays only those areas and occupations for which data are available. The mechanism for entering an area and an occupation are related. If the user chooses an area, the occupation list will show only those occupations for which data are available for that area. Similarly, if the user chooses an occupation, the area list will show only those areas for which data are available for that occupation. These features may be helpful if a user is attempting to find wage data for multiple occupations in the same area or for the same occupation in multiple areas.

Once the user has selected an area and occupation, he or she may select a work level. If wage data by work level are not needed, the automatic default selection is "Overall occupation average (no work level)." At that point, the user can view wage data for the selected area and occupation. If the user needs wage data by work level, he can either designate a specific work level or build a work level by defining each of the nine key job characteristics. In either case, once the work level is determined, the user can view wage data for the selected area, occupation, and work level.

Exhibit 1. Description of work level system

A sample of occupations is selected from each establishment in the National Compensation Survey (NCS). BLS then collects information on the duties and responsibilities involved in these occupations in order to classifiy them into the appropriate detailed occupational categories. In addition, the work level of each selected occupation is determined using the U.S. Office of Personnel Management's Factor Evaluation System, which is the underlying structure for evaluation of Federal General Schedule (GS) employees. The following list includes a brief description of each of the factors:

Knowledge measures the nature and extent of information or facts that the workers must understand to do acceptable work and the nature and extent of the skills needed to apply those knowledges.

Supervision received covers the nature and extent of direct or indirect controls exercised by the supervisor, the employee's responsibility, and the review of completed work.

Guidelines covers the nature of instructions, procedures, and directions and the judgment needed to apply them.

Complexity covers the nature, number, variety, and intricacy of tasks, steps, processes, or methods in the work performed; the difficulty in identifying what needs to be done; and the difficulty and originality involved in performing the work.

Scope and effect covers the relationship between the nature of the work (purpose, breadth, and depth of assignment) and the effect of work products or services both within and outside the organization.

Personal contacts includes face-to-face contacts and telephone dialogue with persons not in the supervisory chain.

Purpose of contacts ranges from factual exchanges of information to situations involving significant or controversial issues and differing viewpoints, goals, or objectives.

Physical demands covers the requirements and physical abilities required by the employee to complete the work assignment.

Work environment considers the risks and discomforts in the employee's physical surroundings or the nature of the work assignment and the safety regulations required.

Within each factor are a number of levels, and each level has an associated written description and point value. The number and range of points differ among the factors. For each NCS occupation, the level and associated point value of each factor is determined on the basis of occupation position descriptions and interviews with survey respondents. The point values are recorded and totaled; the total points determine the overall level (or grade) of the occupation, based on the same 15 levels used for the Federal Government's General Schedule employees. A description of the levels for each factor can be found within the BLS website at the following address: **www.bls.gov/ncs/**.

Using regression techniques, BLS researchers examined the relationship between wages and the nine factors used to determine overall grade level. The analysis showed that several of the factors, most notably knowledge and supervision received, had strong explanatory power for wages. That is, as the levels within a given factor increased, the wages also increased. For additional information see Brooks Pierce, "Using the National Compensation Survey to Predict Wage Rates," *Compensation and Working Conditions*, Winter 1999, pp. 8–16.

Query limitations and complexities

In the NCS, available work levels vary by occupation. For example, clerical workers typically are found in work levels 01 through 08. Alternatively, professional workers typically begin at work level 05 or 07 and can be as high as work level 15. The query system prevents users from requesting data for a work level that is not appropriate for the occupation. In addition, a few occupations—legislators, dancers, artists, athletes, authors, actors, musicians, painters/sculptors, and announcers—are not classified by work level. The Federal Government developed the Factor Evaluation System used in the NCS for the evaluation of white-collar workers. When BLS adopted this system for the NCS, it reviewed the factors to determine their appropriateness for the occupations being surveyed. The nine occupations excluded from the work level process were thought to have other criteria that determined work level and pay. Wage data are available for these occupations, but not by work level.

In some cases, there are insufficient data to publish all work levels for an occupation. For example, of the eight possible work levels for accountants in Miami, in a given year fewer than eight are published. This occurs for two reasons. First, the survey includes only a subset of the occupations in each sampled establishment in a given area, rather than a census of all jobs in every establishment. Second, data for certain work levels may not meet BLS confidentiality and reliability standards. As of June 2001, estimates of average wages for these "missing" work levels within occupations can be obtained using regression models, as described in the section that follows. Wage data by work level displayed in the wage query system are derived either from direct estimation of data or from the regression model. This distinction is clearly marked when users view results of their query.

Model-based estimates

Statisticians use *direct estimation* to produce the series of average wages for area, occupation, and work level that ap-

pear in NCS publications and as part of the wage query system. This method, which refers to the direct computation of an average (or other statistic) using sample data, is the technique used most often in BLS and other statistical agencies. In some cases, however—often because the sample is too small to produce reliable estimates—a different approach is used: *indirect estimation* or *model-based estimation*.[5]

To produce the indirect estimates of hourly wages by area, occupation, and work level that now form part of the wage query system, regression analysis is used. One important aspect of regression methods is that they can be used to produce estimates of *conditional means*, which in this case refer to the average hourly wage for individuals, given the area in which they work, their occupation, and their work level. Clearly, estimates of conditional means are generated by direct methods as well, but there are significant differences between the two techniques.

Before discussing how the regression model works, it may be useful to examine table 1, which displays statistics for hypothetical hourly wage data for three areas (*X*, *Y*, and *Z*), three occupations (*A*, *B*, and *C*) and three work levels (1, 2, and 3). The averages presented have been calculated by the usual method of direct estimation. For the sake of simplicity, employment is distributed evenly across the cells (in the top three panels of the table) that are defined by combinations of these three dimensions. One can see, for example, that the average wages of an individual in area *X*, occupation *A,* and level 1 is $10.00.

The fact that both direct and indirect methods can be used to produce conditional means makes it possible to use this table to give a sense of how the regression model produces its estimates. Before doing so, however, it may be helpful to summarize some key patterns evident in the top three panels of the table. First, for any given occupation and work level, area *Y* tends to have the highest wages and area *X* the lowest wages, while wages for area *Z* are somewhere in the middle. Second, wages by occupation tend to be highest for occupation *C* and lowest for occupation *A*. Third, wages always increase as the level of work increases.

To quantify these trends, one can take an average of the cells by area, occupation, and work level, and then take an average of all cells to obtain a mean for the Nation as a whole. Taking one dimension at a time, one can then calculate differentials with respect to the overall average. For instance, the wages for area *X* are, on average, $1.22 lower than those for the Nation as a whole ($17.56 versus $18.78). Similarly, the wages for occupation *A* are $2.56 lower than the average for all occupations ($16.22 versus $18.78), while those for work level 1 are $5.78 lower ($13.00 versus $18.78).

To provide a simplified example of how the regression model works, let's say one is interested in estimating an average wage for area *X*, occupation *C*, and work level 2. Instead of using the direct estimate in the table, one can construct an estimate in a fashion similar to the way the regression model predicts wages. Using the numbers on the table and making the appropriate subtractions, one sees that average wages in occupation *C* are $2.33 higher than the overall average ($21.11 versus $18.78), and that those in level 2 are $0.11 higher ($18.89 versus $18.78). Remembering that the wages in area *X* are $1.22 lower than the overall average, one can add the differentials to the national average of $18.78, which results in a predicted wage of $20.00 ($18.78 – $1.22 + $2.33 + $0.11 = $20.00).

In this case, the estimate computed indirectly via the model exactly matches the $20.00 that resulted from a direct estimate. Even in this highly artificial example, however, most of the wages predicted by the model would not be exactly right. The reason is that the patterns of wages by occupation and work level are not identical by area. That is, while table 1 was constructed so that the ranking for pay of occupations and work levels is the same for all areas, the exact magnitudes sometimes differ. Thus, the implicit assumption of the model that occupation and work level differentials are identical across areas will, in general, lead to prediction errors.

The regression model used in the wage query system allows wages to differ by area and occupation as in this example. Instead of using work levels as a predictor, however, the model uses scores on the nine factors that are used to calculate the level. Although the example shows that prediction errors come from assuming that differences in wages by occupation and by work level are the same across areas, the regression model used does, in fact, make this assumption.

Table 1. Hypothetical mean hourly earnings by area, occupation, and work level

Items	Area *X*	Area *Y*	Area *Z*	Nation
Occupation *A*				
Level 1	$10.00	$12.00	$11.00	$11.00
Level 2	15.00	18.00	16.00	16.33
Level 3	20.00	23.00	21.00	21.33
Occupation *B*				
Level 1	12.00	14.00	13.00	13.00
Level 2	18.00	20.00	19.00	19.00
Level 3	24.00	26.00	25.00	25.00
Occupation *C*				
Level 1	14.00	16.00	15.00	15.00
Level 2	20.00	22.00	22.00	21.33
Level 3	25.00	28.00	28.00	27.00
Occupation *A*	15.00	17.67	16.00	16.22
Occupation *B*	18.00	20.00	19.00	19.00
Occupation *C*	19.67	22.00	21.67	21.11
Level 1	12.00	14.00	13.00	13.00
Level 2	17.67	20.00	19.00	18.89
Level 3	23.00	25.67	24.67	24.44
Overall	17.56	19.89	18.89	18.78

While the fact that this is not literally true introduces a greater chance of prediction error, not making the assumption means relying on smaller amounts of data to estimate how these areas differ in this regard, which also increases the chances of making inaccurate predictions.[6] It should also be noted that a variety of alternative models were assessed that relaxed the assumption of equality of wage differences by occupation and work level across areas. On average, these models did not have better predictive power than the model that was chosen for incorporation into the wage query system.

Given these errors, one might naturally wonder why it is useful to present estimates generated by the model. First, it is important to keep in mind that even direct estimates contain prediction errors. While they are correct, on average, for the given sample, the average wage is, of course, not the wage that everyone for that job actually receives. In fact, if one could perform a parallel survey, where the respondents are different because the establishments and the occupations within the establishments that are randomly selected are different, the direct estimates also would undoubtedly differ. Second, when using a model, one can combine data from areas with similar labor market patterns to increase the sample size, a process that statisticians refer to as "borrowing strength." While areas can be combined when making direct estimates as well, a model has the advantage of being able to incorporate the ways in which areas differ from each other. Third, a model facilitates the incorporation of auxiliary information to improve the accuracy of its prediction. In this case, using detailed information on factor scores, rather than the work level, which is a kind of summary of the scores, improves the performance of the model.

While it is hoped that this description of where the model-based estimates come from has been of interest (see the appendix for additional technical details), it is not necessary to understand the details of the procedure for generating the estimates in order to make good use of the data. It is important, however, that users know how to view the model-based estimates relative to the directly estimated ones. First, the regression-based estimates should be considered experimental. Though a substantial amount of work has gone into developing, estimating, and validating the model, and such models have a long tradition in the field of labor economics, it has not undergone the scrutiny given to standard BLS products and does not benefit from the years of experience BLS has in direct estimation. Second, the regression-based estimates are being used only in cases where the sample size is too small for direct estimates, indicating greater variability in any estimate, direct or indirect. Work on the model is ongoing, and should, in the future, strengthen users' confidence in the regression-based estimates.

Future enhancements

The BLS wage query system has quickly become a popular Internet tool—nearly 13,000 requests were processed through the system in a recent month. The addition of regression estimates will only enhance the system's usefulness. And BLS is researching additional enhancements as well. Currently, the system is limited to the average wages for all workers in the occupation. Future enhancements will allow users to obtain median and percentile wage estimates, as well as iterations for private sector versus State and local government, and full time versus part time. In addition, some data will be available by union status, industry, and size of establishment. Efforts also are underway to tie the output of the query system to wage escalator calculations from the Employment Cost Index.[7] In this way, detailed occupational wage estimates that may be several months old can be escalated to reflect wage rates in the most recent quarter. □

NOTES

[1] The earnings used to calculate the hourly wage rates are defined as regular payments from the employer to the employee as compensation for straight-time hourly work, or for any salaried work performed. Wage data represent gross pay (that is, prior to taxes) and include incentive pay such as commissions and production bonuses, but do not include overtime or bonuses not directly tied to production, such as hiring and year-end bonuses. For additional details, see *National Compensation Survey: Occupational Wages in the United States, 1999*, Bulletin 2539 (Bureau of Labor Statistics, July 2000). This information is available on the Internet at **www.bls.gov/ncs/**.

[2] Occupations in the National Compensation Survey are defined by the Census Occupational Classification System. The NCS is beginning to reclassify occupations using the new Standard Occupational Classification system. BLS expects to publish NCS wage data with occupations defined using this new system by 2005.

[3] More precisely, for data in a given occupation to meet BLS publication standards, there must be sufficient observations to ensure that no one establishment could be identified, perhaps because data from that establishment dominate a particular estimate. In addition, the relative standard error, calculated as the ratio of the standard error to the mean, must be less than 0.50.

[4] The Federal General Schedule (GS) pay system is used for most white-collar employees of the Federal Government.

[5] Examples of indirect estimation that will be familiar to many BLS data users are the estimates produced by the BLS Local Area Unemployment Statistics (LAUS) program. These data and a description of the estimation methodology may be found within the BLS Internet site at **www.bls.gov/lau/**.

[6] The mean squared error, the measure used to gauge the level of predictive accuracy, is composed of a term for prediction bias and one for the variability of predictions. Restricting certain parameters to be the same across regions imposes some bias, but decreases the variability of the estimates.

[7] The Employment Cost Index is a quarterly measure of the change in employer costs for wages, salaries, and employer-provided benefits. More information may be found at **www.bls.gov/ncs/ect/**.

Appendix: Regression model

The model used to predict wages is of the form

$$W_m = \alpha + \sum_{a=1}^{A-1} \beta_a \, \mathrm{AREA}_{ma} + \sum_{o=1}^{O-1} \chi_o \, \mathrm{OCCUP}_{mo} +$$

$$\sum_{i=1}^{9} \sum_{j=1}^{S_i-1} \delta_{ij} \, \mathrm{FACTOR}_{mij} + \varepsilon_m$$

where W_m is the average hourly wage rate of the mth observation, which is for occupation o in an establishment that is in area a and that has a vector f of scores for each of the nine factors. AREA is a vector of dummy variables indicating area, OCCUP is a vector of dummy variables for occupation, and FACTOR is a matrix of dummy variables representing the different possible scores for each of the nine factors. The corresponding coefficients are β, χ, and δ, while α is a constant term, and ε is the error term. Areas are indexed by a and are numbered from one to A, occupations are indexed by o and are numbered from one to O, i is the index for the nine factors, while S_i is the highest score possible for factor i. The coefficients are calculated by using weighted least squares. An initial weight is determined for each observation by taking into account the probability of selection for the establishment and a given occupation in that establishment, and then corrected for nonresponse. This final employment weight is then multiplied by hours worked per week and weeks worked per year to arrive at an hours weight.

Though it is conventional in labor economics to use the log wage rather than the wage itself, taking logs did not improve the performance of the model significantly. Many different specifications were tried, with most of the variations attempting to see if the predictive accuracy of the model could be improved by allowing either the coefficients on occupation, the coefficients for the factor scores, or both, to vary by area. Using the measures *root mean squared error* and *mean absolute error* to gauge predictive accuracy, it was not possible to find a model that allowed occupation or factor score differentials to vary by locality that substantially outperformed the model.