Log-linear Indexes of Nominal Wage Rates and Employment in Eight U. S. Regions

Eric Blankmeyer
Department of Finance and Economics
McCoy College of Business Administration
601 University Drive
Texas State University – San Marcos
San Marcos, TX 78666

Email eb01@txstate.edu

February 2008

Log-linear indexes

In many areas of economic research, aggregation of goods and services into index numbers is essential to develop models of manageable size and to obtain reliable estimates by reducing the number of unknown parameters and avoiding collinearity among closely-related products. However, conventional index numbers compare prices or quantities over two time periods or between two regions, a procedure that is equivalent to fitting a line to a pair of observations instead of using the entire sample. In general, index numbers are presented as descriptive statistics; and the absence of an inferential framework largely precludes the use of confidence intervals and hypothesis tests available in other areas of quantitative economics.

Drawing on the ideas of Theil (1960) and Banerjee (1977), Blankmeyer (1990) proposes the joint estimation of log-linear price and quantity indexes using a two-way analysis of variance (ANOVA). His central hypothesis is one of proportional variation: if, across two regions, all the prices under consideration were to change in the same ratio, then any item's price would be an exact index of all the prices. Quantities that varied equiproportionally could likewise be indexed by a single item. Given N regions and K goods and services, it would be true that

$$v_{rt} = p_r + q_t + z\,, \qquad r, t = 1, \ldots, N. \tag{1}$$

Here $\exp(v_{rt}) = \Sigma_k P_{rk} Q_{tk}$ is the aggregate monetary value obtained when the K quantities of region t are evaluated at their respective prices in region r; $\exp(p_r)$ is the price level in region r; $\exp(q_t)$ is the quantity level in region t; and

exp(z) is a constant of proportionality. The N^2 sample values v_{rt} are observable, but all the terms on the right-hand side of equation (1) are parameters that have to be estimated. Since the hypothesis of proportional variation cannot be expected to hold exactly, errors e_{rt} must be added to equation (1); they make allowance for the possibility that either prices or quantities may fail to move in lockstep due to effects that are random and self canceling on average.

It is worth mentioning that, when there are only two time periods or regions (N = 2), the contrast $p_2 - p_1$ is simply the logarithm of the ideal price index --the geometric mean of the Laspeyres and Paasche price indexes. Likewise, $\exp(q_2 - q_1)$ is the ideal quantity index. Therefore, the ANOVA model with N > 2 is a generalization of Irving Fisher's formulas. The quantities are the weights for the prices indexes, and the prices are the weights for the quantity indexes.

In this paper, the ANOVA model is used to estimate log-linear indexes of nominal wage rates and employment in eight regions of the United States. A first set of indexes aggregates over 12 retail categories, and a second set of indexes covers 16 service categories. For each category in each region, the "price" is the annual salary per worker averaged over the three years 2003—2005, while the "quantity" is the number of workers employed, also averaged over 2003-2005. The dependent variable in the regression, v_{rt} , is therefore the log of the total payroll when wage rates per worker in region r are multiplied by employment in region t. Thus $\exp(v_{rr})$ is region r's total payroll actually reported in the data set, while $\exp(v_{rt})$ is a counterfactual total payroll if $r \neq t$.

To compute $\exp(v_{rt})$ for the 12 retailing categories, each category's wage rate in region r is multiplied by the same category's employment in region t (r, t = 1, … , 8); and these products are summed across the 12 categories. The same procedure is applied to the 16 service categories. The data set is described and documented in an appendix. In particular, the eight regions used by the U. S. Bureau of Economic Analysis are New England, Mideast (the mid-Atlantic states), Great Lakes, Plains, Southeast, Southwest, Rocky Mountains, and Far West.

Estimation, inference, and robustness

If the unobservable errors e_{rt} added to equation (1) have identical and independent gaussian distributions, then least squares produces maximum-likelihood estimates (MLE) of the regional indexes for wage rates and employment. The MLE for $p_r - p_t$ estimates the log difference between annual wage rates in regions r and t while the MLE for $q_r - q_t$ estimates the log difference in employment between the two regions. As is well known, these contrasts are merely differences between group means. However, we want to explore several variations on the conventional least-squares solution, so it will be convenient to interpret equation (1) in extensive form as a linear regression on a set of regional dummy variables (fixed effects). Then z is the intercept; and a unique solution is obtained by omitting the wage-rate and employment dummies for one of the regions. Specifically, we suppress the dummies for Far West.

Given the MLE for either $p_r - p_t$ or $q_r - q_t$, the standard error of the log difference is estimated by

$$\sqrt{(2s^2 / N)}\ , \qquad (2)$$

where s^2 is the usual unbiased estimate of the error variance and N = 8 regions. For example, to test whether $p_r = p_t$, the MLE of $(p_r - p_t) / \sqrt{(2s^2 / N)}$ has, under the null hypothesis, Student's t distribution with $(N-1)^2$ degrees of freedom. Tables 1 and 2 display the OLS point estimates for retailing and services respectively. In both tables, the regressions are based on $N^2 = 64$ observations v_{rt} ; and R-squared exceeds 0.99, indicating strong support for the hypothesis of proportional variation in wage rates and employment.

Table 1. OLS Estimates of the Retail Indexes

	Log	**Log**
	Wage-rate	**Employment**
Region	**index**	**index**
New England	-0.0695	-1.0466
Mideast	-0.0780	-0.0077
Great Lakes	-0.2013	-0.0233
Plains	-0.2615	-0.7343
Southeast	-0.1932	0.4817
Southwest	-0.1258	-0.3588
Rocky Mountain	-0.1765	-1.4885

Table 2. OLS Estimates of the Services Indexes

	Log	**Log**
	Wage-rate	**Employment**
Region	**index**	**index**
New England	0.0265	-1.0204
Mideast	0.0424	0.1191
Great Lakes	-0.1164	-0.0269
Plains	-0.1900	-0.8527
Southeast	-0.1549	0.3625
Southwest	-0.1431	-0.4903
Rocky Mountain	-0.1588	-1.5817

For example, Table 1 shows that the wage-rate indexes for New England and the Great Lakes have a log difference of -0.0695 – (-0.2013) = 0.1318, so average annual wage rates in New England retailing are estimated to exceed those in the Great Lakes by about 13 percent. On the basis of expression (2), each log difference in Table 1 has a standard error of 0.0016, so a difference of 13 percent is highly statistically significant. It is of course a difference in nominal wages. According to the Bureau of Labor Statistics, the level of consumer prices in New England exceeded that in the Great Lakes region by about 9 percent (for all urban consumers) or 10 percent (for urban wage earners and clerical workers) during 2003-2005. In retailing, therefore, the real wage-rate differential between the two regions is in the range of 3 or 4 percent.

As another example, the standard error in Table 2 is 0.0015; and a comparison of employment in services shows that the log difference between the Plains and the Southwest, 0.3624, is also highly statistically significant. In 2005, the total populations of the Plains and the Southwest were 19.8 million and 35.2 million respectively, a log difference of 0.5752, so the number of service workers per capita is considerably smaller in the Southwest.

To what extent are these results dependent on the assumption of independent and identically distributed gaussian errors in the linear regression model ? In the first place, econometric theory provides a number of tests and adjustments for the types of heteroscedasticity commonly encountered in cross-section data (e. g., Greene 2003, chapter 11). In particular, White's consistent

estimator of the covariance matrix for generalized least squares adjusts the standard errors for heteroscedasticity without altering the estimates of the regression coefficients (Greene 2003, 198-200, 220-221). Expression (2) is no longer applicable, and of course the estimated standard errors are no longer identical. However, it is straightforward to perform hypothesis tests since each log difference is merely a linear restriction on a pair of regression coefficients in the White-adjusted model. For our data set, the heteroscedasticity adjustments have negligible effects on the statistical significance of the various log differences.

The linear regression format also facilitates the use of the bootstrap to validate the point estimates and their standard errors (Efron and Tibshirani 1993, Davison and Hinkley 1997). To preserve the ANOVA structure of the model, one should bootstrap the regression residuals instead of the observations themselves; the wild bootstrap can be used to check for heteroscedasticity (Flachaire 2005). For our data sets, the bootstrap produces some evidence of heteroscedasticity but no indication of bias in the regression coefficients. Again, the OLS results do not appear to be seriously misleading.

Of course, these sensitivity analyses do not have a high breakdown point; in other words, they are not robust if there happen to be several large outliers in the data. Since the ANOVA regressors are dummies, any outliers are confined to the dependent variable; and it seems that they could easily be identified and down weighted by an M estimator of regression (e. g., Rousseeuw and Leroy 1987, 148-150). As Hubert (1997) and Mili and Coakley (1996) have explained,

dummy variables can reduce the breakdown point of regression estimators, including those that are very robust when all the regressors are defined along a continuum of values. Despite this caveat, it seems worthwhile to apply a robust regression method to our index-number model. The algorithm of Yohai, Maronna and Zamar (Insightful Corporation 2002) identifies three moderately large outliers in retailing and one in services. When these observations are down weighted, the estimated regression coefficients are still similar to OLS. Like the White estimator and the bootstrap, robust regression departs from expression (2), potentially producing different standard errors for each log difference. Unlike OLS, the robust regression does not try to accommodate the outliers, so the robust R-squared is smaller: 0.85 for the retail sector and 0.82 for the service sector.

Tests between sectors

Although there is no gain in statistical efficiency since each sector has the same set of regressors (the dummy variables), the seemingly-unrelated regressions procedure (SUR) is a convenient framework for testing hypotheses about the equality of log differences between sectors (Greene 2003, 339-350). The system includes one equation for each sector (e. g., retailing and services). For example, Table 1 shows that the average wage rate in retailing is 6.83 percent higher in the Southeast than in the Plains [-0.1932 –(-0.2615)], while Table 2 shows that the corresponding wage differential in services is 3.51 percent [-0.1549 –(-0.1900)]. SUR indicates that these intersectoral wage disparities cannot be attributed to sampling error alone, since the chi-square statistic with one degree of freedom is 317.46. However, the evidence against

the null hypothesis is tempered somewhat since a sample of 64 observations is hardly of the asymptotic order on which the chi-square test is predicated.

SUR can also be used to explore the aggregation properties of the log-linear indexes. If the raw data on payrolls and employment for retailing and services are combined, then aggregate log-linear indexes of wage rates and employment can be computed; these are displayed in Table 3. Consistent aggregation then requires that, for any pair of regions, a weighted average of the log difference in the retailing wage rate and the log difference in the services wage rate should equal the log difference in the aggregate wage rate. Similarly, a weighted average of the log difference in retailing employment and the log difference in services employment should equal the log difference in aggregate employment between any two regions. In practice, sampling error will prevent these aggregation relationships from holding exactly; but large discrepancies are evidence against consistency in aggregation for the two regions being tested.

Table 3. OLS Estimates of the Aggregate Indexes (Retailing + Services)

	Log	Log
	Wage-rate	Employment
Region	**index**	**index**
New England	0.0097	-1.0247
Mideast	0.0215	0.0988
Great Lakes	-0.1313	-0.0182
Plains	-0.2026	-0.8317
Southeast	-0.1618	0.3836
Southwest	-0.1398	-0.4669
Rocky Mountain	-0.1620	-1.5654

The weights for aggregation might be chosen to reflect the relative importance of each sector in total sales or valued added. Alternatively, the sample data can be used to select the weights. For example, the weight for retailing can be computed by a simple least-squares regression in which the dependent variable is the 14 values in Table 3 minus the corresponding values in Table 2 and the independent variable is the 14 values in Table 1 minus the corresponding values in Table 2. The regression coefficient, 0.174, is the weight for retailing; and the weight for services is $1 - 0.174 = 0.826$.

In the SUR model for all three sectors (retailing, services, and the aggregate), tests of consistency in aggregation can be performed using the computed weights. For example, an hypothesis that wage rates aggregate consistently between New England and the Mideast produces a chi-square statistic whose significance level is 0.725, so consistency in aggregation is not rejected. Similarly, an hypothesis that employment aggregates consistently between New England and the Great Lakes is not rejected since the chi-square statistic has a p-level of 0.312. However, consistency in aggregation is not guaranteed: the p-level is 0.000 for a test that employment aggregates consistently between New England and the Mideast, so the hypothesis is rejected. Since the weights for retailing and services have been estimated from the sample data, a case could be made for bootstrapping the entire SUR aggregation model to assess the adequacy of these hypothesis tests.

Summary

This paper has shown that log-linear indexes are a practical way to summarize certain kinds of cross-sectional data. Because the indexes are computed in a linear-regression framework, a researcher can easily apply the full range of techniques for estimation and hypothesis testing available for the linear statistical model.

References

Banerjee, K. S. 1977. *On the factorial approach providing the true cost of living*. Gottingen, Germany: Vandenhoeck and Ruprecht.

Blankmeyer, E. 1990. Best log-linear index numbers of prices and quantities. *Atlantic Economic Journal* XVIII, 17-26.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

Efron, B. and R. Tibshirani. 1993. *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Flachaire, E. 2005. Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis* 49, 361-376.

Greene, W. H. 2003. *Econometric analysis*. Fifth edition. Upper Saddle River, NJ: Prentice Hall.

Hubert, M. 1997. The breakdown value of the L_1 estimator in contingency tables. *Statistics & Probability Letters* 33, 419-425.

Insightful Corporation. 2002. *S-Plus 6 Robust Library User's Guide Version 1.0*. Seattle, WA.

Mili, L., and Coakley, C. W. 1996. Robust estimation in structured linear regression. *Annals of Statistics* 24, 2593-2607.

Rousseeuw, P. J., and A. M. LeRoy. 1987. *Robust regression and outlier detection*. New York, NY: John Wiley.

Theil, H. 1960. Best linear index numbers of prices and quantities. *Econometrica* 28, 464-480.

Data Appendix

The employment data are from Table SA27 and the salary-disbursements data are from Table SA07 of the Regional Economic Information System of the Bureau of Economic Analysis (BEA) in the U. S. Department of Commerce. Both tables are dated March 2007, and both are based on the 2002 North American Industry Classification System (NAICS). The log-linear index numbers in the paper are computed from the retail and services categories listed below:

LineCode	Retail Categories
701	Motor vehicle and parts dealers
702	Furniture and home furnishings stores
703	Electronics and appliance stores
704	Building material and garden supply stores
705	Food and beverage stores
706	Health and personal care stores
707	Gasoline stations
708	Clothing and clothing accessories stores
709	Sporting goods, hobby, book and music stores
711	General merchandise stores
712	Miscellaneous store retailers
713	Nonstore retailers

LineCode	Service Categories
1200	Professional and technical services
1300	Management of companies and enterprises
1401	Administrative and support services
1402	Waste management and remediation services
1500	Educational services
1601	Ambulatory health care services
1602	Hospitals
1603	Nursing and residential care facilities
1604	Social assistance
1700	Arts, entertainment, and recreation
1801	Accommodation
1802	Food services and drinking places
1901	Repair and maintenance
1902	Personal and laundry services
1903	Membership associations and organizations
1904	Private households

The 2005 population data for the Plains and the Southwest are from the BEA News Release: State Personal Income 2006 (March 27, 2007).

The cost-of-living indexes for the Northeast and the Midwest regions are from the Bureau of Labor Statistics in the U. S. Department of Labor (www.bls.gov/ro1/9140.htm). They are consumer price indexes for all items (1982-84 = 100).

The BEA Regions are listed below.

BEA Regions

New England
- Connecticut
- Maine
- Massachusetts
- New Hampshire
- Rhode Island
- Vermont

Mideast
- Delaware
- District of Columbia
- Maryland
- New Jersey
- New York
- Pennsylvania

Great Lakes
- Illinois
- Indiana
- Michigan
- Ohio
- Wisconsin

Plains
- Iowa
- Kansas
- Minnesota
- Missouri
- Nebraska
- North Dakota
- South Dakota

Southeast
- Alabama
- Arkansas
- Florida
- Georgia
- Kentucky
- Louisiana
- Mississippi
- North Carolina
- South Carolina
- Tennessee
- Virginia
- West Virginia

Southwest
- Arizona
- New Mexico
- Oklahoma
- Texas

Rocky Mountain
- Colorado
- Idaho
- Montana
- Utah
- Wyoming

Far West
- Alaska
- California
- Hawaii
- Nevada
- Oregon
- Washington