Interpreting Bivariate Regression Coefficients: Going Beyond The Average

Dennis Halcoussis, California State University, Northridge, USA
G. Michael Phillips, California State University, Northridge, USA

ABSTRACT

Statistics, econometrics, investment analysis, and data analysis classes often review the calculation of several types of averages, including the arithmetic mean, geometric mean, harmonic mean, and various weighted averages. This note shows how each of these can be computed using a basic regression framework. By recognizing when a regression model is computing one of these averages, students can properly interpret these types of regressions. Similarly, by seeing how these averages fit into a common framework, students can have a better understanding of the different calculations.

Keywords: Econometrics Education, Data Interpretation, Regression Analysis, Geometric Mean, Harmonic Mean, Weighted Mean, Financial Analysis

INTRODUCTION

This note provides several regression-based examples, including the common arithmetic mean, the geometric mean, the harmonic mean, and various weighted averages, and shows the equivalence between the regression specifications and these measures of central tendency. The teaching of these alternative measures of central tendency has been the focus of several recent papers (e.g., Graziani and Veronese (2009), Lann and Falk (2006)). This paper presents alternative ways of finding many of these measures of central tendency that can assist in relating the descriptive statistics portion of a first-year statistics class to the bivariate modeling component typically found towards the end. While we are not advocating that the regression framework is computationally the most appropriate for computing central tendency, we do propose that students should be aware of these relations for purposes of properly interpreting regression coefficients and for better relating their knowledge of descriptive statistics to regression modeling.

ARITHMETIC MEAN

The arithmetic mean is typically computed by summing a series and dividing by the observation count. This is equivalent to the expected value. A regression of the form

$$y_i = c + u_i \quad (1)$$

where c is a constant provides a computation of the arithmetic mean of y_i as c^*, the estimate of c.

Theorem 1: Given a regression of the form in (1), the ordinary least squares estimate of c is the arithmetic mean of y_i.

Proof: Consider an equation like (1) and derive the least squares estimator of c.

$\delta E[(y_i - c)(y_i - c)]/\delta c =$

$2*E(y_i) - 2c = 0 \rightarrow c^* = E(y_i)$ QED.

Although this, in itself, may not seem very interesting and well known to most practitioners, it provides us a starting point for the later discussion.

WEIGHTED ARITHMETIC MEAN

The weighted arithmetic mean is typically computed by multiplying a target series by a weighting series, summing the terms of the product, and dividing by the sum of its weights. Often the weights are assumed to be in the unit simplex and can be summed to one.

The regression form is not quite as transparent. Consider a regression of the form

$$y_i w_i^{\theta/2} = c\, w_i^{\theta/2} + u_i \quad (2)$$

where $w_i^{\theta/2}$ is a weighting series. (A common source of student confusion is to forget that since w is a series, not all elements w_i are necessarily the same.)

Theorem 2: Given a regression of the form in (2), a least squares estimate of the parameter c will be the weighted mean of the target series y_i where the weighting series is w_i^{θ}

Proof: Consider an equation like (2) and derive the least squares estimator of c.

$$\delta E[(y_i w_i^{\theta/2} - c\, w_i^{\theta/2})(y_i w_i^{\theta/2} - c\, w_i^{\theta/2})]/\delta c =$$

$$2*E(y_i w_i^{\theta}) - 2c*E(w_i^{\theta}) = 0 \rightarrow c^* = E(y_i w_i^{\theta})/ E(w_i^{\theta}) \text{ QED.}$$

If $\theta = 0$, this reduces to the standard arithmetic mean. If $\theta = 1$, then the estimated coefficient c^* will be the weighted average using the series w for weighting. A common student error is to multiply the target series by w rather than $w^{1/2}$, essentially using $\theta = 2$, which would result in c^* representing a w-squared weighted average.

Situations where data is analyzed using an implicit $\theta = 2$ are sometimes found in real life but without any evidence that the analyst intended to use a squared-weighting for the computation. Consider, for example, a real estate appraiser attempting to value 100 miles of potential hiking trails from abandoned railroad right-of-way. If the appraiser were to regress selling prices of previous "comparable" transactions against the miles of each previous transaction, there might be a temptation to use the resulting regression coefficient as an estimate of the average price per mile without recognizing that it would actually be a w-squared weighted average. This is similar to an analysis proposed in a recent 7th circuit US Court of Appeals decision by Judge Easterbrook (Guardian Pipeline, LLC v 950.80 Acres of Land et al, 525 F.3d 554;2008 U.S. App. Lexis 9818).

Typically, to use regression to calculate a weighted average one uses the square root of the initial weighting series w, which is the $\theta = 1$ scenario. The case when $\theta = -1$ is a common heteroskedasticity adjustment used with "weighted least squares" in advanced linear modeling classes. (An overview of weighted least squares is provided in Halcoussis (2005), p174-175.)

GEOMETRIC MEAN

The geometric mean, Ψ, is usually defined as the n^{th} root of the product of n observations of a positive valued target series: $(\prod_{i=1}^{n} y_i)^{(1/n)} = \Psi$.

Geometric means can be calculated by way of a regression when the regression is specified in the form

$$\ln(y_i) = c + u_i \quad (3)$$

Theorem 3: Given a regression of the form (3), a least squares estimate of the parameter c will be the arithmetic mean of ln(y) by Theorem 1. The antilog of c^* will be an estimate of the geometric mean.

Proof: $(\prod_{i=1}^{n} y_i)^{(1/n)} = \Psi$. Taking logarithms, $ln\ (\Psi) = (1/n)*\sum_{i=1}^{n} ln(y_i)$ = arithmetic mean of $ln\ (y_i)$. By (1) above, the regression $ln\ (y_i) = c + u$ returns c^* as the arithmetic mean, $ln\ (\Psi)$. Therefore, $e^{c^*} = e^{\psi}$, the geometric mean. QED.

The geometric mean will be given by e^{c*}. Geometric means are commonly used when working with a series of sequential rates, such as inflation rates. The federal government often uses the geometric mean for adjusting various series for differing rates of inflation. This process is referred to as "chain-link pricing." (Clayton and Giesbrecht, pp. 149-51.)

Note that while an adjustment for the nonzero residuals would be needed if the regression were being used to forecast y_i, such an adjustment would not impact the estimate of c^* and therefore is not appropriate here.

HARMONIC MEAN

The harmonic mean, ξ, is the reciprocal of the mean of the reciprocals of a series: $\frac{n}{\frac{1}{y_1}+\frac{1}{y_2}+\ldots+\frac{1}{y_n}} = \xi$.

Harmonic means can be computed using a regression of the form

$$1/y_i = c + u_i \qquad (4)$$

Theorem 4: Given a regression of the form (4), a least squares estimate of the parameter c will be the arithmetic mean of $1/y_i$. The inverse of c^* will be an estimate of the harmonic mean.

Proof: c* is the $E(1/y_i)$, the arithmetic mean of $1/y_i$ by Theorem 1. Since $\xi = 1/E(1/y_i)$ then by definition, $\xi = 1/c^*$. QED.

WEIGHTED HARMONIC MEAN

A generalization of the Harmonic Mean is the weighted harmonic mean, ξ', which is the reciprocal of the weighted average of the weighted reciprocals of a series: $\frac{w_1+w_2+\cdots+w_n}{\frac{w_1}{y_1}+\frac{w_2}{y_2}+\ldots+\frac{w_n}{y_n}} = \xi'$.

Weighted harmonic means can be computed using a regression combining aspects of (2) and (4) as follows:

$$w_i^{\theta/2}/ y_i = c\ w_i^{\theta/2} + u_i \qquad (5)$$

Theorem 5: Given a regression of the form in (5), a least squares estimate of the parameter c will be the inverse of the weighted harmonic mean, ξ'.

Proof: Consider an equation like (2) and derive the least squares estimator of c.

$\delta E[(w_i^{\theta/2} / y_i - c\ w_i^{\theta/2})(w_i^{\theta/2} / y_i - c\ w_i^{\theta/2})]/\delta c =$

$2*E(w_i^{\theta}/ y_i)- 2c*E(w_i^{\theta})= 0 \rightarrow c^*= E(w_i^{\theta}/ y_i)/ E(w_i^{\theta})$.

$1/ c^*= E(w_i^{\theta})/ E(w_i^{\theta}/ y_i) = \xi'$. QED.

Note that when $\theta = 0$, the weighted harmonic mean reduces to ξ, the standard harmonic mean.

While harmonic means are perhaps most commonly thought of in terms of physics problems (e.g., comparing average speed on a trip, computing effectiveness of multiple layers of insulation in an attic, computing equivalent resistance in multiple resistor circuits) some students may appreciate a financial application. When one purchases stocks by investing a constant dollar amount each month regardless of the corresponding price of the stock, some months will result in purchasing more shares and others less. The average price paid per share over

time could be computed by taking the total amount available to invest and dividing by the total number of shares purchased. One can show that this is equivalent to computing the harmonic mean of the prices paid.

An example using weighted harmonic means is to consider a company with two different accounting firms on retainer. One accounting firm is paid $500 a month, another is paid $1000 a month. The amount of services provided each month depends on the billing rates of specific staff. The average price paid for each firm's services is its respective harmonic mean price. The average price paid for an hour of accounting across the firms is the weighted harmonic mean of the firms' prices where the weights are the monthly expenditures per firm.

CONCLUSION

Several common variable transformations used in simple regression specifications are equivalent to computing different types of averages. While the use of regression techniques for computing such measures of central tendency might be viewed as computational overkill, there may be certain circumstances when it is convenient to use a regression program for these computations. More likely, an analyst might specify a regression equation that is a variation of one of these averages. This might be intentional, but we have seen students and practitioners inadvertently propose regression modeling as an alternative to an average when, in fact, their specifications were equivalent to some form of average. Finally, by presenting these basic statistical concepts within the concept of regression analysis, faculty may assist students to better relate the "end of the semester" ordinary least squares model to descriptive methods taught earlier in the semester.

AUTHOR INFORMATION

Dennis Halcoussis is Professor of Economics at California State University, Northridge, where he teaches courses in econometrics and data analysis. He is the author of the popular textbook "Understanding Econometrics" and is the author of numerous papers applying econometric methods to diverse economic topics. He received his Ph.D. from the University of Pennsylvania in 1992. He was raised in the Pittsburgh area.

G. Michael Phillips is Professor of Finance at California State University, Northridge, where he specializes in online education and serves as the Director of the Center for Financial Services and Insurance. His research focuses on valuation in incomplete markets, macroeconomic impacts on financial markets, and advanced forecasting methods. He received his Ph.D. from the University of California, San Diego, in 1982. He was raised in central Illinois.

REFERENCES

1. Clayton, G., and Giesbrecht, M. (2003), *Guide to Everyday Economic Statistics* (6th ed.), New York, NY: McGraw-Hill/Irwin.
2. Easterbrook, F (2008), Guardian Pipeline, LLC v 950.80 Acres of Land et al, 525 F.3d 554;2008 U.S. App. Lexis 9818.
3. Graziani, R., and Veronese P. (2009), "How to Compute a Mean? The Chisini Approach and Its Applications", *The American Statistician, 63(1),33-36*
4. Halcoussis, D. (2005), *Understanding Econometrics* (Mason, OH:South-Western/Thomson)
5. Lann, A., and Falk, R. (2006), "Tell Me the Method, I'll Give You the Mean" *The American Statistician, 60(4),322-327*