P E R S P E C T I V E S

PHOENIX CENTER FOR ADVANCED LEGAL & ECONOMIC PUBLIC POLICY STUDIES

Econometric Analysis of Broadband Subscriptions:

A Note on Specification

George S. Ford, PhD*

May 12, 2009

Broadband subscriptions are data, and where there is data, there is likely an economist or other interested party ready to subject that data to econometric analysis. Econometrics is an exceedingly important tool for public policy analysis, and the application of statistical techniques to broadband subscription data is now old hat. In this regard, the Phoenix Center has used econometrics to develop improved methods for evaluating the performance of OECD countries in terms of broadband subscriptions, shedding new light on the issue of adoption.[1]

Econometric analysis is used to quantify the relationship between one thing and one or more other things. In our work on broadband performance, we consider the role that income, income inequality, education, age, and other factors play in the observed variations in broadband adoption across countries and U.S. states. These models have proven very potent at explaining such variations, implying that much of the variation across geo-political units is explained by differences in economic and demographic endowments.

When estimating such relationships, proper technique is important. If a statistical model is incorrectly specified, then the results obtained may be meaningless, and may improperly guide policymakers. In this PERSPECTIVE, I will discuss one important model mis-specification that is common in the literature evaluating fixed broadband adoption.

The econometric model of broadband subscriptions typically specifies broadband subscriptions (per capita or per household) as a function of such things as price, income, population density, age, education, and so forth. An additional regressor in some models in a measure of "market potential," which is an effort to account for the number of subscriptions at the market saturation point. This last variable is an important one, particularly in cases where saturation rates may differ across countries.[2] My focus here is on this last regressor. While saturation is important to consider, improperly choosing the measure can lead to substantial problems.

In POLICY PAPERS 29, 30, and 33, we measured saturation using the number of (fixed and mobile) telephones per-capita in a country (the variable *PHONE*).[3] My research in POLICY PERSPECTIVES 08-03[4] and 09-01[5] suggests that fixed-line telephone penetration in the mid-1990's may be a preferred measure of market potential until better measures can be created, and I continue to investigate the implications of this alternative.

Other studies have used different measures of saturation. In a study by the Organization of Economic Cooperation and Development (OECD 2007), market saturation is measured as total Internet subscribers (broadband + dialup) in a previous period.[6] A similar measure is used in the report by the Information Technology and Innovation Foundation (ITIF 2008).[7] In this PERSPECTIVE, I demonstrate that including a variable adding broadband and fixed connections, whether contemporaneous or lagged, is a model mis-specification that may lead to significant bias and inefficiency in the estimated coefficients of a regression model. As discussed below, the mis-specification is in the form of a illegitimate parameter restriction.[8] In other words, this specification assumes that broadband and dialup subscriptions have identical effects on broadband subscriptions.

Saturation as a Model Mis-Specification

To ease exposition, say that broadband subscriptions per capita (y) is a function of market potential or saturation (m) and one other variable (x). As with some earlier studies, market potential m is measured as the sum of dialup (d) and broadband subscriptions (b). Note that both d and b could be lagged values and this does not meaningfully impact the analysis under most conditions. Ignoring the observation subscript, we have

$$y = \beta_0 + \beta_1 x + \beta_2 m + \varepsilon\,; \qquad (1)$$

where the β are the coefficients and ε is the econometric disturbance term. Simple substitution allows us to rewrite (1) as

$$y = \beta_0 + \beta_1 x + \beta_2 (d + b) + \varepsilon\ ; \qquad (2)$$

and by the distributive property we have

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_2 b + \varepsilon\,. \qquad (3)$$

Observe in Equation (3) that the coefficient β_2 is attached to both dialup (d) and broadband (b) customers. Thus, this specification includes as a constraint the requirement that the effect on broadband subscriptions of dialup and broadband subscriptions be identical. This legitimacy of this constraint is highly improbable.

For example, if broadband subscription b is contemporaneous to the dependent variable (that is, $b = y$), then the coefficient on b is 1.0 by definition. Yet, the general expectation is that higher numbers of dialup users reduce broadband subscriptions, so that the coefficient on d is expected to be negative ($d < 0$). From Equation (3), we see that the specification requires the two coefficients to be identical. This illegitimate constraint on the coefficient values leads to biased coefficients and inefficient estimates—the worst of all possible worlds in econometric research.[9]

Even if b is a lagged value of y, as in OECD (2007) and ITIF (2008), the coefficient restriction remains improbable. Lagged values in well behaved time series (such as broadband subscriptions) typically have positive values that are often close to 1.0. For the OECD data covering 2005 through the first half of 2008, a regression of broadband subscriptions on lagged subscriptions renders a coefficient of 1.04.[10] Yet, it is clear from the data that broadband subscriptions rise, dialup subscriptions fall, indicating a negative relationship.

What is the effect of such mis-specification? The answer is somewhat apparent in the ITIF (2008) and OECD (2007) studies themselves. One consequence of mis-specification is that some or all of the estimated coefficients of the model are biased toward zero and, consequently, less likely to be statistically significant. In the OECD (2007) report, very few variables are statistically significant. When the saturation variable is excluded from the model, the significance of the remaining regressors rises (e.g., the t-statistic on *AGE* rises from 1.47 to 1.93).[11] Likewise, very little statistical significance is found for the regressors in the ITIF (2008) study.

Monte Carlo Evidence

When the combination of broadband and dialup subscriptions is used as a measure of maturity, the estimated coefficients on the other variables are generally biased (they do not approximate their true values) and they are inefficient (the t-statistics are too small). We can demonstrate this effect with a Monte Carlo study.[12] In Monte Carlo analysis, we create a dataset of known properties so that we can assess the accuracy of statistical techniques.

To being, I make up some data with known properties. Say we have a data generating process ("DGP") described by

$$y = 1.0 + 0.5x_1 - 0.2x_2 + 0.8x_3 + e\,; \quad (4)$$

where x_1, x_2, and x_3 are all uniform random numbers and $\varepsilon \sim N(0, 0.05)$. We generate 100 observations based on this template. It may help to think of x_1 as income, x_2 as dialup, and x_3 and a legitimate measure of market potential (such as telephone subscriptions). Using the data generated by Equation (4), we know exactly how the variable y is formed, and can evaluate whether specific techniques render those results accurately.

The correct econometric specification given the DGP is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + v\,; \quad (5)$$

where v is the econometric disturbance. Estimating Equation (5) 1,000 times using 100 observations each time, we get the average coefficient vector [1.0005, 0.498964, -0.19933, 0.799833].[13] As expected, the correct model renders the correct coefficient values expressed in Equation (4).

Now consider altering the model to include the maturity variable used in some studies, including OECD 2007 and ITU 2008. The practice described above would turn Equation (5) into

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 - \hat{\beta}_2 (x_2 + y) + e\,; \quad (6)$$

where the legitimate maturity variable x_3 is replaced with (x_2 + y), that is (dialup + broadband). Running the model 1,000 times, we find the average coefficient vector is [0.56959, 0.013222, 0.514642]. Focusing on the coefficient β_1 on x_1, the only slope coefficient in both models, we see it is substantially undersized relative to its true value (0.013 versus the true value of 0.50).[14] Increasing the sample size to 1,000 does not help. The average of β_1 falls to 0.006.[15] The coefficient on x_1 is biased.

Econometric Evidence

Using the least squares model specification and data from Policy PAPER NO. 33 as a guide, we can assess the legitimacy of the coefficient constraint from ITIF (2008) and OECD (2007). Since these studies are dated now, the broadband data used to test the coefficient restriction for the period June 2007 (30 observations).

The regression is specified as the natural log of subscriptions per capita regressed on the (natural log of) broadband price, GDP per capita, the GINI coefficient, the percent of the persons age 65 and over, the percent of population living in urban areas, average household size, and the two variables measuring market potential or saturation (dialup and broadband subscriptions).

As in POLICY PAPER NO. 29 and NO. 33, we use weighted least squares regression to account for the dependent variable being heteroscedastic.[16] All the estimated models perform well, with most variables being statistically significant at the 5% level or better. We cannot reject the null hypothesis of RESET or White's test for any of the estimated regressions.[17]

In order to test the validity of the coefficient restriction implied by OECD (2007) and ITIF (2008), we consider two sources for the market potential data. First, we use 2005 data on broadband and dialup users from the OECD, which is data similar to the OECD (2007) study. Using this data, the estimated coefficient on (lagged) broadband subscriptions is 0.405 (t = 5.03) and on (lagged) dialup is -0.022 (t = -0.85). As expected, the coefficient on historical broadband subscriptions per capita is positive and the coefficient on historical dialup is negative. A Wald test that the two coefficients are equal is easily rejected (F = 47.93, Prob < 0.001).

Second, to better match up with the study by the ITIF (2008), we use 2006 data on broadband and dialup users as published by the ITU. The estimated coefficient on broadband is 0.668 (t = 12.91) and on diaup is -0.036 (t = -0.63). Again, the coefficient on broadband subscriptions per capita is positive but the coefficient on dialup is negative. A Wald test that the two coefficients are equal is again easily rejected (F = 51.62, Prob < 0.001).

As mentioned above, in POLICY PAPERS NO. 29 and NO. 33, we measured market potential as the number of telephones, including both fixed and mobile lines. Therefore, we likewise imposed a potentially invalid restraint on the coefficients in our regressions. We can likewise test the validity of this restriction.

Using separate variables for fixed and mobile connections from the OECD for 2005, we test for the equality of the coefficients using the same data and model as above. The coefficient on fixed lines is 0.263 (t = 1.47) and on mobile lines is 0.164 (t = 1.28). The Wald test does not allow for the rejection of the null hypothesis the two coefficients are equal (F = 0.38, Prob = 0.55).[18] Honestly, this outcome is more by luck than by design. Nevertheless, the restriction we imposed is not problematic from a statistical standpoint, while the others most likely are, and this is supported by the low t-statistics in the OECD (2007) and ITIF (2008) reports.

Conclusion

Regression analysis of broadband subscription should take on an increasingly important role for public policy. Broadband is approaching maturity in many countries, yet there remains substantial variability is subscription rates. Much of these differences can be explained by economic and demographic differences, rather than some significant variation in public policy. Econometric analysis can be very useful, however, in quantifying the impacts of known policy interventions, and the Phoenix Center hopes to remain on the frontier of that analysis.

Quantification of any factor on broadband subscriptions requires a correct specification of the econometric models. Here, I have shown that some early work in this area may have suffered from significant specification error, rendering the results of questionable value. We must keep in mind, however, that models can rarely be specified perfectly, and econometric analysis is largely an attempt to minimize the problems with poor or lacking data and inadequate estimation methods. For policy relevance, statistical analysis is often required sooner rather than later, and haste can lead to specification problems.

As we proceed with the analysis of broadband adoption, attention must be paid to sound econometric technique and good model specification. As the data get richer, a wider range of techniques can be applied, and more and more researchers will turn to the data to ask and answer interesting questions.

NOTES:

[*] **Dr. George S. Ford** is the Chief Economist of the Phoenix Center for Advanced Legal and Economic Public Policy Studies. The views expressed in this PERSPECTIVE do not represent the views of the Phoenix Center, its staff, its Adjunct Follows, or any if its individual Editorial Advisory Board Members.

[1] G.S. Ford, T.M. Koutsky and L.J. Spiwak, *The Broadband Performance Index: A Policy-Relevant Method of Comparing Broadband Adoption Among Countries*, PHOENIX CENTER POLICY PAPER NO. 29 (July 2007) (available at: http://www.phoenix-center.org/pcpp/PCPP29Final.pdf); G.S. Ford, T.M. Koutsky and L.J. Spiwak, *The Demographic and Economic Drivers of Broadband Adoption in the United States*, PHOENIX CENTER POLICY PAPER NO. 31 (November 2007)(available at: http://www.phoenix-center.org/pcpp/PCPP31Final.doc); George S. Ford, Thomas M. Koutsky and Lawrence J. Spiwak, *The Broadband Efficiency Index: What Really Drives Broadband Adoption Across the OECD?* PHOENIX CENTER POLICY PAPER NO. 33 (May 2008)(available at: http://www.phoenix-center.org/pcpp/PCPP33Final.pdf).

[2] G.S. Ford, PHOENIX CENTER PERSPECTIVES NO. 08-03 (Second Edition), *Broadband Expectations and the Convergence of Ranks* (October 1, 2008)(available at: http://www.phoenix-center.org/perspectives/Perspective08-03Final.pdf).

[3] *Supra* n. 1.

[4] *Supra* n. 2.

[5] G.S. Ford, PHOENIX CENTER PERSPECTIVES NO. 09-01, *Normalizing Broadband Connections* (May 12, 2009)(available at: http://www.phoenix-center.org/perspectives/Perspective09-01Final.pdf).

[6] J. de Ridder, *Catching-Up in Broadband—What Will It Take?* Working Party on Communications Infrastructure and Services Policy, OECD DSTI/ICCP/CISP(2007)8/FINAL (July 25, 2007)(available at: http://www.oecd.org/dataoecd/34/34/39360525.pdf). A critical review of this study, see G. Boyle, B. Howell, and W. Zhang, Catching *Up in Broadband Regressions: Does Local Loop Unbundling Really Lead to Material Increases in OECD Broadband Uptake?*, Unpublished Manuscript (July 28, 2008) (available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1184339).

[7] R. D. Atkinson, D. K. Correa, J. A. Hedlund, *Explaining International Broadband Leadership*, Information Technology and Innovation Foundation (May 2008) (available at: http://www.itif.org/files/ExplainingBBLeadership.pdf).

[8] W. Greene, ECONOMETRIC ANALYSIS (2003), Ch. 8.

[9] D. Gujarati, BASIC ECONOMETRICS (1995) at 204-7.

[10] This is from a regression $y = a_0 + a_1y_{t\text{-}1}$ for all 30 OECD countries.

[11] *See* OECD 2007, *supra* n. 6.

[12] Gujarati (1995), *supra* n. 9, at 85-86.

[13] The R^2 of the regressions are about 0.98.

[14] A Wald Test on the equality of the estimate to its true value is rejected for 100% of the simulations.

[15] A Wald Test on the equality of the estimate to its true value is rejected for 100% of the simulations.

[16] G. S. Maddala, LIMITED DEPENDENT AND QUALITATIVE VARIABLES IN ECONOMETRICS (1983), 29. This specification is the minimum chi-square method for the linear and log-linear model.

[17] The same size is small, so the power of these tests is likely to be low. A failure to reject must be weighted accordingly.

[18] Using 2003 data from the ITU, similar results are found. I cannot exclude the possibility that this failure to reject is due to low power of the test in small samples, though the test statistics for the alternative measures of saturation are very high.