PERSPECTIVES

PHOENIX CENTER FOR ADVANCED LEGAL & ECONOMIC PUBLIC POLICY STUDIES

A Quality Check on Form 477 Data: Errors, Subsidies, and Econometrics

George S. Ford, PhD*

October 27, 2021

Broadband availability data collected by the Federal Communications Commission ("FCC") using its Form 477 are heavily criticized as inaccurate. These data certainly overstate availability since a census block is indicated as having broadband throughout the block even if only a single home in the block is served.[1] Despite the criticism and inherent flaws, these data are used to distribute billions in subsidy dollars for the construction of broadband networks. Absent a better measure of the "true" extent of broadband deployment, quantifying the errors in the 477 data has been difficult.[2]

Recently, the State of Georgia released the results of a detailed survey in which broadband availability was assessed at all household and business locations. These Georgia data, which provide served and unserved locations by census block, may be viewed as a closer surrogate to the "true" availability rate. Thus, comparing the two data sources permits an assessment of the accuracy of the FCC's 477 data and possibly some of their quirks.

In this PERSPECTIVE, I compare the two data sources. Several findings are worth mentioning.

First, the availability rate by census block from the Form 477 is highly correlated with actual availability. At the total state level, the error in the availability is small (3.5 percentage points), but at the census block level, which is important for identifying where it might be necessary to subsidize broadband deployment, the Form 477 data may both over- and under-state availability.

Second, the error in the availability rate, which for a census block is a percentage in the Georgia data but dichotomous in the Form 477 data, is related (on average) to the geographic size of the census block. This is because census blocks tend to become larger the more rural the geography and the more disperse are residential and business locations.

Third, using the estimated relationship between block size and availability, I estimate there are 14 million unserved locations in the U.S., though about 5 million of these were addressed in the recent Rural Digital Opportunity Fund ("RDOF") auction, leaving approximately 9.1 million unserved locations.

Fourth, I estimate that significant subsidies are required to get broadband to the remaining 9.1 million unserved locations. According to my calculations, if the average subsidy is $2,000 (which is the average of the RDOF auction), then the additional subsidy required to reach unserved households is $18.2 billion. If the average subsidy level is $3,000, then $22.8 billion is needed. And at a very high average subsidy of $5,000, getting broadband to every location requires approximately $45.5 billion.

Finally, I show that the use of the Form 477 in econometric analysis of broadband's effect on outcomes leads to biased coefficients. That said,

using and Instrumental Variables ("IV") approach to account for measurement errors reduces that bias and may be a sensible way to use the 477 data in empirical analysis, though I recommend further work in this area.

The Georgia Broadband Map

Filling in broadband availability gaps requires knowledge of where broadband is located and where it is not. Today, the FCC's Form 477 data are used to do so, but the quality of those data are questionable. The Form 477 data are collected from providers. If any provider has a single customer in an area (or the ability to serve a customer in sixty days), then this is sufficient to label the entire census block as "served." As such, the 477 data will overstate availability, a problem expected to be more serious when census blocks are geographically large. But this source of overstatement is not the only issue with the data. There is also no rigorous check on whether a block labeled "served" is, in fact, served, and no check on whether an unserved block is truly unserved. Also, the (un)availability rates calculated from the Form 477 are based on population (estimates) and not actual residential or business locations where broadband may be provided. As the relationship between population and locations is not a constant, the population-based (un)availability rates may be inaccurate.

Several states have begun their own effort to map broadband availability. Georgia leads the way with a detailed assessment of broadband availability. According to the Georgia Broadband Program's webpage, Georgia's map project:

> represents a location-level methodology that precisely maps the availability of broadband services to every home and business in the State [] by overlaying: (1) all the locations of homes and businesses in the State of Georgia, and (2) broadband provider service availability for those locations within the State. There are over 5 million locations used in the mapping process.[3]

In so doing, broadband services are defined using the 25 Mbps down and 3 Mbps up speed thresholds. The Georgia Map differs from the Form 477 data in that the former provides an accounting of served and unserved locations thus permitting the construction of a "homes passed" rate. Alternately, the Form 477 data provide only a dichotomous indicator of availability at the block level.

... the use of the Form 477 in econometric analysis of broadband's effect on outcomes leads to biased coefficients.

Georgia reports its findings at the census block level thereby permitting a comparison to the Form 477 data, with two caveats. The Georgia data are dated 2021, while the latest 477 data are from June 2020. Nonetheless, it is worth comparing the two datasets. Once broadband is available it presumably remains so, thus a difference between areas labeled "served" in the Form 477 data (in 2020) and the "served" penetration rate from the Georgia map (in 2021) is informative. There may be differences in "unserved" areas, though given the short interval between the two datasets suggests such differences should be small. The second caveat is that the Georgia data report availability based on number of locations within a census block that are served. This may differ from a population-based availability measure that would correspond to Form 477 data.

Data

Several datasets are used in this analysis. First, I use the census block level data from the Georgia Map.[4] There are 291,086 census blocks in these data, though many do not contain either households or businesses (about 34.2%). Second, I use the Form 477 census block data from June-2020 (excluding satellite services).[5] Third, some

census block demographic data is obtained from the Census Bureau.

Definitions

For each census block, the Georgia data provide the number of served and unserved locations (the sum of which is total units) at the 25/3 Mbps level. A location may be either a household or a business. Let s_i and u_i be the number of served and unserved locations in census block i. The estimated "true" location-passed rate, t_i, is the number of served locations divided by total locations

$$t_i = s_i/(s_i + u_i) \quad (1)$$

At the state level, the locations-passed rate is

$$T = \Sigma s_i/\Sigma(s_i + u_i). \quad (2)$$

Form 477 data include only a dichotomous indicator for "served" census blocks (f_i), which is also a statistic for t_i. The statistic for the statewide availability rate, F (the "false" rate), is

$$F = \Sigma[f_i(s_i + u_i)]/\Sigma(s_i + u_i). \quad (3)$$

For census block i, the number of served locations for the Georgia data is s_i and for the Form 477 data is $f_i(s_i + u_i)$, where for any block served according to the FCC data $f_i(s_i + u_i) \geq s_i$.

Comparing the Data

Perhaps the most public use of the Form 477 data is quantifying the share of U.S. population with services available satisfying the FCC's definition of broadband (now, 25/3 Mbps). In the Commission's latest *Broadband Progress Report* (using 2019 data), 95.6% of the population had access to broadband services (93.8% in Georgia).[6] Given the one-location/all-locations assumption underlying the data, this share is certainly overstated. But by how much? This question can be answered in Georgia with some precision.

Rather than population, I measure the availability rate across the two data sources using locations. At the statewide level, the share of locations passed from the Georgia Map (T) is 90.93%, while the Form 477 data indicate an availability rate (F) at locations of 94.16% (based on 2010 population). Thus, the Form 477 data overstates, as expected, availability by about three percentage points; a relatively small difference.[7] Using earlier data, Ford (2019) estimates an overstatement of 5.6 percentage points (using 2019 data) and Busby, Tanberk, and Cooper (2021) estimate an overstatement of 11.2 percentage points.[8] The true difference is smaller than these predictions estimate and is much less than the prediction by Busby, Tanberk, and Cooper (2021).

At the statewide level, and perhaps at the national level, the error in the aggregate availability rates of the Form 477 data may not be very large. The more important use of the Form 477 data is allocating subsidy dollars. Thus, it is errors at the census block level and not the state level that are important. There are two types of errors: Type I and Type II errors. A Type I error occurs when the Form 477 data indicate broadband is available when it is not (an overstatement). A Type II error occurs when the Form 477 data indicate broadband is not available, but it is (an understatement).

The following tables provide a feel for these errors. Some basic descriptive statistics are provided in Table 1. There are 291,086 census blocks in the Georgia Map data of which 99,512 have no locations (34.2%). Of the 191,574 location-populated blocks, 16.4% have zero broadband and 83.6% have at least some broadband. This latter figure is comparable to the Form 477 data which indicates a single location has broadband available in the census block. Of served blocks, 70.7% of the state total have 100% coverage and 12.9% have non-zero coverage less than ubiquitous coverage.

Table 1. Census Block Statistics

	GA Data		477 Data	
$t_i = 0\%, f_i = 0\%$	31,402	16.4%	37,567	19.6%
$t_i > 0\%, f_i = 100\%$	160,172	83.6%	154,007	80.4%
$t_i = 100\%$	135,376	70.7%		
$0 < t_i < 100$	24,796	12.9%		

For the Form 477 data (restricting to populated blocks), 19.6% have no availability (higher than the Georgia data) and 80.4% have at least some availability (lower than the Georgia data). The Georgia data indicates only 70.7% of blocks are completely served, which is lower than the 80.4% from the Form 477 data (which, by design, implies 100% coverage). Part of these discrepancies, at least for the unserved blocks, may be related to the timing difference in the data sets. Also, these statistics are for blocks, not locations.

Complaints about errors in the Form 477 data, at least for allocating subsidy dollars, appear justified.

Table 2 provides a consistency matrix of served and unserved locations based on the two data sources. Recall that s_i is the served and u_i the unserved locations from the Georgia data. For the five million census blocks marked as served by the Form 477 data ($f_i = 1$), the Georgia data indicate that 4.76 million locations are served (95% accurate) while 252,690 are unserved (5% inaccurate). For the 0.31 million blocks marked unserved by the Form 477 data ($f_i = 0$), 73.9% are accurately marked as unserved. In all, the Form 477 data has an accuracy rate of 90.9%, which, while high, has a Type I error rate affecting about 250,000 locations. Summing horizontally, there are about 4.84 million served locations and 0.480 million unserved locations in Georgia.

Table 2. Consistency Matrix

	$f_i = 0$	$f_i = 1$	Sum
s_i	81,042 (26.1%)	4,756,520 (95.0%)	4,837,562
u_i	229,684 (73.9%)	252,690 (5.0%)	482,374
Sum	310,726	5,009,210	

In Table 3, the accuracy matrix is presented for blocks comparing f_i to t_i at the extreme values of t_i (0 and 1). When the Form 477 data mark a block as entirely unserved, the Georgia also mark the block as entirely unserved in 70.5% of cases. Alternately, when the Form 477 data mark a block as entirely served, the Georgia data also mark the block as entirely served in 82.3% of cases. Nearly 5,000 blocks are improperly labeled as having broadband when no broadband is available, but 8,672 blocks are fully covered when the Form 477 data states there is no broadband at all. In all, the accuracy rate for entirely served or unserved is 83.3%.

Table 3. Consistency Matrix, Blocks
(Dichotomous Indicators)

	$f_i = 0$	$f_i = 1$
$t_i = 0$	26,468 (70.5%)	4,934 (3.2%)
$t_i = 1$	8,672 (23.1%)	126,704 (82.3%)

Table 4 summarizes the availability rates from the Georgia data by f_i. The (unweighted) average availability rate in Form 477 unserved blocks is 26.1% while in served blocks is 92.1%. Weighting by locations, the average availability rate in Form 477 unserved blocks is 26.1% while in served blocks is 95.0% (as in Table 2). The Form 477 data does distinguish between served and unserved blocks, albeit imperfectly. Availability rates of both 0% and 100% are found in blocks marked as served or unserved by the Form 477 data (also see Table 3).

Table 4. Availability Rates (t_i)

	$f_i = 0$	$f_i = 1$
Mean	26.1%	92.1%
Min	0%	0%
Max	100%	100%

The polychoric correlation coefficient between t_i and f_i is 0.67.[9] Marking as served any block where $t_i > 0$ (a dichotomous variable matching the 477 indicator), the correlation coefficient is 0.64. Again, the Form 477 is not "vastly inaccurate" or "fake news," but nor it is entirely accurate.[10]

While the Form 477 data and Georgia's more accurate measure of availability are highly correlated, they are not equivalent. Such discrepancies may be important for subsidy determinations.

While the Form 477 data and Georgia's more accurate measure of availability are highly correlated, they are not equivalent. Such discrepancies may be important for subsidy determinations. The location-based survey conducted by Georgia is certain to provide better guidance on subsidy allocations and other states should follow suit. Measuring the unserved using estimates of population rather than locations is certain to introduce inaccuracies. In all, the Form 477 data is not entirely useless, but strict adherence to its inferences is ill-advised.

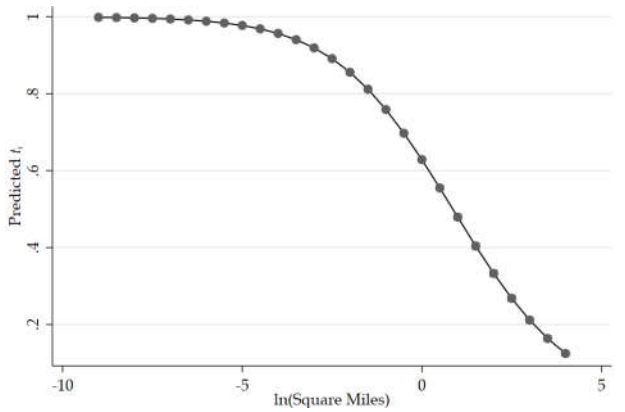
Errors and Block Size

Census blocks are often geographically small (e.g., a city block). In rural areas, however, census blocks can be quite large. Thus, because it is more reasonable to expect partial broadband coverage in large blocks than in small blocks, it might be expected that the overstatement inherent to the Form 477 data will be more pronounced in larger blocks. It is possible to assess this possibility using the regression,

$$t_i = \beta_0 + \beta_1 f_i + \beta_2 f_i \cdot m_i + \beta_3 f_i \cdot n_i + \varepsilon \ , \qquad (4)$$

where m is the natural log of block i's square mileage and n_i is total locations in block i [$= s_i + u_i$]. Since t_i lies on the unit interval, Equation (4) is estimated by a General Linear Model of the binomial family with a logit link (thus constraining predictions to the unit interval).[11] The β_2 coefficient is negative, indicating that the penetration rate is lower in larger census blocks marked served by the Form 477 data. The coefficients β_1 and β_3 are both positive.

Figure 1. Availability and Block Size



The relationship between t_i and block size m is illustrated in Figure 1. As block size increases, the predicted availability rate falls. The relationship is nonlinear, falling off sharply for the largest blocks. In the largest blocks, the predicted availability rates are often well below 50%. About 80% of blocks in Georgia have a $m > -5$ containing 93% of locations. About 7.5% of blocks have a $m > 0$ containing 11.8% of the locations. These larger blocks are common and contain large shares of locations.

It may be possible, using the Form 477 data, to approximate the overstatement due to block size using Equation (4) or something like it. Note that many western states have very large block sizes, many much larger than the largest block in

Georgia. Thus, using the Georgia data only leads to predictions outside the range. As suggested in Figure 1, it might do little harm to assume very large blocks have very low availability.

If the average subsidy to reach the net 9.1 million locations is $2,000 (the average from Auction 904), then the subsidy required to reach unserved households is $18.2 billion. If the average subsidy level is $3,000 (a fifty-percent increase over Auction 904), then $22.8 billion is needed. At a very high average subsidy of $5,000, getting broadband to every location requires funding of $45.5 billion.

Extrapolation to the Nation

An estimate of the number of unserved locations at the national level is conducted using Equation (4). The parameters are estimated using the Georgia data and then predictions are made using national data. Location counts in other states (a regressor) are approximated using the ratio of locations to population in Georgia (0.685). An estimate of locations served is the location weighted average of the predictions from Equation (4).

This approach predicts that 93.4% of locations across the U.S. are served, meaning 6.6% are unserved.[12] The total number of unserved locations is estimated to be 14.3 million. This figure is very close to that estimated by Chaplin, Kurn, Harlalka, and Burnett (2021) at New Street Research (14 million).[13] With the recent Auction 904 for the RDOF set to make broadband available to 5.2 million homes with $9.2 billion in subsidies, the *net* unserved locations equal 9.1 million locations. About $11 billion remains in the RDOF budget for future auctions.

How much subsidy is required to get broadband to the remaining locations? From Auction 904, the average subsidy was about $2,000. Presumably, reaching the last of the unserved locations may cost more. Table 5 summarizes the subsidies required at various average subsidies.

Table 5. Subsidies Required
(Billions)

Mean Subsidy	Subsidy
2,000	$18.2
2,500	$22.8
3,000	$27.3
3,500	$31.9
4,000	$36.4
4,500	$41.0
5,000	$45.5

If the average subsidy to reach the net 9.1 million locations is $2,000 (the average from Auction 904), then the subsidy required to reach unserved households is $18.2 billion. If the average subsidy level is $3,000 (a fifty-percent increase over Auction 904), then $22.8 billion is needed. At a very high average subsidy of $5,000, getting broadband to every location requires funding of $45.5 billion.[14] Note that about $11 billion remains in the RDOF budget, but if the Infrastructure Bill is passed (with $45 billion in funding), then it may be possible to reduce RDOF funding.[15]

Econometric Analysis of Broadband

The Form 477 serves as a measure of broadband availability in many empirical studies of the broadband marketplace and its influence on economic outcomes. Since t_i and f_i are correlated, the Form 477 data may serve as a useful proxy for broadband availability. Still, f_i contains measurement error and mismeasured regressors may lead to biased coefficients.

A Monte Carlo analysis may shed light on the magnitude of the problem and potential solutions. Let z be a simulated outcome determined by broadband availability and a

continuous variable X that is correlated with t_i, so that,

$$z_i = 1.0 + 0.10t_i - 0.05X_i + e_i \quad , \qquad (5)$$

where ε is a normally distributed random error term.[16] I let X be correlated with t_i since the bias of the estimated coefficients is related to correlations among the covariates. By using t_i, the z is the "true" effect, *if it is the availability rate rather than an indicator of any availability that determines the outcome*.

The effect of error in the measurement of the broadband availability from the Form 477 data can be assessed by estimating the regression,

$$z_i = \alpha_0 + \alpha_1 f_i - \alpha_2 X_i + v_i \quad , \qquad (6)$$

and comparing the α coefficients to their true values in Equation (5).

Table 6 summarizes the estimated coefficients for the true model and Equation (2) for various correlations of the t_i and X. As the correlation rises, the bias in β_2 should increase. The coefficients are the mean from 100 simulations.

Table 6. Simulation Results

$\rho(t_i, X_i)$	β_1 0.100	β_2 -0.050
0.00	0.066	-0.050
0.25	0.064	-0.035
0.50	0.057	-0.023
-0.50	0.057	-0.077

The results are as expected. When t_i and X_i are uncorrelated, only the β_1 coefficient is biased, and materially so (about a 35% difference). As the correlation between t_i and X_i rises, both β_1 and β_2 are biased. For a mild correlation of 0.25, the β_1 coefficient shrinks and the β_2 coefficient is -0.035, well below its true value of -0.050. With a moderate correlation of 0.50, the β_1 coefficient is 0.057 and β_2 is -0.023. These are material departures from the true values. When the correlation is negative (-0.50), the direction of the bias on β_2 is reversed, and the coefficient is more negative than its true value. It appears that using the Form 477 data leads to downward biased estimates of the effect of broadband on outcomes.

Since f_i is a mismeasured substitute for t_i, it may be possible to reduce the bias using Instrumental Variables regression ("IV").[17] Here, IV regression is not intended to address endogeneity but measurement error (though it may do both). Following Ford (2019), the excluded instruments are the natural log of the block's population, the f_j for block i's three closest neighbors j, and the interaction of these indicators with the distance between these blocks in miles based on centroid latitude and longitude.[18] These variables may be obtained from Census data or the Form 477 data.

Table 7 summarizes the results. Two models are estimated: (1) traditional IV regression ("IV"); and (2) IV regression weighted by the natural log of population ("IV-W"). With zero correlation, the β_1 coefficient is 0.091 and β_2 equals its true value for the standard IV method. The bias in β_1 is much smaller (about 10%, versus 35%). Weighting by population increases β_1 to 0.105, which is only 5% larger than its true value. As the correlation between t_i and X_i rises, the β_1 coefficient shrinks and β_2 becomes biased. For the most part, the weighted IV regression tends to produce better results, at least for the broadband variable.

Table 7. Simulation Results, IV

	$\rho(Y_T, X)$	β_1 0.100	β_2 -0.050
IV	0.00	0.091	-0.050
IV-W	0.00	0.105	-0.050
IV	0.25	0.089	-0.040
IV-W	0.25	0.103	-0.043
IV	0.50	0.082	-0.031
IV-W	0.50	0.099	-0.038
IV	-0.50	0.082	-0.069
IV-W	-0.50	0.099	-0.062

It appears that IV regression may be a suitable approach to address the errors in the Form 477 data when conducting empirical analysis on the effects of broadband. At least, it may produce a better estimate of the effect of broadband, though

any correlated variable may have a materially-biased coefficient. It is perhaps worth further study on a proper set of excluded instruments and to otherwise refine modeling choices.

Conclusion

Form 477 data take a lot of criticism for being an inaccurate measure of broadband availability. Georgia's recent mapping effort provides an opportunity to quantify the accuracy of the Form 477 data. At the aggregate level, the difference between the Form 477 data and actual availability is somewhat small—about 3.5 percentage points for Georgia. But, subsidy dollars are not allocated on aggregates but at the census block level. At the census block level, the Form 477 data is prone to mislead in some cases. State efforts to improve maps should lead to the better allocation of subsidy dollars. Complaints about errors in the Form 477 data, at least for allocating subsidy dollars, appear justified.

Instrumental Variables regression to account for measurement error appears to be a sensible approach, though further study is warranted.

The Georgia map sheds light on other important issues. As expected, the Form 477 data's zero-one approach to availability at the census block overstates broadband availability at high rates in larger census blocks. Also, assuming Georgia is representative of the nation, the number of unserved locations, net of the effect of the recent RDOF expenditures, is about 9.1 million locations. Assuming a very-high average subsidy level ($5,000 per location), the amount of subsidy dollars required to serve unserved locations is $45.5 billion, which is about equal to the funds allocated in the Infrastructure Bill (assuming no waste, which is unrealistic).

As for the empirical analysis of the effects of broadband on various outcomes, the errors in the Form 477 data appear to bias the effects of broadband downward. If any covariates are correlated with broadband availability, then the bias is worsened and extends to the coefficients on the correlated regressors. Instrumental Variables regression to account for measurement error appears to be a sensible approach, though further study is warranted.

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 or its staff. Dr. Ford may be contacted at*** ford@phoenix-center.org.

1 G.S. Ford, *Quantifying the Overstatement in Broadband Availability from the Form 477 Data: An Econometric Approach*, PHOENIX CENTER POLICY PERSPECTIVE No. 19-03 (July 11, 2019) (available at: https://www.phoenix-center.org/perspectives/Perspective19-03Final.pdf).

2 Ford, *id.*; *see also* J. Busby, J. Tanberk, and T. Cooper, *BroadbandNow Estimates Availability for all 50 States; Confirms that More than 42 Million Americans Do Not Have Access to Broadband*, Broadbandnow.com (August 21, 2021) (available at: https://broadbandnow.com/research/fcc-broadband-overreporting-by-state).

3 https://broadband.georgia.gov/maps.

4 Data available at: https://broadband.georgia.gov/maps/map-data.

5 Data available at: https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477.

6 *Fourteenth Broadband Deployment Report*, Federal Communications Commission (rel. January 19, 2021) at Fig. 4 and Appendix A (available at: https://docs.fcc.gov/public/attachments/FCC-21-18A1.pdf).

7 Ford, *supra* n. 1.

8 BroadbandNow, *supra* n. 2.

9 F. Drasgow, *Polychoric and Polyserial Correlations*, in ENCYCLOPEDIA OF STATISTICAL SCIENCES, Vol. 7. (S. Kotz, N. Balakrishnan, C.B. Read, B. Vidakovic & N.L. Johnson eds.) (1986) pp. 68–74.

10 B. Nuelle, *Senators Push FCC to Improve Broadband Data Maps*, AGRI-PULSE (March 20, 2019) (available at: https://www.agri-pulse.com/articles/12010-senators-push-fcc-to-improve-broadband-data-maps).

11 The estimated coefficients are: $\beta_0 = -1.055$; $\beta_1 = 0.402$; $\beta_2 = -0.681$; and $\beta_3 = 0.530$. All robust t-statistics are statistically significant at the 1% level or better.

12 Counting residential and business locations may be subject to error depending on the counting method. For instance, homes and buildings may be abandoned, some structures may be out buildings or barns, and so forth.

13 J. Chaplin, S. Kurn, V. Harlalka, and P. Burnett, *Biden's Choice: Infrastructure Investment or Lower Prices; Pick One*, New Street Research (May 19, 2021).

14 Chaplin, *et al.,* (2021), *id.,* also estimate that $35 billion will be required to cover the nation with broadband service.

15 R. Layton, *What's In The Broadband Component Of The Infrastructure Bill*, FORBES (September 2, 2021) (available at: https://www.forbes.com/sites/roslynlayton/2021/09/02/whats-in-the-broadband-component-of-the-infrastructure-bill/?sh=461a5cb22362).

16 The disturbance is a random number with mean 0 and standard deviation 0.10. The R^2 of the models is about 0.65.

17 A.C. Cameron and P.K. Trivedi, MICROECONOMETRICS (2005) at Ch. 26.

18 Ford, *supra* n. 1.