Quantifying the Overstatement in Broadband Availability from the Form 477 Data: An Econometric Approach

Dr. George S. Ford*

July 11, 2019

Like most things today, the nation's efforts at conducting communications policy receives considerable criticism, and often deservedly so.[1] Among other issues, a major point of contention for broadband policy is way the Federal Communications Commission ("FCC") collects and reports broadband availability data—a problem conceded by the FCC itself.[2] Broadband availability data is collected from broadband providers at the census block level, which is the smallest geographic unit used by the Census Bureau for data tabulation.[3] In collecting and reporting these data, it is assumed that if a single home in a census block has access to broadband (however defined), then every home in the census block has broadband. Plainly, this "*all-in assumption*," as I will call it, overstates broadband availability to some degree, and this overstatement is the source of much discontent and debate—so much so that some have even gone so far as to describe these data as "fake news."[4]

While the all-in assumption certainly overstates broadband availability to some extent, to date there has been no effort to quantify the magnitude of the overstatement. Census blocks are typically (though not always) fractions of a square mile with an average household count of only about 20 homes. We might expect, therefore, the overstatement to be small, but that would be a guess. While efforts are now underway to develop a new data collection strategy that offers greater precision, for now the magnitude of the overstatement is a mystery.[5]

While the all-in assumption certainly overstates broadband availability, to date there has been no effort to quantify the magnitude of the overstatement.

These new approaches to data collection will take some time to implement, but, as observed by FCC Commissioner Geoffrey Starks, the consequences of the all-in assumption "are worth exploring in more detail."[6] Such an exploration is the purpose of this PERSPECTIVE, in which I attempt to quantify, by statistical means, the magnitude of the overstatement of broadband availability resulting from the all-in assumption. Given the computational demands, I estimate the overstatement rate using the Form 477 data for six states (California, Georgia, Pennsylvania, Texas, West Virginia, and Wyoming) and then extrapolate to the nation.

With respect to the prediction model, the results are encouraging—model diagnostics are favorable and the results comport with expectations. Extrapolating to the nation, the overstatement rate is 3.3 percentage points, suggesting that about four million homes are said to have broadband that may not. Overstatement rates for counties in Wyoming are also calculated; the overstatement rates vary widely.

More important than the national coverage rate, which is useful mainly for speeches and reports, not knowing the specific geographic areas where broadband is unavailable impedes efforts to close the availability gap, including where subsidy dollars are spent.[7] As shown here, my statistical method may be useful in evaluating the accuracy of existing availability data prior to the implementation of new collection methods. I leave that task to future research.

An Empirical Strategy

Broadband availability statistics reported by the FCC are based on the Form 477 data.[8] These data are collected and reported at the census block level. A census block—created to manage the collection of data by Census Bureau personnel for the decennial census—is the smallest area defined by the U.S. Census Bureau for tabulating statistics. A block is typically a small geographic area with boundaries formed by streets, roads, railroads, lakes, rivers, and other physical and cultural features.[9] In some cases, a census block is literally a city block. In 2010, there were 11,078,300 census blocks in the U.S., although 4,871,270 were unpopulated. On average, there are about 50 people and 21 households in a block.

While census blocks are typically quite small, they are not uniformly sized and vary from fractions of a square mile in more urban areas to hundreds of square miles in more rural areas. While the all-in assumption undoubtedly overstates broadband availability, when blocks are very small and found in urban areas (a non-trivial share of which contain only a single home), the overstatement of the all-in assumption is likely small. In the larger blocks found in rural areas, however, the overstatement could render large discrepancies between reported and actual broadband availability. Any overstatement is, therefore, likely to be larger in rural than in urban areas.

When a census block is indicated as having broadband by the Form 477 data, all we know is that at least one household in the block has broadband. Underlying this zero/one indicator is an *unobserved* (or latent) broadband coverage rate, which is likely less than 100% in many blocks. My empirical analysis aims to exploit the differences in areas with and without broadband to recover this unobserved latent coverage rate for each block using only the Form 477 zero/one indicator of availability.[10] This approach is not novel; it is based on the standard latent variable interpretation of econometric models.[11] With an estimate of the latent coverage rate obtained from this model, it is possible to compute the overstatement rate of the all-in assumption at any chosen level of geographic aggregation from the census block to the nation.

…not knowing the specific geographic areas where broadband is unavailable impedes efforts to close the availability gap, including where subsidy dollars are spent.

More specifically, let r_i be an indicator variable that equals 1.0 if *any* home in a census block has access to broadband (however defined), 0 otherwise. The actual coverage rate that is used to determine whether a home has service is r_i^*, which is an *unobserved* latent variable. Under the all-in assumption, the broadband indicator, r_i takes on a value of 1.0 if $r_i^* \geq h_i$, where h_i is the coverage rate if only one home is served; r_i has a value of 0 otherwise. The overstatement rate for a *served* census block i is $1 - r_i^*$. For unserved blocks, the overstatement rate is presumed to be zero.[12]

In an effort to recover the latent coverage rate, my empirical model takes the general form,

$$r_i^* = \alpha C_i + \beta D_i + \lambda_0 r_j + \lambda_1 d_j + \lambda_2 r_j d_j + \varepsilon_i , \qquad (1)$$

where if $r_i^* \geq h_i$, $r_i = 1$, and if $r_i^* = 0$, $r_i = 0$, C_i is a measure of cost of deployment to the block, D_i is a set of demand-side factors that determine availability (e.g., income), r_j indicates broadband availability for block i's nearest neighbor j (a 0/1 indicator), d_j is the distance in miles to the nearest neighbor, and ε is the disturbance term. The spatial terms (r_j, d_j) account for the likelihood that the coverage in block i is higher if broadband is also available in a neighboring block j, where the influence is allowed to vary by distance to that neighbor.[13]

Using the predictions of the latent coverage rate based on the estimates of the model, the final estimate of the coverage rate is calculated using,

$$\dot{r}_i^* = \max(h_i \cdot r_i, \ \hat{r}_i^* \cdot r_i), \quad (2)$$

where r_i is the indicator of broadband availability (either 0 or 1), $\hat{r}_i^*$ is the prediction from Equation (1), and h_i is the coverage rate if a single home in block i has access to broadband (*i.e.*, the minimum coverage rate). Multiplying h_i by r_i forces blocks indicated as having no broadband to have a coverage rate of zero (which is not required by the model).[14] For blocks indicated as having broadband, the final estimated coverage rate is equal to $\hat{r}_i^*$.

With these estimates, the overstatement rate for any geographic areas (e.g., county, state, or nation) is obtained by multiplying the estimated coverage rate for each block by the number of households in the block and summing across all blocks. This coverage rate may then be compared to the all-in aggregate coverage rate to quantify the overstatement of the number of homes with access to broadband service arising from the all-in assumption.

Estimating the Model

Spatial analysis is computationally burdensome, so analyzing data for the entire nation is beyond the scope of this analysis. Instead, I limit my attention to six states: California, Georgia, Pennsylvania, Texas, West Virginia, and Wyoming. California and Texas are both large states in terms of population and geographic footprint with an array of both urban and rural census blocks. West Virginia and Wyoming are chosen because of their relatively large rural populations. I expect that the overstatement rate is larger for rural areas, and this may manifest at areas across and within states. Georgia and Pennsylvania are moderately sized states with fairly typical levels of rural population.

Broadband availability is obtained from the latest available Form 477 data, which measures availability in December 2017 (excluding satellite services).[15] Version 2 of the data, which corrects from some known errors, is used. I limit the data to consumer service alone.

Explanatory variables in vector C include an indicator variable for the cost of serving each block. The cost variable is categorical and is obtained from the analysis of costs for the Connect America Fund ("CAF").[16] The categories include costs less than \$30, between \$30 and \$40, between \$40 and \$50, between \$50 and \$200, and greater than \$200.[17] Missing values for the cost variable are imputed and a dummy variable for blocks with missing data included in the model.[18]

Demand-side factors (vector D) include four income variables: (1) the share of persons with incomes less than \$25,000; (2) the share of persons with incomes between \$25,000 and \$49,999; (3) the share of persons with incomes between \$50,000 and \$99,999. The share of persons with incomes of \$100,000 or more are excluded to avoid the dummy trap. A fourth variable is equal to the ratio of persons with incomes greater than \$100,000 to those with incomes less than \$25,000.[19] Income is measured at the census tract level.

For the spatial component of the model, I find the nearest neighbor for each census block and include as a regressor the zero/one indicator for broadband availability in that neighbor's block, the distance to that neighboring block, and the interaction of the two variables.[20] Variables for three broadband speeds are included: (1) maximum speeds less than 10 Mbps; (2) maximum broadband speeds of 25 Mbps or greater (matching the dependent variable); and (3) maximum broadband speeds of 100 Mbps or greater.

There are multiple models that may be used to estimate the coverage rate by Equation (1). Logit and Probit Models are obvious choices, since the dependent variable is dichotomous and they have a natural latent variable interpretation.[21] Here, the Logit Model is used, though the Probit Model provides nearly identical results.

Results

Summary statistics for the six states evaluated are provided in Table 1.[22] California and Texas are the nations two most populous states with relatively high urban populations. West Virginia and Wyoming have small populations and below average urban populations. Georgia and Pennsylvania lie in between these extremes. Wyoming has the largest average block size at 1.06 square miles; California has the smallest average block size at only 0.13 square miles. These six states represent a variety of sizes and urbanicity, which is intended.

Table 1. Summary Statistics for States

State	Pop (mil)	Homes (mil)	Urban	Block Sq.Mi.
CA	37.3	12.542	95.0%	0.130
GA	9.71	3.518	75.1%	0.262
PA	12.7	4.958	78.7%	0.132
TX	25.2	8.886	84.7%	0.312
WV	1.85	0.741	48.7%	0.316
WY	0.56	0.223	64.8%	1.063

Many census blocks are unpopulated and these blocks are excluded from the sample. A few observations in each state are lost due to missing data. Given the importance of the cost data and the relatively large number of missing values (though less than 10%), I impute costs for missing values.

... there is a near linear relationship between the overstatement rate and the share of rural population in the state, despite the fact that the Logit Model does not include a specific indicator for urban-or-rural status.

As is standard, I look at the availability data for a broadband service satisfying a 25-3 Mbps (up-down) service threshold. Equation (1) is estimated by the Logit Model and the predictions are computed, which allows for the calculation of the coverage rate in Equation (2). For present purposes, I am interested only in the predictions of the model. The sign and significance levels of the regressors are unimportant, though all the regressors are statistically different from zero. For all states, the spatial terms are potent determinants of availability and greatly improve the fit of the model.

The coverage rate of Equation (2) ranges from nearly zero (h_i) to one. For California, as an example, in areas with broadband the coverage rate has a minimum of 0.03 and a maximum of 1.0; other states have similarly wide ranges. In many instances, the coverage rates are below what the all-in assumption implies, and in many cases is well below 100%.

Table 2 summarizes some basis diagnostics from the Logit Models. Sample sizes are very large in all cases and the Pseudo-R^2 of the models are all relatively large. The model does a good job at predicting both blocks with and without broadband at the 25 Mbps level. The value of the Receiver Operator Curve ("ROC") are all very large, indicating the model provides an "outstanding" discrimination.[23] The Hosmer-Lemeshow Goodness-of-Fit test is favorable for all states.[24]

Table 2. Model Diagnostics

State	N	Pseudo R^2	Correct [1, 0]	ROC
CA	380,242	0.73	[0.988, 0.865]	0.977
GA	150,988	0.46	[0.946, 0.665]	0.923
PA	277,365	0.41	[0.971, 0.538]	0.908
TX	435,377	0.65	[0.955, 0.904]	0.962
WV	59,290	0.49	[0.955, 0.800]	0.927
WY	23,005	0.66	[0.936, 0.915]	0.963

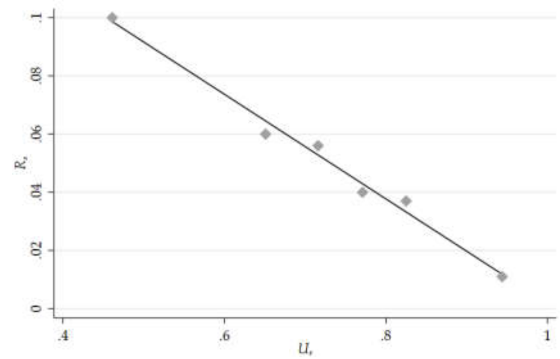
Turning the predictions of the model, Table 3 summarizes the results. Under the all-in assumption 98% of homes in California have access to broadband at the 25-3 Mbps level. The predicted coverage rate, however, is only 96.9%, so the overstatement rate is 1.1 percentage points. This overstatement seems rather small. For Georgia, however, with its slightly larger rural population and larger block sizes, the overstatement of availability is 5.6 percentage points, with the all-in assumption reporting a coverage rate of 94.8% whereas the predicted coverage rate is only 89.1%.

Table 3. Predicted Coverage Rates

State	All-In Coverage	Predicted Coverage	Over-Statement	R_s
CA	0.980	0.969	0.011	0.014
GA	0.948	0.891	0.056	0.050
PA	0.961	0.921	0.040	0.043
TX	0.938	0.901	0.037	0.033
WV	0.867	0.767	0.100	0.098
WY	0.854	0.794	0.060	0.069

For the more rural states, the predicted overstatement rates are much larger. For West Virginia, the overstatement is a sizable 10 percentage points, and for Wyoming is 6 percentage points. The error rate is largest for West Virginia, which is the state with the largest share of rural population. Likewise, the smallest error rate is for California, the state with the highest share of urban population.

Figure 1. Predicted Coverage Rates



In fact, as shown in Figure 1, there is a near linear relationship between the overstatement rate and the share of rural population in the state, despite the fact that the Logit Model does not include a specific indicator for urban-or-rural status. A reasonable approximation of the overstatement rate for a state may be approximated by the formula,

$$R_s = 0.186 - 0.181 \cdot U_s, \quad (3)$$

where R_s is the overstatement rate for state s and U_s (where $0 \le U_s \le 1$) is the share of urban population for state s. These coefficients are based on a least-squares regression (with $R^2 = 0.987$). The values for the six states are provided in Table 3; the approximation is imperfect but a reasonable back-of-the-envelope estimate. Be aware, however, that there are different measures of urbanicity and the coefficients of Equation (3) are based on a specific one. Thus, when making the calculation of Equation (3) it is essential to use the same source as used here.[25] Also, the formula is not accurate for any geographic aggregation other than a state.

Looking within states, Table 4 provides the overstatement rates for urban and rural blocks. There are large differences in the overstatement rates between urban and rural areas across all six states. In California, the overstatement rate is only 0.8 percentage points, but is 6.1 percentage points in rural areas. The largest spread is for Wyoming, with a 1.5 percentage point error rate in urban areas and a 14.3 percentage point error rate in rural areas. In West Virginia, the urban overstatement rate is relatively high at 4.5 percentage points.

Table 4. Overstatement by Urban Status (Percentage Points)

State	Statewide	Urban	Rural
CA	1.1	0.8	6.1
GA	5.6	2.8	14.0
PA	4.0	1.8	11.8
TX	3.7	2.3	10.6
WV	10.0	4.5	15.5
WY	6.0	1.5	14.3

While I report the data at the state level above, it is possible to calculate the overstatement rates for any level of geography. Table 5 provides the estimated overstatement rates for the 23 counties in Wyoming.

Table 5. Overstatement in Wyoming Counties

County	Over-Statement	County	Over-Statement
Albany	0.019	Natrona	0.030
Big Horn	0.042	Niobrara	0.028
Campbell	0.099	Park	0.074
Carbon	0.039	Platte	0.060
Converse	0.052	Sheridan	0.095
Crook	0.254	Sublette	0.170
Fremont	0.061	Sweetwater	0.087
Goshen	0.024	Teton	0.149
Hot Springs	0.033	Uinta	0.020
Johnson	0.090	Washakie	0.053
Laramie	0.069	Weston	0.217
Lincoln	0.206		
		State-Wide	0.060

While the overstatement rates for many counties are relatively small, many others are quite large. Crook, Lincoln, Sublette, Teton and Weston Counties have particularly large overstatement rates. These counties have below average household counts and typically consist of blocks with above average square mileage, below average population density, and above average costs. In Crook County, for instance, 41% of homes are said to have broadband at the 25 Mbps level where only about 15% are predicted to have coverage. Broadband availability in these counties may warrant special attention.

Extrapolating to the Nation

In order to obtain an estimate of the overstatement rate for the nation, I combine the data from the six states and estimate the model. I assume these states are a suitable proxy for the nation. The estimation sample includes 1,326,267 observations, the Pseudo-R^2 of the model is 0.57, and the ROC is 0.95.

...the all-in assumption overstates the number of homes with access by about four million homes, thus overstating availability by 3.45%. I suppose the question of whether or not this is a "large" discrepancy is in the eyes of the beholder, though it is less than 5%.

For these states, 95.7% of households have broadband service at the 25-3 Mbps level under the all-in assumption. The model predicts that only 92.4% have broadband for an overstatement rate of 3.3 percentage points. Across the six states, the overstatement rate for rural blocks is 11.5 percentage points but only 1.9 percentage points for urban blocks.

In 2017, there were 126 million households across the country. Under the all-in assumption, about 120.6 million could purchase 25-3 Mbps broadband service. Of these, my estimate approach suggests that only 116.4 actually have such access, so the all-in assumption overstates the number of homes with access by about four million homes, thus overstating availability by 3.45%. I suppose the question of whether or not this is a "large" discrepancy is in the eyes of the beholder, though it is well less than 5%. There are, however, far more important reasons to collect better data than tabulating statistics for the Commission's *Broadband Deployment Report*.

Conclusion

Dissatisfaction with broadband availability data has led to significant interest in developing new (and potentially costly) methods for data collection. At the center of the frustration with current methods is the inability of the data to determine how many homes within a census block have broadband. At present, if a single home in a census block has access to broadband, then it is assumed all homes in the block have the same access. Though unlikely to be true, there has been no effort yet to quantify the magnitude of the overstatement caused by this all-in assumption, and thus no basis for concluding the data offers a substantially inaccurate summary of broadband availability.

In this PERSPECTIVE, I estimate the coverage rate of broadband service, applying a Spatial Logit Model to the Form 477 data for the states of California, Georgia, Pennsylvania, Texas, West Virginia, and Wyoming. The all-in assumption, as expected, is found to overstate broadband availability, especially for more rural states and rural census blocks. At the national level, there are about 4 million homes said to have broadband that may not, which is less than a five-percentage point overstatement rate.

At the national level, the overstatement from the all-in assumption is, in my view, not terribly large, though there may be disagreement on that assessment. In any case, efforts to improve data collection have little to do with national reporting, however, but aim to better target broadband deployment efforts and subsidy funds to specific geographic areas. As demonstrated here, the statistical approach used here may be useful for predicting the overstatement rate for specific geographies, at least until the enhanced data collection efforts are fully implemented.

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 *See, e.g.*, G.S. Ford and M. Stern, *Ugly is Only Skin Deep: An Analysis of the DE Program in Auction 97*, PHOENIX CENTER POLICY PERSPECTIVE NO. 15-04 (July 20, 2015) (available at: http://www.phoenix-center.org/perspectives/Perspective15-04Final.pdf); G.S. Ford and L.J. Spiwak, *The Unpredictable FCC: Politicizing Communications Policy and its Threat to Broadband Investment*, PHOENIX CENTER POLICY PERSPECTIVE NO. 14-05 (October 14, 2014)(available at: http://www.phoenix-center.org/perspectives/Perspective14-05Final.pdf); G.S. Ford and L.J. Spiwak, *Justifying the Ends: Section 706 and the Regulation of Broadband*, 16 JOURNAL OF INTERNET LAW 1 (January 2013) (available at: http://www.phoenix-center.org/papers/JournalofInternetLawSection706.pdf).

2 M. Robuck, *FCC's Pai Proposes New Order To Gather Better Broadband Mapping Data*, FIERCETELECOM (June 14, 2019) (available at: https://www.fiercetelecom.com/regulatory/fcc-s-pai-proposes-new-order-to-gather-better-broadband-mapping-data).

3 *FCC Form 477: Local Telephone Competition and Broadband Reporting Instructions*, Federal Communications Commission (December 5, 2016) at p. 17 (available at: https://transition.fcc.gov/form477/477inst.pdf) ("Report a list [] of all census blocks in which the filer (including affiliates) makes broadband connections available to end-user premises …")

4 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); C.L. Rachfal, *Broadband Deployment: Status and Federal Broadband Programs*, CONGRESSIONAL RESEARCH SERVICE (June 6, 2019) (available at: https://crsreports.congress.gov/product/pdf/IF/IF10441); R. Molla, *The Government is Using the Wrong Data to Make Crucial Decisions About the Internet*, VOX (February 8, 2019) (available at: https://www.vox.com/2019/2/8/18211794/government-data-internet); T. Jett, *Is Bad FCC Data Holding Back Georgia's Rural Internet Push?*, GOVERNMENT TECHNOLOGY (March 2, 2018) (available at: https://www.govtech.com/network/Is-Bad-FCC-Data-Holding-Back-Georgia-s-Rural-Internet-Push.html).

5 J. Spalter, *Mapping the Broadband Gap*, RICHMOND TIMES DISPATCH (April 29, 2019) (available at: https://www.richmond.com/opinion/their-opinion/jonathan-spalter-column-mapping-the-broadband-gap/article_bfedd5fe-7745-5402-90a9-0fa209bafd04.html); C. Wood, *NTIA Partners with 8 States for Better Broadband Coverage Map*, STATESCOOP (February 13, 2019) (available at: https://statescoop.com/ntia-partners-with-8-states-for-better-broadband-coverage-map); J. Eggerton, *NCTA Pitches FCC on 3-Step Method for Improved Broadband Mapping*, MULTICHANNEL NEWS (February 22, 2019) (available at: https://www.multichannel.com/news/ncta-pitches-fcc-on-3-step-method-for-improved-broadband-mapping); J. Eggerton, *Broadband Mapping Bill Introduced*, MULTICHANNEL NEWS (May 9, 2019) (available at: https://www.multichannel.com/news/broadband-mapping-bill-introduced).

6 *In the Matter of Inquiry Concerning Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion*, FCC 19-44, 2019 BROADBAND DEPLOYMENT REPORT, __ FCC Rcd. __ (rel. May 29, 2019) (available at: https://docs.fcc.gov/public/attachments/FCC-19-44A1.pdf) (hereinafter "*2019 Broadband Deployment Report*"), Dissent of Commissioner Geoffrey Starks.

7 *See, e.g.*, A. Nasr, *Rural Americans Suffer the Costs of Faulty FCC Broadband Data*, PACIFIC STANDARD (October 1, 2018) (available at: https://psmag.com/social-justice/take-that-for-fcc-data).

8 Fixed Broadband Deployment Data from FCC Form 477 (available at: https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477).

9 https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf.

10 J.S. Long, REGRESSION MODELS FOR CATEGORICAL AND LIMITED DEPENDENT VARIABLES (1997) at Ch. 3.

11 *Id.*

12 Though it may be that some blocks marked unserved may be partially served, I ignore that possibility here.

13 The command -geonear- in Stata 15 is used to find the nearest neighbor. Using shape files, it is also possible to select blocks that border block *i*. I leave that alternative to future research.

NOTES CONTINUED:

[14] The adjustment by h_i is for logical consistency and has no effect on the reported results.

[15] Data are available at: https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477.

[16] Data are available at: https://www.costquest.com/blog/news-and-events/post/pc-block-analysis. I thank James Stegeman at CostQuest for pointing me to the data.

[17] Continuous cost data is not publicly available but using continuous cost data may (or may not) improve the predictions.

[18] Imputation is conducted by ordinary least squares regression of the cost variable on the natural log of households, square mileage of the block, the population density of the block, and a rural block dummy variable. The predictions of the model are rounded to the nearest integer to comport with the format of the cost variable.

[19] Data are available at: https://ciser.cornell.edu/data/data-archive/census-2010-sf1-download-center.

[20] The distance to the neighboring block is correlated with the size of block *i* and *j*.

[21] Long, *supra* n. 10. A Linear Probability Model is another option, but it can predict coverage rates outside the 0/1 interval so is not used. For a digestible discussion of the different approaches, *see, e.g.*, P. Van Hoppel, *Linear vs. Logistic Probability Models: Which is Better, and When?*, STATISTICAL HORIZONS (July 5, 2015) (available at: https://statisticalhorizons.com/linear-vs-logistic).

[22] Data are available at: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk; https://www.indexmundi.com/facts/united-states/quick-facts/all-states/households; https://www.icip.iastate.edu/tables/population/urban-pct-states; https://www.census.gov/geographies/reference-files/time-series/geo/tallies.html.

[23] *See*, *e.g*., D.W. Hosmer Jr. and S. Lemeshow, APPLIED LOGISTIC REGRESSION (2000) at pp. 160-4. The ROC has values ranging from 0.50 to 1.00. The standard interpretation is ROC values between 0.70 and 0.80 provide "acceptable" discrimination, values between 0.80 and 0.90 provide "excellent" discrimination, and values in excess of 0.90 provide "outstanding" discrimination.

[24] *Id*. at pp. 147-8.

[25] The data are available at: https://www.icip.iastate.edu/tables/population/urban-pct-states.