The Empirics of the Digital Divide: Can Duration Analysis Help?

Wei-Min Hu and James E. Prieger*

21 October 2008

Abstract:

Accurate measurement of digital divides is important for policy purposes. Empirical studies on broadband subscription gaps have largely used cross-sectional data, which cannot speak to the timing of technological adoption. Yet, the dynamics of a digital divide are important and deserve study. With the goal of improving our understanding of appropriate techniques for analyzing digital divides, we review competing econometric methodology and propose the use of duration analysis. We compare the performance of alternative estimation methods using a large dataset on DSL subscription in the U.S., paying particular attention to whether women, blacks, and Hispanics catch up to others in the broadband adoption race. We conclude that duration analysis best captures the dynamics of the broadband gaps and can be a useful addition to the analytic tool box of digital divide researchers. Our results support the official collection of broadband statistics in panel form, where the same households are followed over time.

A revised version of this paper will appear in *Overcoming Digital Divides: Constructing an Equitable and Competitive Information Society*, edited by E. Ferro *et al.*, Hershey, Pa: IGI Global. After publication, please cite the book chapter instead of this working paper.

*Hu: Assistant Professor, Peking University Shenzhen Graduate School of Business. Prieger: Associate Professor, School of Public Policy, Pepperdine University. This chapter was written while the second author was visiting the Federal Communications Commission. The views expressed in this chapter are those of the authors and do not necessarily reflect the views of the FCC or any of its Commissioners or other staff.

INTRODUCTION

Digital divides are among the most pressing concerns in telecommunications policy. If attention paid to the divides is to generate light, and not just heat, then policy-makers require accurate measurement of the gaps in question. In this chapter, we assess some of the statistical tools that empirical researchers use to measure digital divides. Our focus is on econometric regression studies using data from many individuals, households, or geographic areas.[1] Many empirical studies of the digital divide analyze a cross-section of data[2] on the extent of digital deployment or use. Studies of broadband Internet access are a leading example (refer to the next section for citations). In these studies, researchers regress broadband subscription on characteristics of the household or the area, depending on the nature of the available data. Methods used for the binary access decision range from OLS regression, probit, and logit to more complicated estimators tailored to unique features of the data at hand (Flamm & Chaudhuri, 2007; Prieger & Hu, 2008). Researchers and policy-makers often use the results to identify subpopulations that are prone to lie on the wrong side of the digital divide.

What is missing from most of these approaches is the ability to say much about the *timing* of technological adoption. For example, take one of the results from Prieger & Hu (2008): blacks in the U.S. subscribe to broadband DSL at a lower rate than do whites. Unanswered are the questions of whether this divide is only temporary, as predicted by the catch up hypothesis, and how rapidly the gap will close if so. These questions are close to the heart of public policy toward digital divides. If gaps exist but close quickly

without intervention, policymakers may better direct public resources elsewhere. Persistent gaps, on the other hand, may warrant further study and action.

We aim to improve our understanding of appropriate techniques used to analyze the digital divide and policies aimed at reducing it. We use data on DSL adoption in the U.S. to compare the policy implications deriving from traditional cross-sectional analysis with that from duration analysis, an appropriate but under-used statistical technique in digital divide research. Our work contributes to the policy literature on the digital divide in three ways. We begin by clarifying the potential limitations of cross-sectional analysis. We also propose and explore the performance of duration analysis applied to broadband take-up data. Often data are available (or could easily be gathered) on how long a household has subscribed to broadband, even in cross-sectional data sets. Appropriately conducted duration analysis can then clarify the temporal dimension of the digital divide. Finally, we compare duration analysis to other methods used to examine the temporal dimension of the gap. Previous studies such as Whitacre (2008) and Flamm & Chaudhuri (2007) have analyzed data collected from different time periods. We explore whether duration analysis yields different conclusions than does panel data[3] analysis and whether results are more easily interpretable for policy makers.

In our empirical section, we examine the demand for DSL broadband in five U.S. states. To compare traditional cross-sectional analysis with duration analysis and panel data methods, we focus attention on groups prone to the digital divide: racial minorities and women. We assess the gaps three ways. Ordinary least squares and probit regression using the cross-sectional data, which is closest to what is done in most studies, establishes a baseline for our results. Next, we use duration (also known as survival) analysis to

speak to the pace at which gaps can be expected to close. Finally, we can use the cross-sectional data, coupled with the information on when households subscribed, to create a synthetic panel data set on subscription stretching back from the date of the cross-section to when DSL was first deployed in the neighborhood. These are the data that would have been available had subscription been surveyed periodically to create a panel dataset, for example. The latter two methods address the temporal dimension of the broadband gap for these groups. Although there is no policy variable in the models, the techniques we use also apply to policy analysis. We conclude that duration analysis best captures the dynamics of the broadband gaps and can be a useful addition to the analytic tool box of digital divide researchers.

We describe the statistical models without assuming that the reader is familiar with advanced econometric techniques. The chapter thus serves as both a reference for practitioners and as a blueprint for future research.

BACKGROUND

Literature

Over the past several years, broadband adoption has been widely studied. We look into main examples of the previous research in this section, emphasizing the methodology and nature of the data used in their estimations. Most authors use cross-sectional data, although a few take advantage of repeated cross-sectional data. We pass over the earlier generation of studies looking at pre-broadband digital divides (e.g., Fairlie, 2004), as well as studies not using individuals or households for the unit of analysis.

Among cross-sectional studies, the logit model for binary household choices is the most common methodology employed (Duffy-Deno, 2000; Kridel, Rappoport, and Taylor, 2001; Rappoport *et al.*, 2003; Stanton, 2004). Crandall, Sidak, & Singer (2002) use nested logit, an extension of the logit model, to estimate broadband demand. None of these examine the impact of gender or race on demand. The probit model (Leigh, 2003; Savage & Waldman, 2005) is less commonly employed. Leigh (2003) included variables for race, but failed to find significant differences in adoption (but also could not control for broadband availability). Prieger & Hu (2008) use a probit model adapted to aggregations of household data to find that women, blacks, and Hispanics have lower demand for DSL.

A smaller set of studies uses more than one cross section, collected at different times. Chaudhuri, Flamm & Horrigan (2005) analyze data from the Pew Internet and American Life Project with a logit model and another of its extensions, the ordered logit model, and find that women and blacks are less likely to subscribe to broadband. The ordered logit model is used to look at the related, hierarchical choice of no Internet access vs. dial-up access vs. broadband access. Flamm & Chaudhuri (2007) come to the same conclusion with later data from the same source examined with another ordered logit model. Finally, Whitacre (2007) uses the logit model to uncover shifts in the influence of household characteristics and telecommunications infrastructure on residential broadband adoption decisions. We did not find studies of broadband adoption that use panel data methods or that employ duration analysis. We show below that duration analysis is a useful tool to address the evolution of digital divides.

The catch up hypothesis

The temporal dimension of a digital divide is paramount when mapping statistics to the realm of policy. Persistent digital divides are cause for concern, while evanescent gaps are of little consequence. Central to our analysis is the notion of the adoption curve: the fraction of potential broadband adopters who have already adopted, plotted over time (Figure 1). The theory of technological diffusion (see Whitacre (2007) for a summary specific to broadband) explains the commonly observed ogive (S-shaped) adoption curve by learning. When few have adopted a new technology, few other will learn about it (or be convinced they should adopt) and the adoption curve rises slowly. Howell & Oren (2002) also highlight the roles of informational barriers and learning effects in DSL adoption. As time passes and more are exposed to the innovation, the adoption curve increases more rapidly. Eventually, the pace of adoption slackens when most have adopted and the remaining holdouts adopt slowly.

Textbook diffusion theory case thus posits the catch-up hypothesis. Even if a group has a lower adoption rate than the rest of the population, as depicted by the heavy adoption curve in Figure 1, both curves converge at full adoption eventually. While full convergence undoubtedly exists only in the realm of mathematical modeling, the catch-up hypothesis usefully highlights three points. First, adoption gaps today may disappear tomorrow. Second, even when ultimately converging, the adoption rates may diverge among groups initially. In such cases, cross-sectional analysis will uncover these gaps. Temporary gaps are not necessarily without policy concern. Given evidence that the way adopters use the Internet changes as they gain experience online (e.g., Weiser, 2000), differences in the timing of adoption may lead to differences in Internet usage among

groups even after most have adopted. Third, a key question is not merely *if* divides will disappear but *when*.

COMPARISON OF METHODS

Theory

We compare several econometric methods in our exploration. Econometrically experienced readers can skip to the next section to see our empirical results. For those wishing a brief review, this section sets out the basics of linear and probit cross-sectional regression models for binary dependent variables, duration analysis, and panel data methods.

Cross-sectional methods

Least satisfactory for purposes of investigating the catch-up hypothesis are cross-sectional methods, which attempt to uncover the determinants of adoption with data reflecting only one point in time. In terms of Figure 1, cross-sectional data is taken from households (or other units of observation) all at one point on the time axis. Cross-sectional studies can uncover disparities in adoption among subsets of the population, but generally cannot address how quickly the gaps developed or might close.

The cross-sectional studies reviewed in the literature section above model the mean adoption rate for a household as a function of (a linear combination of) explanatory variables (the *regressors*):

$$\mathrm{E}(y_i|x_i) = f(x_i'\beta) \tag{1}$$

where y_i is a binary variable taking values 1 if household i has adopted, and 0 otherwise, x_i is a vector of regressors, and β is a vector of coefficients to be estimated. In the case of ordinary least squares (OLS) regression, termed the *linear probability model* when

applied to binary dependent variables, the function f is the identity function. We assume readers are familiar with OLS methods.

In probit (or logit) models, f is the cumulative density function of the Normal (or logistic) distribution. Probit models have an advantage over the linear probability model, which is not commonly used in the broadband adoption literature. Unlike the linear probability model, the predicted probability of adoption from the probit model is bounded between zero and one, as a probability should be. While the probit model is no more difficult to implement with modern statistical software than is OLS, the interpretation of the coefficients is less obvious. In OLS, β_j (the coefficient for the jth regressor) is also the *marginal effect*, the effect on $E(y_i|x_i)$ of a unit increase in regressor j. In the probit model, the coefficient gives the sign of the marginal effect but not its level, which is $\beta_j\phi(x_i'\beta)$, the derivative of the conditional mean (1) with respect to regressor j.[4] Since the marginal effects depend on the data (i.e., x_i appears in the expression), they are typically computed at either the mean of the regressors or by averaging the marginal effect for each observation over the sample. We do the former below.

Cross-sectional methods, lacking a temporal dimension, cannot speak to the catch-up hypothesis and are of important but limited use for policy purposes. Nevertheless, these methods can document and partly explain the determinants of digital divides at any point in time, which is not without value. Furthermore, many times, only cross-sectional data are available, particularly when new technology is first available.

Duration analysis

When the purpose of the analysis is to estimate adoption curves, a natural method to use is duration analysis. Given its long association with biostatistics, duration analysis

is also commonly known as survival analysis. There are many excellent textbook treatments of duration analysis (e.g., Kalbfleisch & Prentice, 2002), and here we present only the basics needed to understand our estimations. In our context, the duration of interest is the time from the availability of broadband until a household adopts the technology. The fundamental notion in duration analysis is the *hazard rate*, $h(t)$, the rate at which adoption occurs given that it did not occur before time t. In exponential duration models, the hazard rate for household i is modeled as a function of explanatory variables (often called *covariates* instead of regressors in duration analysis):

$$h(t_i) = \exp(x_i'\beta) \tag{2}$$

Exponentiating $x_i'\beta$ ensures that the hazard rate is non-negative. If no functions of time are included among the covariates, the hazard rate in the exponential model is constant. The inverse of the hazard rate is the mean duration for the exponential model.[5] The interpretation of coefficients in specification (2) is thus as follows: a positive β_j implies that increases in the associated covariate increase the hazard and decrease the expected time until adoption. Coefficients can also be interpreted as in a log-linear regression model: a one unit increase in x_j increases the hazard by approximately $\beta_j \times 100$ percentage points. The exponential model is the simplest of the *proportional hazard* models, so-named because covariates have a proportional effect on the hazard rate.

The exponential model in its simple form, with its constant hazard rate, is not flexible enough to investigate the catch-up hypothesis. However, by splitting each duration into month-long intervals and adding dummy variables for the month, the baseline hazard rate can be modeled nonparametrically. Let $D_m(t)$ be a dummy variable for month m with coefficient α_m. More precisely, D_m is a step function that is zero

outside month m (timed from the start of the duration, not the calendar) and one within. Collect these into vectors $D(t)$ and α. Then our semiparametric[6] exponential model has hazard rate for t_i in month $m = 1,\ldots,M$ of

$$h(t_i) = \exp(D(t)'\alpha)\exp(x_i'\beta) = h_0(t)\exp(x_i'\beta) \qquad (3)$$

The baseline hazard h_0 is piecewise constant and can take any shape, nonparametrically accounting for the basic duration properties of the data. We constrain h_0 to be constant within a month only because DSL adoption in our data is observed at the monthly level and any additional α's that further partition time would be unidentified.[7] Since the α's vary during the time until adoption for any duration lasting longer than one month, we now have *time-varying covariates* (TVCs). Explicit treatment of TVCs complicates notation, and we ignore the issue here (except when presenting the formula for the adoption curve in equation (4) below). For the practitioner, the pressing question is how to set up the data for estimation when there are TVCs, and the answer depends on which software package is used.[8]

While the addition of h_0 makes the baseline hazard flexible, specification (3) (as well as other common semiparametric hazard models such as the Cox model) still imposes proportionality on the impact of the covariates. If a coefficient for Hispanics is –0.1, for example, then their hazard rate is constrained to always be (about) 10% lower than non-Hispanics in all months. To relax proportionality, we interact the covariates of interest (in our case, the variables *Female*, *Hispanic*, and *Black*) with the monthly constants. With a new set of covariates $D_1(t)x,\ldots, D_M(t)x$ (where x stands for the female, Hispanic, and black variables), the impact of these variables on the baseline hazard can vary freely among months. While greatly increasing the number of coefficients to be

estimated, the added flexibility is essential to investigate the catch-up hypothesis. Our enormous number of observations makes estimating the additional coefficients no problem. In smaller datasets the degrees of freedom may be used up rapidly, since interacting a variable adds *M-1* coefficients to be estimated.

With an estimate of the (time-varying) hazard rate, calculation of the adoption curve is straightforward. The adoption curve is formally the cumulative density function of the durations given the observed covariates. Standard results from survival analysis (Kalbfleisch & Prentice, 2002) show that for our model, the adoption curve F is found from the hazard rate as

$$\begin{aligned} F(t \mid x_i) &= 1 - \exp\left(- \int_0^t h(s) ds \right) \\ &= 1 - \exp\left(- \exp(z_i' \beta) \sum_{m=1}^{M} \exp(\alpha_m + \gamma_m w_{im}) \Delta t_m \right) \end{aligned} \tag{4}$$

where x_i is partitioned into TVCs w_{im} and other covariates z_i, M is the number of months spanned by t, and Δt_m is the amount of time spent in month m. With estimates of α, β, and γ in hand, predicted adoption curves can be generated for any subgroup of the population by setting the covariates to the appropriate values.

Repeated cross section and panel data

A few papers in the literature (e.g., Flamm & Chaudhuri, 2007; Whitacre, 2008) use repeated cross sections to address the digital divide. Repeated cross sections are cross sectional data gathered at multiple times, where the individuals or households differ each time. Repeated observations on the same units of analysis, known as panel data, enable more sophisticated modeling than do single or repeated cross-sections; see Hsiao (2002) for an excellent treatment of methods suitable for panel data. We are not aware of previous panel studies of broadband demand using individual- or household-level data.

The greatest advantage of panel data is the ability to control for unobserved factors specific to the unit of observation (e.g., households) that may render cross-sectional estimation results invalid through the use of random or fixed effects. Perhaps more important for present purposes, panel data can shed light on the dynamics of a divide since households are followed over time. Panel methods are available for linear, probit, and logit models. The interested reader is referred to Hsaio (2002) for descriptions of these and other panel models.

An empirical application of the methods

The data we analyze is from 1998-2000, the early years of DSL adoption. The vintage of the data limit the applicability of our results to present digital divides. However, the dataset has other advantages that make it suitable to demonstrate the candidate methods.[9] The data cover households in over 50,000 Census blocks in four Midwestern U.S. states. For each Census block, the dependent variable is whether at least one household subscribes to the incumbent phone company's DSL service. Only blocks where DSL is available are in the data. Since it is unlikely that DSL from any other provider would have been offered without the incumbent's service available, the data give a good measure of DSL adoption. Cable modem subscription and other forms of broadband Internet access are not covered in the data, however. While the data are not at the household level, the geographic fineness of the data[10] and the large number of observations make these data unique. Prieger & Hu (2008) describe the construction of the dataset more fully, and analyze it using a cross-sectional method.

Joined to the dataset are Census variables measuring the number of households in the block, the racial and ethnic composition, the fraction of women, and income. The

latter is aggregated in the Census data to the block group level, and so in all estimates we cluster the observations at that level when calculating the standard errors of the estimates.[11] Each Census block is also matched to the phone company's local service area into which it falls.

For the cross-sectional analysis, we use the snapshot of DSL adoption as of March 2000 provided in the data, at which time 85% of blocks had a household subscribing to DSL. For the duration analysis, we create observations on the time until initial adoption by a household in the block.[12] Time elapsed is measured from to the initial availability of DSL in the block, and so durations for blocks in different local service areas are not necessarily occurring at the same calendar time. Blocks that never subscribed are durations for which the ending time is not known, and are marked as right-censored observations.[13] For the panel analysis, we create monthly panel data from the March 2000 data and the information on when first adoption occurs in each block. A block that could have subscribed to DSL a year before any household actually did, for example, will have zeroes for the adoption variable for 12 months before it changes to one upon adoption and thereafter. The data are equivalent to a monthly adoption survey of areas where DSL is enabled.

Results from cross-sectional models

To establish a baseline for DSL adoption, in this section we present the linear probability model (OLS) and probit regression results from the cross sectional data from March 2000. Results are presented in Estimations 1 (OLS) and 2 (probit) in Table 1. Our main independent variables of interest in our estimations are the fraction within each Census block that are female, identify with a racial minority, or claim Hispanic ethnicity.

In estimation 1, aside from race, ethnicity, and gender we control for the log of income (in levels and squares to test for non-linearities), average household size, the number of households within the census block, and a set of indicator variables for the local telephone service (central office) areas. The role of the central office indicators is to hold constant all unobserved factors common to all households in the area. Such factors include how long DSL has been available in the central office, the availability of competitors also offering DSL in the area, and the average value of all other unobserved factors that vary among households.

The coefficients from OLS, reported in Table 1, are similar to the marginal effects from the probit estimation, and we discuss the latter. In comparison to whites, the excluded category, only Asians and other races have significantly lower probability (13.1% and 6.6% respectively) of DSL adoption. That adoption is lower for Asians is the opposite of national statistics (Prieger & Hu, 2008). We have few Asians and "other races" in our Midwestern sample (3.7% and 7.7% of people, resp.), and our results may not be representative. The negative coefficients for women and blacks reveal adoption gaps, but are insignificant. Surprisingly, the coefficient for Hispanics is positive (but not significant). Blocks with more and larger households are more likely to contain a household adopting DSL, as expected. Income has no significant effect, probably because the central office fixed effects remove the variation in average income among local service areas.

The cross-sectional estimations yield a few results of note. First, as one commonly finds with binary dependent variable models, it matters little whether one uses probit or the linear probability model.[14] More interesting is that the analysis does not

uncover digital divides where other studies lead us to look for them, except for the "other race" category. It may be that broadband diffusion was fairly even among the population in the states represented in the data. However, our analysis in the next section leads us to conclude instead that the cross-sectional analysis fails to find broadband gaps that do exist for women, blacks, and Hispanics. Finally, there is no way to speak to the catch-up hypothesis with these results, because there is no temporal dimension in the data.

Results from duration analysis

We now consider whether duration analysis sheds additional light on the adoption experience of women and minorities. Two specifications of the duration model are compared in Table 2: one in which the variables for blacks, Hispanics, and females are constrained to affect the hazard rate proportionally (Estimation 3), and another in which they are not (Est. 4). The coefficients on the monthly dummy variables are largest in month one (showing that many households adopt DSL immediately upon availability) and overall create a rough U-shaped hazard rate.[15] In both estimations, the hypothesis that the coefficients on the monthly dummy variables are equal to each other is rejected. Thus the baseline hazard rate of adopting DSL is not constant (or even monotonic) in these data, which makes our semiparametric approach a appropriate choice.

Estimation 3 shows that women, blacks, Hispanics, and Asians have significantly lower hazard rates for DSL adoption. Thus, in contrast to the suggestions of the cross-sectional results, these groups take longer on average to adopt after DSL becomes available to them. The coefficient for other races is not significant. The coefficients for log income imply that as income increases the time to adoption decreases (for all but the bottom 0.7% of incomes). Larger households also decrease the time to adoption.

Estimation 3 constrains the female, black, and Hispanic variables to affect the hazard proportionally regardless of elapsed time. We relax this assumption to investigate the catch-up hypothesis in Estimation 4, in which we allow the impact of these three variables to differ in each month elapsed after DSL is available. A hypothesis test for the three variables that the coefficients in the expanded set are equal in all months, which tests the assumption of proportionality, is soundly rejected for each variable. The impact and significance of the other variables is similar to that in Estimation 3.

Catch-up is most easily investigated via the adoption curves implied by the coefficients. The adoption rates for women, blacks, Hispanics, and others are graphed in Figure 2. We limit the graphs to the first nine months after DSL becomes available because no further adoption is observed until month 22. The adoption curves are thus flat until month 22, and then the coefficients are either insignificant (*black*M22*) or large and negative (*Female*M22 and Hispanic*M22*), so that the adoption curves remain nearly flat.[16] Two curves are calculated in each graph to compare the group of interest with everyone else.[17] In the top panel of the figure, the adoption rate for women starts at 2.1% after the first month of availability and rises to 3.3% after nine months. Men start out with a 10.3% adoption rate, and the absolute difference between men and women stays relatively constant across the graph. Thus, there is no evidence of women catching up to men during the first year of availability.

The adoption curve for blacks in the middle panel shows a different story. The adoption rate for blacks starts at 3.7% after the first month of availability, compared to 4.8% for non-blacks. The adoption rate for blacks rises to 4.3% after nine months, but the gap between blacks and others doubles over time, from 1.1 percentage points after

one month to 2.2 points after nine months. Not only do blacks fail to catch up during this period of initial DSL availability, they fall further behind.

Hispanics fare differently than women and blacks concerning adoption. Hispanics have an adoption rate of 4.1% initially, compared to 10.7% for non-Hispanics. Hispanic household adoption rises to 6.2% after nine months. After a slow start, their gains in adoption are greater than that for non-Hispanics, and their adoption gap narrows by 18% (from 6.6 to 5.4 percentage points) during the time. Hispanics do begin to catch up even during our relatively brief period.

Results from panel data models

In this section we repeat the OLS and probit regressions using panel data to compare the duration analysis with another way of studying digital divide dynamics. In addition to the set of regressors we use in the cross-sectional estimations, we include dummies for each calendar month in the estimation. We again include indicators for the local service areas to control for unobserved, time-invariant factors specific to the central office area.[18] Since, as before, the coefficients from OLS estimation for DSL adoption are similar to the marginal effects from the probit estimation we present and discuss only the latter. Two specifications are compared in Table 3. In Estimation 5, the adoption gap between women and men is constrained to be constant over time, and same for the gaps between the minority groups and their non-minority counterparts. In Estimation 6, the adoption gaps for women, blacks, and Hispanics are allowed to vary as time progresses.

In both estimations, the coefficients for the monthly indicator variables are positive, significant, and generally increasing over time. The month coefficients by themselves represents the baseline adoption trend for all groups in Estimation 5 and for

non-black, non-Hispanic males in Estimation 6. The data thus show that over the period December 1998-March 2000, demand for DSL services progressed in the region. Recall that since only Census blocks where DSL is available at the household are included in each month's data, the results do not merely pick up that DSL becomes more widely available.

In Estimation 5, the signs and significance of the coefficients for income, number of households, and household size are the same as in the corresponding cross-sectional estimation (Estimation 2), and we focus on the variables of interest instead. The marginal effects show that Asians and those in the "other race" category have significantly lower (the latter only at the 10% level) adoption rates than the whites. Women also have lower adoption rates (10% significance level) than men. The gaps are sizeable: 17.2% for Asians, 6.6% for women, and 11.0% for other races. There are small, insignificant adoption gaps between blacks and whites and between Hispanics and non-Hispanics.

Compared to the results from the cross-sectional estimation, the estimated adoption gaps are larger for all groups of interest, except for blacks. The larger number of observations also leads to statistical significance (albeit only at the 10% level) for the gaps for women and other races. In addition, using the panel data removes the anomalous positive coefficient for the Hispanic group. Since these results are more in line with results found in the literature, the case is strong for using panel data over a cross-section, even before moving to the augmented set of variables in Estimation 6. We compare the panel data results to the results of the duration analysis below.

In Estimation 6, we interact the month indicators with the variables for women, blacks, and Hispanics to evaluate how their adoption gaps evolve. The marginal effect

for, e.g., blacks in month 5 is the difference in the level of broadband adoption between blacks and whites in the same month. If the marginal effect is negative, there is a broadband gap that month for the group in question. The gaps in DSL adoption are depicted in Figure 3.[19]

The estimation indicates that, as with the duration estimations, there are significant differences in the evolution of the broadband gaps for women, blacks, and Hispanics. In month one, only the gap for women is significant. However, as time passes the gaps reverse. For women, the gap narrows significantly after eight months, with a few months of reversal mixed in, until the final month, which shows women strongly ahead of men. Hispanics start with essentially no gap, show stronger adoption than non-Hispanics through the first year, and then begin to lag sharply in the last few months. The pattern for blacks is similar to that of Hispanics, except that they have only one month of significantly more adoption than whites. The pattern of catch up overall, therefore, is present for the women but absent for blacks and Hispanics. The impacts of the other variables are generally similar to those in Estimation 5.

To compare with the results of the duration analysis, consider the message Figure 1 suggests if it is truncated at nine months. One would conclude that female broadband adoption not only catches up to the baseline, but surpasses it. Blacks apparently start out ahead of others but slowly lose their advantage and maybe fall behind. Hispanics also start out ahead and increase their broadband lead over others during the next eight months. These conclusions differ starkly with the patterns revealed by the duration analysis. The previous section showed that women exhibit no evidence of catching up to men, that adoption by blacks was never ahead of others, and that Hispanics narrow their

broadband gap but do not erase it. Furthermore, the panel results do not seem plausible in their own right, given other estimates of broadband demand (Prieger & Hu, 2008; Flamm & Chaudhuri, 2007; Stanton, 2004).

Why do the panel results mischaracterize the dynamics of the broadband adoption gaps? The comparison to the results from the duration model is not exact, since the estimates in the previous section are at the household level and those here are at the level of the Census block. However, aggregation alone should not create such widely differing results. To test this, we aggregated the data to the block group level, and re-ran Estimation 6. Although the levels of the broadband gaps for women, blacks, and Hispanics differed somewhat from Figure 3, the general shape of the curves was the same.

A more likely reason that the panel data—and also the cross-sectional data—do not properly capture the dynamics of the adoption gaps is that DSL becomes available at different times in different areas. Time in the duration model is time elapsed since availability, whereas in the panel data it is calendar time. The panel estimations thus suffer from composition effects, since in any calendar month there are new areas added to the sample as DSL becomes available. Furthermore, in any cross-section of the panel, some areas will have had access to DSL for months, while it will be newly introduced in other areas. Of course, the panel data can be re-organized to have the same timing convention as do the duration models. However, probit estimation of monthly observations on time to adoption is merely a duration model itself. However, discrete duration models estimated by probit are neither as easy to interpret nor as naturally linked to the underlying duration process as is our duration model.[20]

CONCLUSION

Duration analysis can be a useful analytic implement in the tool box of digital divide researchers. Cross-sectional studies may highlight the existence of divides at a point in time (although they did not here), and indeed may be all that is possible in the initial stages of monitoring adoption of a new technology. However, with our DSL adoption data duration analysis gives a more complete picture. In particular, duration analysis sheds light on how groups progress along their adoption curves. Policymakers can use the information to identify groups for which the adoption gap is widening rather than closing. While some of the inner workings of duration analysis may appear arcane to policymakers without substantial econometric foundations, the results can be presented in adoption curves, which are easy for anyone who can read a graph to interpret and understand.

Although we have concerned ourselves in this chapter primarily with methodological issues, our work suggests one policy recommendation. For duration analysis to be performed, longitudinal data must be available on households. To the extent that duration analysis proves useful for analyzing digital divides, it follows that priority in data collection should go to following the same people or households over time, rather than merely surveying differing cross-sections. Thus, official broadband statistics collected in panel form should be supported and expanded. The U.S. Federal Communications Commission recently recommended to the Census Bureau that the American Community Survey (ACS) questionnaire be modified to gather information about broadband availability and subscription in households.[21] However, given that the ACS does not resample the same households, perhaps official support would be better directed to panels such as the Current Population Survey from the U.S. Bureau of Labor

Statistics, a longitudinal survey which has asked questions relating to broadband in the past.

APPENDIX

This technical section deals with adapting the Census block-level observations to a household-level analysis for maximum likelihood estimation (MLE). The issues discussed here are unique to our dataset and can be ignored if household observations are available.

Let the number of households in a Census block be N. Define (compound) event A as the first household adoption of DSL not occurring until time interval $[t,t+\Delta)$, event B as the first adoption occurring before t, and event C as the first adoption not occurring until after $t+\Delta$. Since events A, B, and C are mutually exclusive and exhaustive, we have:

$$\Pr(A) + \Pr(B) + \Pr(C) = 1 \qquad \text{(A1)}$$

Since the complement of B is that all adopt after t, which has probability $S(t)^N$, we have $\Pr(B) = 1 - S(t)^N$. Similarly, $\Pr(C) = S(t+\Delta)^N$. Combining these facts with (A1) implies

$$\Pr(A) = S(t)^N - S(t+\Delta)^N \qquad \text{(A2)}$$

Taking a second-order Taylor's expansion shows that

$$S(t+\Delta)^N = S(t)^N + \Delta N S(t)^{N-1} S'(t) + o(\Delta^2) \qquad \text{(A3)}$$

where $o(x)$ means "terms of order x". Expressing the right side of (A2) as a rate, applying (A3), and noting that $S(t) = -f(t)$ (the p.d.f.) gives

$$\frac{S(t)^N - S(t+\Delta)^N}{\Delta} = NS(t)^{N-1} f(t) + o(\Delta) \qquad \text{(A3)}$$

Taking the limit of (A3) as $\Delta \to 0$ and explicitly noting the dependence of S and f on coefficients β gives the likelihood for an observation:

$$L_i(\beta) = N_i S(t_i;\beta)^{N_i - 1} f(t_i;\beta) \qquad \text{(A4)}$$

where the subscript denotes quantities and functions pertaining to observation i. Since $L_i(\beta)$ is proportional to N_i, that term can be ignored when maximizing the log likelihood. Dropping N_i, the rest of (A4) is equivalent to the likelihood of observing one household adopting at t_i and the other $N_i - 1$ households adopting after t_i. We can thus expand each Census block observation into separate, identical observations for each household, mark all but one of them as censored, and perform MLE on the expanded dataset. The block characteristics are assigned to each household for their covariates. To account for the fact that only one observation per Census block is available, the standard errors must account for clustering at (at least) the Census block level. In fact, we cluster at a higher level of observation in the text.

REFERENCES

Chaudhuri, A., Flamm, K., & Horrigan, J. (2005). An Analysis of the Determinants of Internet Access. *Telecommunications Policy, 29*, 731–755.

Crandall, R. W., Sidak, J. G., & Singer, H. J. (2002). The Empirical Case Against Asymmetric Regulation of Broadband Internet Access. *Berkeley Technology Law Journal, 17*(1), a, 953-987.

Duffy-Deno, K. T. (2000, September). *Demand for high-speed access to the internet among internet households*. PowerPoint Presentation at ICFC 2000, Seattle, WA.

Retrieved October 20, 2006, from http://www.icfc.ilstu.edu/icfcpapers00/duffy-deno.PDF.

Fairlie, R. W. (2004). Race and the Digital Divide. *Contributions to Economic Analysis & Policy, Volume 3*, Issue 1, Article 15.

Flamm, K., Chaudhuri, A. (2007). An analysis of the determinants of broadband access. *Telecommunications Policy, 31*, 312–326.

Howell, B., & Obren, M. (2002). *Broadband Diffusion: Testing for Vintage Capital, Learning by Doing, Information Barriers and Network Effects*. New Zealand Institute for the Study of Competition and Regulation Working Paper BH02/10.

Hsiao, C. (2002). *Analysis of Panel Data*, 2nd ed. Cambridge University Press.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Hoboken, NJ: Wiley.

Kridel, D., Rappoport, P., & Taylor L. (2001). The Demand for High-Speed Access to the Internet: The Case of Cable Modems. In Loomis, D., & Taylor, L. (Ed.), *Forecasting the Internet: Understanding the Explosive Growth of Data Communications*. Boston: Kluwer.

Leigh, A. (2003). *Digital Divide and Broadband Divide – Some Multiple Regression Results.* unpublished manuscript, from http://econrsss.anu.edu.au/~aleigh/pdf/Digital%20divide%20update.pdf

Moulton, B. R. (1990). An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *Review of Economics and Statistics, 72*, 334-38.

Prieger, J., & Hu, W. (2008). The Broadband Digital Divide and the Nexus of Race, Competition, and Quality. *Information Economics and Policy*, *20*(2), 150-167.

Rappoport, P. N., Kridel, D. J., Taylor, L. D., Alleman J. H., & Duffy-Deno, K. T. (2003). Residential demand for access to the Internet. In Madden, G. (Ed.), *Emerging Telecommunications Networks: The International Handbook of Telecommunications Economics (Vol. 2)* (pp. 55-72). Cheltenham, U.K.: Edward Elgar.

Savage, S. J., & Waldman, D. (2005). Broadband Internet Access, Awareness, and Use: Analysis of United States Household Data. *Telecommunications Policy, 29*, 615-633.

Stanton, L. J., (2004). *Factors influencing the adoption of residential broadband connections to the internet.* Proceedings of the 37th Annual Hawaii International Conference on System Sciences, from http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/05/205650128a.pdf .

Sueyoshi, G. T. (1995). A Class of Binary Response Models for Grouped Duration Data. *Journal of Applied Econometrics, 10*(4), 411-431.

Weiser, E. B. (2000). Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison. *CyberPsychology & Behavior*. April 1, *3*(2), 167-178.

Whitacre, B. (2007). Factors influencing the temporal diffusion of broadband adoption: evidence from Oklahoma. *Annals of Regional Science, 42*(3), 661-679.

Figure 1: S-shaped Adoption Curves

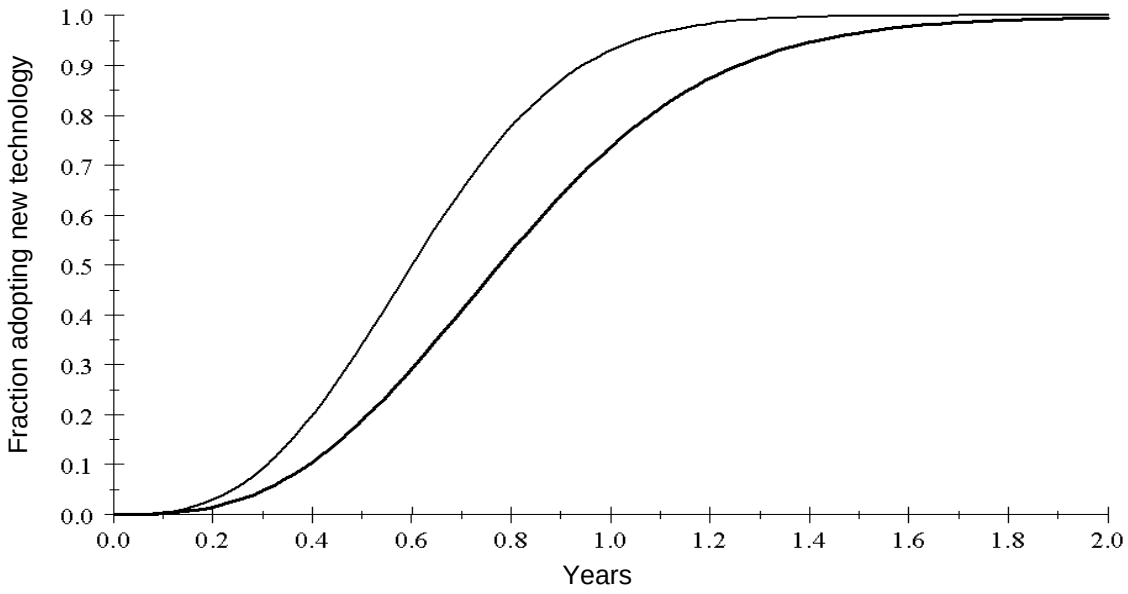

Figure 2: Estimated DSL Adoption Curves from the Duration Model

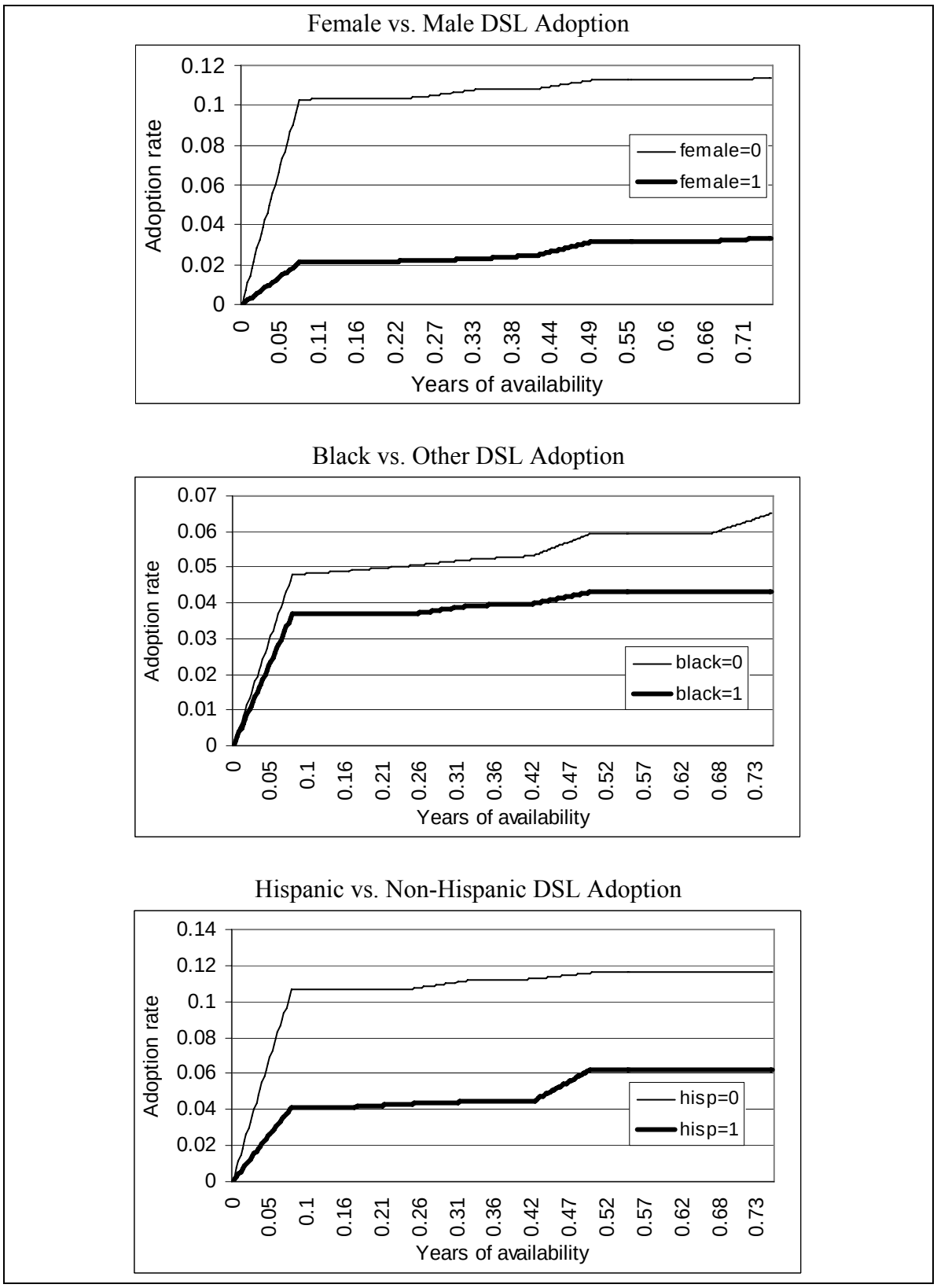

Figure 3: Estimated DSL Adoption Gaps from the Panel Data Model

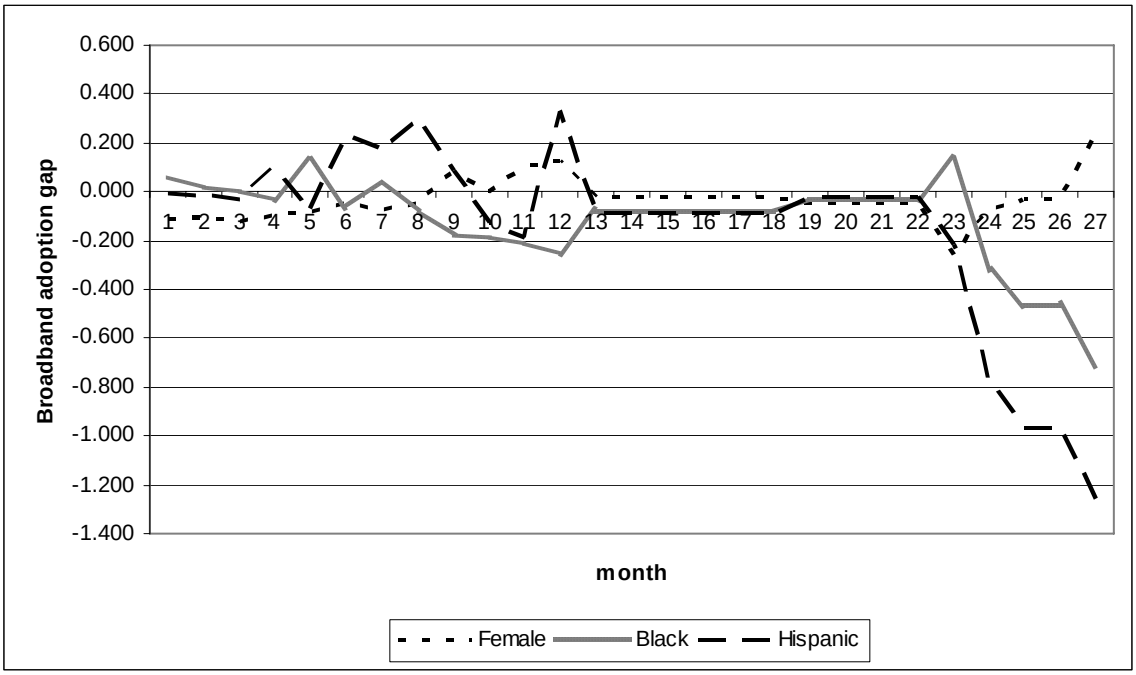

Table 1: DSL Adoption: Cross-Sectional Estimation Results

	Estimation 1 OLS		Estimation 2 Probit	
	coefficient	s.e.	marginal effect	s.e.
Female	-0.026	0.022	-0.026	0.020
Black	-0.033	0.022	-0.030	0.020
Hispanic	0.045	0.029	0.039	0.027
Asian	-0.147***	0.035	-0.131***	0.027
Other race	-0.075***	0.028	-0.066***	0.024
Income (log)	0.232	0.228	0.243	0.219
Income (log) squared	-0.012	0.011	-0.013	0.010
Household size	0.011***	0.004	0.012***	0.003
Number of Household	4.13E-4***	3.71E-5	0.001***	7.09E-5
R^2 (OLS)/Pseudo-R^2 (probit)	0.0657		0.0658	
N	51,796		51,796	

* = significant at the 10% level; ** = significant at the 5% level; *** = significant at the 1% level.

Notes: both estimations include local telephone service area fixed effects, not shown in the table. Standard errors are robust to heteroskedasticity and clustering at the block group level.

Table 2: DSL Adoption: Duration Analysis Results

	Estimation 3		Estimation 4	
Variable	coefficient	s.e.	coefficient	s.e.
Female	-1.353***	0.232		
Female*M1			-1.633***	0.253
Female*M2			0.349	0.865
Female*M3			-0.184	0.949
Female*M4			-1.557*	0.867
Female*M5			1.364	1.062
Female*M6			0.365	0.637
Female*M9			0.754	0.936
Female*M22			-3.343**	1.526
Black	-0.333***	0.101		
Black*M1			-0.268**	0.109
Black*M2			-10.417**	4.250
Black*M3			-3.041***	0.548
Black*M4			0.266	0.238
Black*M5			-0.504	0.395
Black*M6			-0.685***	0.217
Black*M9			-10.431	7.284
Black*M22			0.306	0.377
Hispanic	-0.674***	0.184		
Hispanic*M1			-0.982***	0.209
Hispanic*M2			-0.995	0.776
Hispanic*M3			1.536***	0.308
Hispanic*M4			-1.498***	0.550
Hispanic*M5			-4.326***	1.147
Hispanic*M6			1.582***	0.428
Hispanic*M9			-0.778	0.630
Hispanic*M22			-1.568**	0.625
Asian	-2.801***	0.381	-2.765***	0.380
Other race	0.189	0.246	0.309	0.245
Income (log)	-2.510***	0.266	-2.650***	0.261
Income (log) squared	0.136***	0.014	0.142***	0.013
Household size	0.433***	0.028	0.422***	0.028
Month 1	8.847***	1.340	9.819***	1.334
Month 2	4.786***	1.333	5.108***	1.360
Month 3	5.071***	1.351	5.258***	1.434
Month 4	5.762***	1.349	6.543***	1.401
Month 5	5.057***	1.363	4.718***	1.492
Month 6	6.782***	1.342	6.529***	1.367
Month 9	6.433***	1.350	6.431***	1.431
Month 22	5.845***	1.357	7.576***	1.543
χ^2 stat (*p*-value)	4,855.3	(0.000)	5,562.3	(0.000)
Pseudo-likelihood	-253,041.8		-252,085.9	
N	1,917,724		1,917,724	

* = significant at the 10% level; ** = significant at the 5% level; *** = significant at the 1% level

Notes: both estimations include state and calendar year fixed effects, not shown in the table. Standard errors are robust to heteroskedasticity and clustering at the block group level.

Table 3: DSL Adoption: Panel Probit Estimation Results

	Estimation 5		Estimation 6	
Variable	marginal effect	s.e.	marginal effect	s.e.
Female	-0.066*	0.035		
Female*M1			-0.111***	0.028
Female*M2			-0.107***	0.029
Female*M3			-0.118***	0.030
Female*M4			-0.099***	0.033
Female*M5			-0.079**	0.037
Female*M6			-0.052	0.046
Female*M7			-0.072	0.049
Female*M8			-0.049	0.051
Female*M9			0.086	0.062
Female*M10			0.002	0.071
Female*M11			0.096	0.076
Female*M12			0.125	0.084
Female*M13-M18			-0.021	0.079
Female*M19-M22			-0.045	0.079
Female*M23			-0.261**	0.123
Female*M24			-0.070	0.160
Female*M25-M26			-0.033	0.174
Female*M27			0.243	0.218
Black	-0.023	0.038		
Black*M1			0.056	0.037
Black*M2			0.018	0.037
Black*M3			0.004	0.038
Black*M4			-0.030	0.039
Black*M5			0.124***	0.042
Black*M6			-0.064	0.044
Black*M7			0.040	0.051
Black*M8			-0.084*	0.046
Black*M9			-0.180***	0.048
Black*M10			-0.189***	0.052
Black*M11			-0.216***	0.053
Black*M12			-0.256***	0.054
Black*M13-M18			-0.083	0.054
Black*M19-M22			-0.031	0.053
Black*M23			0.136	0.076
Black*M24			-0.327	0.094
Black*M25-M26			-0.466***	0.095
Black*M27			-0.711***	0.101
Hispanic	-0.005	0.060		
Hispanic*M1				
Hispanic*M2			-0.009	0.051
Hispanic*M3			-0.012	0.052
Hispanic*M4			-0.029	0.052
Hispanic*M5			0.104	0.059
Hispanic*M6			-0.084	0.054
Hispanic*M7			0.232***	0.078
Hispanic*M8			0.176**	0.079
Hispanic*M9			0.296***	0.087
Hispanic*M10			0.086	0.097

Continued from previous page

Variable	Estimation 5 marginal effect	s.e.	Estimation 6 marginal effect	s.e.
Hispanic*M11			-0.131	0.096
Hispanic*M12			-0.184*	0.101
Hispanic*M13-M18			-0.090	0.126
Hispanic*M19-M22			-0.021	0.127
Hispanic*M23			-0.219	0.135
Hispanic*M24			-0.781***	0.153
Hispanic*M25-M26			-0.966***	0.156
Hispanic*M27			-1.259***	0.169
Asian	-0.172***	0.047	-0.174***	0.048
Other race	-0.110*	0.060	-0.115*	0.060
Income (log)	0.218	0.472	0.202	0.476
Income (log) squared	-0.013	0.022	-0.012	0.022
Household size	0.004	0.006	0.002	0.006
Number of households	0.001***	0.835E-5	0.001***	0.850E-5
Month 2	0.027***	0.003	0.032***	0.007
Month 3	0.037***	0.004	0.051***	0.010
Month 4	0.084***	0.006	0.084***	0.013
Month 5	0.116***	0.006	0.100***	0.015
Month 6	0.174***	0.007	0.155***	0.017
Month 7	0.202***	0.007	0.187***	0.017
Month 8	0.198***	0.007	0.180***	0.017
Month 9	0.215***	0.007	0.178***	0.020
Month 10	0.216***	0.007	0.214***	0.019
Month 11	0.222***	0.007	0.203***	0.022
Month 12	0.249***	0.006	0.220***	0.021
Months 13-18	0.180***	0.007	0.171***	0.027
Months 19-22	0.164***	0.005	0.150***	0.029
Month 23	0.242***	0.008	0.260***	0.018
Month 24	0.283***	0.006	0.295***	0.007
Months 25-26	0.287***	0.006	0.297***	0.007
Month 27	0.287***	0.006	0.296***	0.007
Pseudo-R^2	0.3224		0.3296	
N	411,477		411,477	

= significant at the 10% level; ** = significant at the 5% level; *** = significant at the 1% level.

Notes: both estimations include local telephone service area fixed effects, not shown in the table. Standard errors are robust to heteroskedasticity and clustering at the block group level.

[1] We thus do not comment on the methodology of the many important case studies and qualitative analyses of the digital divide.
[2] With cross-sectional data, observations are taken from a single period, and the sample comprises different individuals, households, or geographic areas. Cross-sectional data thus provides a point-in-time snapshot of the phenomenon under study.
[3] Panel data consist of repeated observations on the same units of analysis in the cross-section.
[4] In this notation, ϕ is the Normal density function. Modern statistical software packages can calculate marginal effects automatically for probit and logit models.
[5] This simple relationship between the hazard rate and the mean holds only when the former is constant.
[6] The term *semiparametric* has different meanings in the statistics literature. Here we mean that the baseline hazard is modeled effectively nonparametrically and the effect of the covariates on the hazard rate is modeled parametrically.
[7] The nature of our data also lends itself to a discrete-time hazard model (see Kalbfleisch & Prentice, 2002), but the results would differ little.
[8] The authors have found both S-Plus and Stata to be particularly easy to use in this regard. We use the latter for this article.
[9] Chief among the advantages of the data are the large number of observations, the accurate information on the availability of DSL, and the fine geographic detail. See Prieger & Hu (2008) for a discussion of the strengths and weaknesses of these data.
[10] Census blocks are the smallest unit of Census geography, and there are only 23 households in the median block in our data.
[11] When estimating the effect of aggregated variables on a dependent variable at a lower level of aggregation, standard errors can be artificially small unless corrected by clustering methods. See Moulton (1990) for an illustration of the principles involved.
[12] Initial availability is determined by the first date any household in the local service area subscribes. Initial adoption in the block is available in the data.
[13] See Kalbfleisch & Prentice (2002) for a complete discussion of censoring in duration models. For the practitioner, the statistical software takes care of the details.
[14] The difference between the two models is likely to be more pronounced when the mean of the dependent variable is near zero or one.
[15] Only those months for which adoption is observed are represented with dummy variables in the specification. With no adoption observed in month 8, for example, the maximum likelihood estimate of the coefficient on the month dummy is negative infinity, and the hazard rate is zero for the month.
[16] The adoptions after 22 month all come from a single area in Detroit, the only area with DSL available for more than two years. Thus the results after nine months are likely to be unrepresentative anyway.
[17] The curves are calculated at the mean values of the other variables, which are month-specific in the case of the interacted variables.
[18] We cannot estimate a panel fixed effects model by adding a dummy variable for each Census block, because only the dependent variable varies over time, and none of the coefficients on the regressors would be identified. We can, in theory, estimate a panel random effects model (in which the intercept for each block is treated as a random variable to capture unobserved heterogeneity). However, our large sample size and number of regressors precluded estimation of a panel probit random effects model. In a half-sample version we did estimate, the results for the gender, race, and ethnicity coefficients were similar to that of Estimation 5 below.
[19] In the figure, the gap for women is with reference to men, the gap for Hispanics is with reference to non-Hispanics, and the gap for blacks is with reference to whites.
[20] The probit discrete duration model implies a lognormal, rather than constant, hazard rate within each period and has covariate effects that are far from proportional. Given that within-period hazard rates cannot be identified nonparametrically and that assumptions on their shape cannot be tested with discrete data, we assume the simplest possible form: constant (Sueyoshi, 1995).
[21] See *Report and Order and Further Notice of Proposed Rulemaking*, FCC 08-89, released June 12, 2008.