How does income relate with households call pattern?

Wenxuan Gu

Abstract

Landlines use was popularized after 1960s. A dramatic increase of landline demand occurred in 1970s. In late 1984, the Kentucky Public Service Commission (KPSC) made an economic investigation by measuring local telephone rates mainly in Louisville. This paper uses datasets from KPSC to study the relation between local households' income and Louisville households' call patterns with detailed demographic factors.

1. Introduction

In mid-1950s, the landline phones played an important role in political movements. Protestors applied landlines to stay informed in certain political movements such as Montgomery bus boycott. (The Telephone and Its Uses in 1980s U.S. Activism, 2018) The frontier telephone service companies Wide Area Telephone Service (WATS) and the well-known American Telephone and Telegraph (AT&T) were both introduced soon after 1960s. (The Telephone and Its Uses in 1980s U.S. Activism, 2018) Both of these companies provided two kinds of service plans. One plan allowed unlimited long-distance calls for a fixed charge and another plan was set to be more economy which charged with a cheaper fixed cost plus other variable costs such as duration, peak periods, and distance.

In 1984, landline phones became a popular and dispensable communication tool for household's daily use. An increase demand for telecommunication occurred since 1970s. Analyses found that the service price change has a strong effect to households' demand, especially to low-income households. More specifically, analyses showed that local telephone rates played a more important role than fixed charge in regarding to household demand. (Telephone pricing structures: The effects on universal service, 1991) The census report showed that the median family income was $26,430 and the poverty threshold for a family of four in 1984 was $10,609. (Money Income and Poverty Status of Families and Persons in the United States: 1984, 1985)

A one-year unlimited telephone plan would be a total of $224.4 in 1984 which was about a fixed 2% of total income in a low-income family. Since choosing a measured plan might be even cheaper which indicated that the maximum telephone cost would be no greater than 2% of a low-income family. Questions might be raised: "Is it true that low-income households make less phone calls than higher income households during all periods even though the cost of telephone service was relatively cheap in 1984?" "How do households income impact their call pattern?"

2. Data

2.1 Data Source

The dataset used in this paper is mainly assembled from one primary source. The Administrative Case No. 285 economic study established by Kentucky Public Service Commission (KPSC) contains a total of 1998 randomly sampled households in Louisville. The sampled households' demographic information was pre-surveyed in July 1986 as the economic study firstly took in place. Case No. 285 awarded the sampled households before the experiment officially started and recorded households' detailed telephone usage for three continuous months from October 1986 to December 1986.

2.2 Summary Statistics

The experiment started in the last three months of 1986. An unlimited option with a monthly fee of $18.70 or a measured service option with a monthly fee of $14.02 were randomly chosen by sampled households in Louisville before the study officially began. The measured service tariff option was set to include four different parts: a fixed amount of $5 setup fee, duration, periods, and distance. Duration was set to measure the total time elapse in an individual call which had unit of per minute. Period was designed to differentiate the three different periods in a call which included: peak hours from 8 a.m. to 5 p.m. on weekdays; shoulder hours from 5 p.m. to 11 p.m. on weekdays and Sundays; off–peak hours from any other excluded time periods. Distance was set to measure two different origin bands in individual calls which includes Band A and Band B.

As an overview, I implement two tables with useful variables that might be use in further studies as presented. Two tables below provide brief descriptions based on summary statistics on demographic variables and call pattern variables:

Table 1
Summary statistics for significant dummy variables

Variable	Mean	Std. Dev.	Min	Max
tariff	0.335	0.472	0	1
age	3.369	1.655	1	6
hhsize	2.45	1.434	1	16
educ	2.609	1.379	1	6
married	1.824	0.842	1	3
race	1.198	0.615	1	5
income	4.656	2.257	1	9
benefits	0.284	0.451	0	1

Notes: All variables contain 1998 non-missing observations. All variables are dummy variables expect one continuous variables which is hhsize(# in household).

Table 2
Summary statistics for call pattern variables

Variable	Mean	Std. Dev.	Min	Max
CAP	45.733	51.689	0	480
CAS	37.504	46.925	0	609
CAO	30.656	37.874	0	533
CBP	12.897	18.256	0	272
CBS	13.001	17.211	0	245
CBO	9.059	11.657	0	127
MAP	180.271	222.654	0	1930
MAS	191.327	252.086	0	2222
MAO	132.576	189.643	0	1899
MBP	62.498	103.279	0	1490
MBS	84.991	131.681	0	1255
MBO	48.387	81.943	0	920

Notes: All variables contain 1998 non-missing observations. Each abbreviation variable has its meaning which can be interpreted as the following: M = minutes, C = number of calls; A = band A, B = band B, P = call occurs during peak period, S = call occurs during shoulder period, and O = call occurs during off-peak period.

Table 1 shows a summarization of households' general demographic information which includes important variables such as households' age, households' size, income, education level, marital status, and race. Tariff and benefits are listed but are not considered as main factors in further modeling process because of their small effects. **Table 2** generates an overview of all call pattern variables. This paper mainly focuses on using overall calls instead of minutes as factors to generate further regressions. Therefore, "minute-based" call pattern variables are listed here for reference instead of regression use. This paper expects to find the relations of income level and related factors in both tables. Income predictions will also be presented after regression analyses.

Figure 1 and figures 2 are two bar graphs that describe general overviews of the average total calls occurred and total call minutes with different levels of income during all time periods. The first figure shows a total average of over 225 calls occurred during all periods with level 1 income households which are households who make less than $5000 annually per year. In comparison, households who make between $20000 and $25000 annually (level 9 households) tend to make a 75 less calls on average which is about a 400 minute less than households with relatively lower income. I present these two figures because further regression process will mainly focus on specific implications and interpretations between income and call patterns.

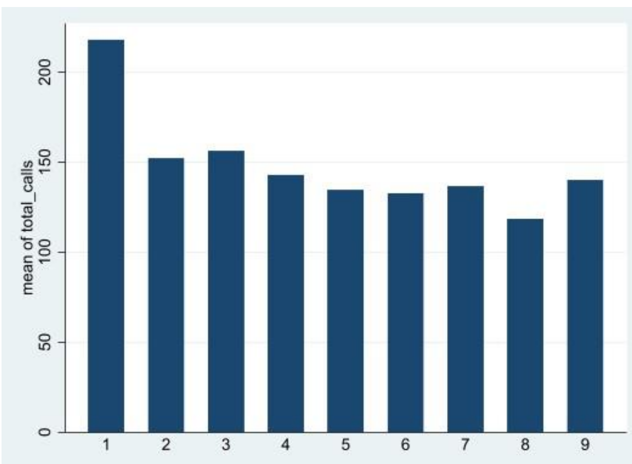

Figure 1

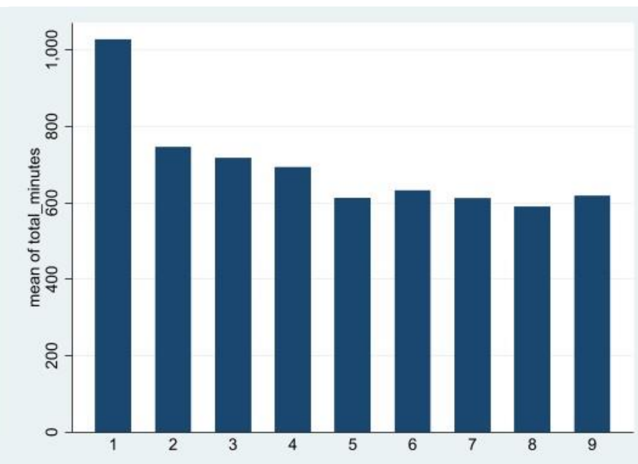

Figure 2

3. Methods

3.1 Population Model

My primary topic mainly focuses on studying the impact of household's income to specific call patterns. In order to dive into more specific findings regarding income and call patterns. I simply begin with deriving the relationship between people's income and the proportion of peak hour calls. By solving this question, variables of *CAP*, *CAS*, *CAO*, *CBP*, *CBS*, *CBO* are used to generate a new variable named *prop_peakcalls ((CAP + CBP) / (CAP + CAS + CAO + CBP +*

CBS + CBO)) which means the proportion of average total calls that occurred based on different income-level households. By taking a detailed view of how households income level correlate with *prop_peakcalls*, a basic regression of *prop_peakcalls* on *i.income* is derived. My first basic regression is implemented to provide a general overview of the trend of calling pattern during peak hours. By implementing similar steps of generating new variables *prop_offpeakcalls* and *prop_shouldercalls* by using the same method as how I generate *prop_peakcalls* help to provide an overall insights of three different proportions during three periods separately corresponding to the households with categorized annual income level. Three basic overview regression models on proportion of calls over income level are presented as following:

(1) prop_peakcalls = ß0+ß1i.income+u

(2) prop_shouldercalls = ß0+ß1i.income+u

(3) prop_offpeakcalls = ß0+ß1i.income+u

By observing the *prop_peakcalls* on *i.income* regression result, I find that all coefficients present negative signs on a total of 9 different level income households except income-level 3 households which indicates that when other variables are set fixed, households income level increasing from level 2 to level 3 is correlated with a 0.006 increase of call proportion in peak hours. However, households with higher income-levels tend to call less during peak hours. In other words, it shows that as households income level increases, the proportion of calls occur during peak hours decreases which also forms a negative relationship. In order to find that if a similar pattern occur in the proportion of calls during off-peak hours, the basic regression presents a similar result which shows a positive relationship between call proportions and income levels except the gap between income-level 2 and 3.

In order figure out more specific relationships between income level and call patterns, I implement 4 new interaction terms in my model which are: *CAP*age*, *CBP*age*, *CAO*age*, *CBO*age*, *CAS*age*, and *CBS*age* since I assume that age is significantly related to amount of calls. (Phone Behavior and its Relationship to Loneliness in Older Adults, 2017) In order to satisfy the zero-conditional mean (ZCM) assumption — E(unobserved error | calls during all periods) = 0, my first fundamental model must satisfy two conditions: (1) Cov(income, calls during all periods) = 0 and (2) Cov(income, all terms besides calls during all periods) = 0. In order to satisfy two conditions in ZCM, my first model (**M0**) is presented as following:

$$Income = ß0+ß1CAP*age+ß2CBP*age+ß3CAO*age+ß4CBO*age+ ß5CAS*age+ß6CBS*age+u$$

By observing the coefficient of income on *CAP*age*, and *CBP*age*, all age groups have positive relationship with income which means in general a higher income level is associated with more calls made during peak hours. By observing the coefficient on *CAO*age and CBO*age*, I find that in band A all coefficient signs are positive, however, in band B, age group of 2, 3 and 4 (households age from 35 to 64) has a negative relationship with income. This evidence implies that younger age households who call during off peak hours are associate with a higher income level and households who call during off peak hours are associated with a high retirement pay.

After observing the t-statistics in all interaction terms, I find that the magnitude of t-values vary a lot from each other in three different time periods which indicates that the age of households might not completely explain a static calling pattern. In addition, the R-squared is 0.2 which means that model one can only explain 20% correlation between income level and regressors. Thus, this fact motivates me to generate a more comprehensive model that contains more related variables. In order to make my model more holistic and convincible, I apply race and education as two another important variables since a research based on different racial groups using phones shows the evidence that "Blacks and Hispanics reach for their phones more often than whites when it comes to looking up information about health conditions, jobs or educational content." (Racial and Ethnic Differences in How People Use Mobile Technology, 2015) In addition, a research concludes that "couples share a bank account to pay for shared bills such as rent, utilities or phone plans in a long-term relationship." Thus, I assume that marriage and size of household should also be included as a variable that might related with phone bills and income level. I implement a new model that contains 4 new variables which are: *black, educ, married and hhsize.* I originally assume that household's race and their education level might have a correlation together in terms of income level. However, after several tests on STATA by regressing with new variable *black*educ*, it specifically shows collinearity problem in certain race groups. In order to avoid omitted variable problems, I present my second model (**M1**) as following:

*Income=ß0+ß1CAP*age+ß2CBP*age+ß3CAO*age+ß4CBO*age+ß5CAS*age+ß6CBS*age+ ß7black+ß8educ+ß9married+ ß10hhsize+ß11age+u*

By observing my **M1**, R-squared also increase twice much from 0.20 to 0.40 comparing with **M0.** The magnitude of t-values become relatively smaller on interaction terms but still not desirable. The significance on interaction terms are not normal which means that some P-values on interaction terms are too large. After running a basic regression on age with my call pattern variables, I find that variable *age* does not explain call patterns in region "B," in other words, using age to create interaction terms with *CBP*, *CBS* and *CBO* might not be "good" indicators. Thus, I implement three new interactions with education which are: *CBP*educ*, *CBS*educ* and *CBO*educ* as my new explanatory variables in my **M2.** In addition, I do not choose race to interact with call pattern variables because they generate multiple omitted variable problems.

Therefore, my adjusted **M2** is presented as following:

*Income=ß0+ß1CAP*age+ß2CBP*educ+ß3CAO*age+ß4CBO*educ+ß5CAS*age+*

*ß6CBS*educ+ß7educ+ß8married+ ß9hhsize+ß10black+ß11age+u*

The regression result shows a same R-squared value which is about 40% and the newly created interaction terms do present some relatively smaller P-values, but some are still large which will be discussed further in conclusion.

3.2 Model Assumptions

By comparing with my simple model **M0 assumption** as below**:**

Income=ß0+ ß1CAP+ß2CBP+ß3CAO+ß4CBO+ß5CAS+ß6CBS+ß7educ+ß8married+

ß9hhsize+ß10black+ß11age+u

M1 and **M2** both satisfy 7 assumptions of the classical linear regression model; I implement the following steps to satisfy the required assumptions:

The first step is two verify that each model is specified and linear with error term. The second step is to verify each model does not violet ZCM. In order to satisfy ZCM, I only apply a total of 6 "C(calls)" call pattern variables as my regressors to avoid repeatedly using "M(minutes)" and "C(calls)" variables in a same model. The third step is to avoid linear collinearity problem which indicates neither all my explanatory variables are not correlated with my error term nor they are correlated with each other. The fourth step is to fix the problem of heteroskedasticity by adding "robust" command in both of my models. The last step is to confirm that error terms are normally distributed in **M1** and **M2**.

3.3 Estimations

In order to show how **M1** and **M2** works, my estimation of using “lncom” command to predict income is explained explicitly as following in part *(1)*:

*(1) lncome=_cons+2*age+2*(i.age#c.CAP)+ 2*(i.age#c.CAO)+2*black +2*educ+ 2*married+ hhsize*

Note: example (1) assumes a household’s age at level 2(between 25 to 34), call pattern contains CAP and CAO, race is black, education level is “high school graduate,” marital status is “divorced”, and household size is 2.

The prediction result shows the mean is 3.36 which means this model predicts that a household with above “conditional” features will fall into the income level between level 3 and level 4 ($7500<annual income<14999).

In order to show the estimation of each variable’s marginal effect, command “margins, dydx” is implemented to generate an overview of each discreate variables partial effect of predicting regression outcome with 95% confidence intervals as shown in part *(2)*:

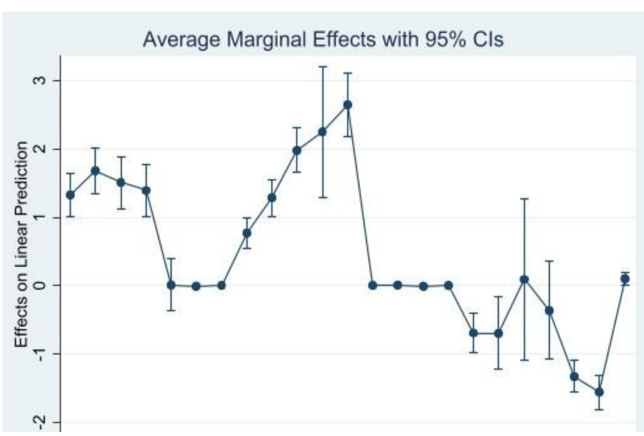

Figure 3

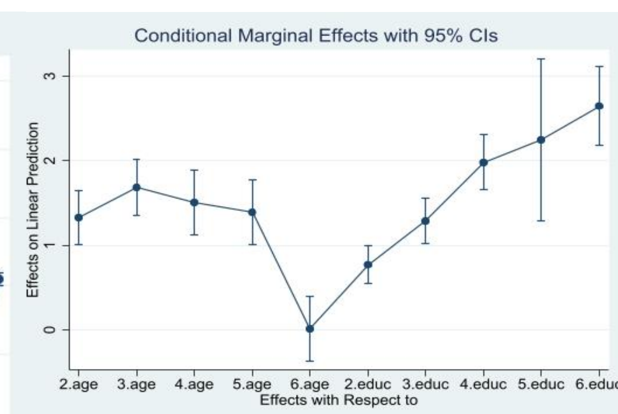

Figure 4

Note: Figure 3 shows the average marginal effects with respect to each variable in M2. Figure 4 shows a conditional marginal effects with respect to variable age and educ as other variables set at means in M2.

3.5 Result

Variable	*M1*	*M2*
income		
CAP*age	***0.021 (0.028)***	***0.048 (0.028)***
CAO*age	***0.038 (0.036)***	***0.031 (0.038)***
CAS*age	***-0.093 (0.030)***	***-0.077 (0.035)***
CBP*age	***0.087 (0.070)***	
CBO*age	***0.002 (0.076)***	
CBS*age	***0.105 (0.048)***	
CBP*educ		***-0.049 (0.046)***
CBO*educ		***0.131 (0.116)***
CBS*educ		***0.015 (0.078)***
age	***0.924 (0.198)***	***1.129 (0.188)***
educ	***1.717 (0.131)***	***1.637 (0.164)***
black	***0.393 (0.199)***	***-0.415 (0.304)***
married	***-1.437 (0.103)***	***-1.442 (0.094)***
hhsize	***0.108 (0.047)***	***0.104 (0.049)***
observations	***1,998***	***1,998***
F	***50.51***	***46.19***
p-value	***<0.001***	***<0.001***
R2	***0.402***	***0.401***

Table 3 – Regression output for models M1 and M2.

Note: Estimated regression coefficients are reported with robust standard errors in parentheses for both models. Standard errors are presented with significance at 5% level.

4.Conclusion

In general, **M1** and **M2** are both reliable in predicting different household's income level based on call patterns and their categorized demographic information. I find that asian and white households with age between 35 to 44 and over 65 years old who are married would like to call relatively more during peak and off-peak hours from band A and they are usually associated with a relatively higher income. However, as I stated in **M2**, I find that education level might be more related with call patterns and income level. My data shows that asian and white households with education of "some college" and "graduate school" from band B who are married tend to call more frequently during peak hours but have a negative relationship with income level. Besides these two unique finds, my result of off-peak call patterns in both bands shows a positive relationship with income level but slightly impact on income which does meet my expectation of considering off-peak hours as a minor factor in terms of determining household's income level.

Besides the findings, further discussions can also be made here. As I construct **M2**, I specifically find that age group 6 might be an arguable group. After several tests by interacting age group 6 with other variables, all results show large significant values which is not expected. However, the other arguable fact shows that some of my interaction terms also show large significant values, but it turns out that my prediction result is normal. I find that the main reason is because variable *age* and *educ* are closed related call pattern variables but call pattern variables might not closely relate to income level which ends that *age* and *educ* becoming the main effects to income level, however, significance (P-value) on interaction terms can hardly be interpreted. Again, my improved model **M2** does provide a job of predicting households' income by relating with interaction call pattern terms and the output results also show rational prediction values in terms of given telephone data.

References

Bureau, US Census (1985). “Income and Poverty Status of Families and Persons: 1984 (Advance Data).” *The United States Census Bureau*, p60 – 149

Barbara Keys (2018) "The Telephone and Its Uses in 1980s U.S. Activism," Journal of Interdisciplinary History, volume 48, p.485 – 509.

Cain, P., Macdonald, J.M. Telephone pricing structures: The effects on universal service. *J Regul Econ* **3,** 293–308 (1991). https://doi.org/10.1007/BF00138470

Petersen, Johanna, et al. “Phone Behavior and Its Relationship to Loneliness in Older Adults.” *Aging & Mental Health*, U.S. National Library of Medicine, Oct. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4698244/.

Anderson, Monica. “Racial and Ethnic Differences in How People Use Mobile Technology.” *Pew Research Center*, Pew Research Center, 30 May 2020, www.pewresearch.org/fact-tank/2015/04/30/racial-and-ethnic-differences-in-how-people-use-mobile-technology/.

“How Marriage Affects Your Legal and Financial Status.” *LawDepot Blog*, 8 June 2020, www.lawdepot.com/blog/how-marriage-affects-your-legal-and-financial-status/.