Explaining the Length of Housing Choice Voucher Waiting Lists

Josh Pelayo, Ben Westermeyer, and Luke Zimmerman

12/18/2017

# Introduction

In 2016, the U.S. Department of Housing and Urban Development (HUD) released information about the subsidized housing programs that it sponsors around the country. The data set contains 13,639 observations, each one representing a specific affordable housing program in a specific U.S. county. An example observation is Public Housing in Bibb County, AL. Of all the different programs, we are particularly interested in Housing Choice Vouchers (HCV). According to HUD, “HCV, formerly known as Section 8, is the federal government's major program for assisting very low-income families, the elderly, and the disabled to afford decent, safe, and sanitary housing in the private market" (HUD.gov). Under the HCV program, participants can find and lease any private housing that meets HUD's health and safety requirements. Local public housing agencies (which receive funding from HUD) then pay the landlord a subsidy to lower the payment required of the participant. The flexibility that HCV grants to its participants makes it the most popular form of affordable housing today.

The 74 variables in the data set provide information on the demographics, health, age, income, and other circumstantial information of each program in each county. Because the demand for Housing Choice Vouchers usually exceeds the supply, the public housing agencies that administer the subsidies often have extensive waiting lists. Using the months\_waiting variable as a response, we investigate which demographic, geographic, and socioeconomic factors influence the amount of time that people spend on HCV waiting lists across U.S. counties through multiple linear regression.

Because 74 variables were more than we needed for this analysis, we selected just 23 to explore in depth. We made this discernment by reading the meaning of each variable and deciding which we thought would be related to waiting times. After creating some new variables from the existing data, the number of variables of interest rose to 29. Although not all 29 of these variables would be included in our final model, they were all of interest to us initially. Below is a table of these variables.

### Definitions of Important Variables

|  |  |
| --- | --- |
| Variable.Name | Meaning |
| months\_waiting | avg months a household spends on waiting list before getting housing |
| logmonths | log(months\_waiting) |
| pct\_movein | % of households that entered the program less than a year before this snapshot |
| person\_income | avg household income per person per year |
| pct\_wage\_major | % of households where the majority of income is derived from wages |
| tpct\_ownsfd | % of households who are occupants of single-family detached homes |
| rent\_per\_month | avg household contribution towards rent per month |
| pct\_occupied | % of available units that are occupied |
| people\_per\_unit | avg people per unit |
| pct\_white\_nothsp | % of households in which the head is white |
| pct\_black\_nonhsp | % of households in which the head is black |
| pct\_native\_american\_nonhsp | % of households in which the head is Native American |
| pct\_black\_hsp | % of households in which the head is black hispanic |
| pct\_wht\_hsp | % of households in which the head is white hispanic |
| pct\_other\_hsp | % of households in which the head is other-hispanic |
| pct\_asian\_pacific\_nonhsp | % of households in which the head is Asian or Pacific Islander |
| hh\_income | avg household income per year |
| state | U.S. state the county belongs to |
| northeast | indicator variable that equals 1 if county is in the northeast U.S. |
| west | indicator variable that equals 1 if county is in west U.S. |
| southeast | indicator variable that equals 1 if county is in southeast U.S. |
| southwest | indicator variable that equals 1 if county is in southwest U.S. |
| midwest | indicator variable that equals 1 if county is in southwest U.S. |
| pct\_female\_head | % of households with a female as the head |
| pct\_disabled\_lt62 | % of households in which the head is disabled and younger than 62 |
| pct\_disabled\_ge62 | % of households in which the head of house is disabled and 62 or older |
| pct\_age62plus | % of households in which the head is 62 or older |
| pct\_1adult | % of households in which a spouse is not present and 1 or more children are present |
| pct\_median | avg household income of participants divided by median household income in local area |

### Data Cleaning: Removing Outliers and Filtering Missing Values

Once we narrowed the data set down to just include the 3067 Housing Choice Voucher observations and the 29 explanatory variables we were interested in, further data cleaning was necessary.

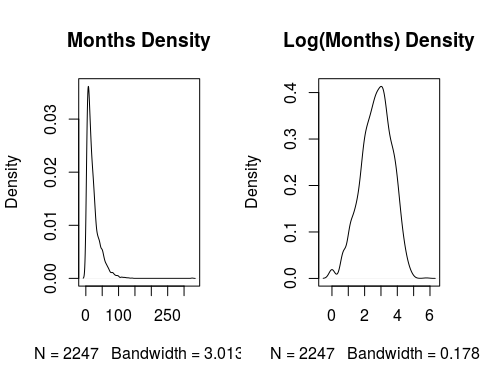
We filtered out all observations in which the months\_waiting response variable was either suppressed since there were less than 11 reported families in the county (represented by -4) or non-reporting (represented by -5). We also chose to remove all missing values (represented by -1), which eliminated 237 more observations. It is possible that -1s indicate that the average waiting time was 0 months. This would make sense given that there are 0 observations in which months\_waiting is 0. However, since there is no way for us to make a generalized inference about all of these missing data, we removed all of these -1 observations, leaving us with 2353 observations of Housing Choice Voucher programs in American counties.

When we looked at the distribution of months\_waiting in this HCV data set, we noticed one alarming outlier: Putnam County, FL had an average months\_waiting of 4830 months, which is equal to 402.5 years. The next highest average months\_waiting was 325. Putnam was clearly an error, so we removed it from the data set. After this removal we also noticed some strange pct\_occupied values greater than 100%. At first we thought this might be a signal of an overhousing problem in some counties. However, the definition of the pct\_occupied variable is occupied units as the percentage of units available. Since there cannot be more occupied units than available ones, we inferred that these were errors and we removed them from our data set. This left us with our final number of 2247 observations.

After removing problematic observations, we shifted our attention to recoding. We found that the data set had a large amount of -1s for the seven variables pertaining to the percentage of a certain ethnicity in a county. In our data set, a -1 implied that the data was missing from the model. However, we inferred the value of many of these -1s were 0 by adding the percentages of the other ethnic groups together, and finding that they were . For example, if pct\_white\_nothsp is 100, we are able to infer that the other six ethnicity percentage variables of pct\_black\_nonhsp, pct\_native\_american\_nonhsp, pct\_asian\_pacific\_nonhsp, pct\_black\_hsp, pct\_wht\_hsp, and pct\_oth\_hsp, were all 0, even if the original data set had a -1. For all of these variables, we converted all -1s to 0. We found that observations of the observations in our data set have a perfect ethnicity percentage. We inferred that the data were rounded to the nearest percentage point, given that observations have a total of and observations with a total of across these variables. The total spread is to , which we attribute to rounding errors within the data collection.

### Transforming Months Waiting to Log(Months)

We noticed the distribution of months\_waiting was skewed right, with the highest density right above zero. Trying to create a linear regression for a non-normal response variable would cause problems with our residual assumptions. Without a normally distributed response, our residuals would not be normally distributed, while linearity and equal variance may be compromised as well. One possible reason for this is the large variance in months waiting, even among similar areas. A 1997 study found that the median waiting time in the San Francisco area varied from months in San Mateo County to months in Richmond County. Furthermore, the median waiting time in all metropolitan areas was months (Painter, G, 1997). To correct our response variable, we performed a logarithmic transformation on months\_waiting to create a more normal distribution. The graphs below show that this transformation was effective. Thus, we created a logmonths variable and moved forward with this as our response variable.



### Sorting States into Regions to Create Categorical Variables

One of our primary research goals was to examine the variation of waiting times between different locations in the United States. If we used each of the 50 states as categorical explanatory variables, we would have a superfluous model. Instead, we created five new indicator variables to represent the five regions of the country: northeast, west, midwest, southwest, and southeast. We sorted all 50 states into these categories. See the appendix for which states are included in each region.

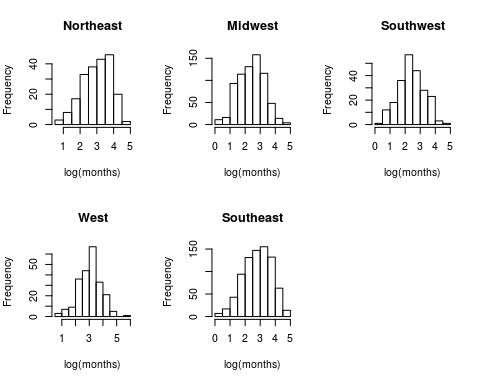
# Research and Results

Research Question: What determines the average length of time that people in a given county spend on the Housing Choice Voucher waiting list?

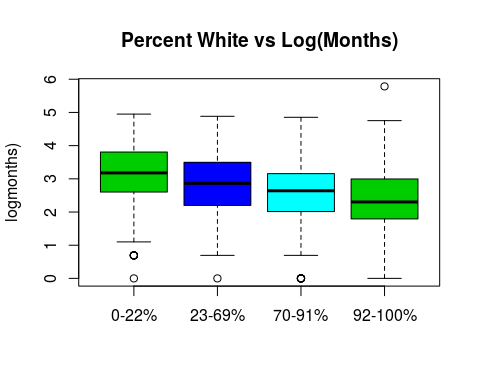
Null hypothesis: none of the coefficients of our explanatory variables will be significantly different from zero.

Alternative hypothesis: at least one of the coefficients will be significantly different from zero.

### Exploratory Data Analysis

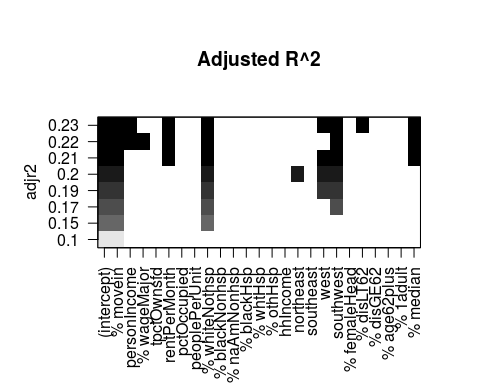


The above histograms show clearly that not all of the regions have the same distribution of waiting times. Midwest, Southwest, and West have relatively normal distributions while Northeast and Southeast both appear skewed left. Also, northeast and west have mean waiting\_times above , while midwest and southwest have values less than . Southeast is somewhere in between, with a mean of . This finding signals to us that these region variables will be valuable predictors of the variation in average months spent on HCV waiting lists.



To examine the relationship between the percentage of whites in a county and the average waiting time, we partitioned counties into four groups. The first group includes counties that are 0-22% white, the second is 24-69%, the third is 70-91%, and the last is 91-100%. The range of logmonths look fairly equal between the four groups, but as the percentage of whites increases, the average waiting time visibly decreases.

### Best Subsets Using our Variables of Interest (Excluding Reference Level Variables)



Notice that pct\_asian\_pacific\_nonhsp and midwest were not included in this best subsets analysis because we used these as reference levels. The best subsets analysis includes every other race percentage and every other region.

As shown in the adjusted graph above, the best subsets method suggested that we include the southwest and west region variables but exclude northeast and southeast. This is an unreasonable suggestion, since these indicator variables are essentially just components of one region variable. We either must include all of them or exclude all of them. In order to decide whether to include or not to include, we performed a nested F-test. The reduced model included all the terms that the best subsets method suggested except for southwest and west. The full model then added southwest, west, northeast, and southeast. The null hypothesis states that adding the region indicator variables explains an insignificant amount of variation in logmonths, making the F statistic zero. The alternative is that the F statistic > 0, meaning that the region variables influence logmonths significantly.

The nested F-test shows that the F statistic is 22.675, so we reject the null hypothesis. The p-value on such a large F statistic is less than 2.2e-16, which implies that it is extremely unlikely that regions would explain so much variation in logmonths by chance.

### Our Model

We decided to create our final model to predict logmonths with pct\_movein, person\_income, rent\_per\_month, pct\_white\_nothsp, northeast, west, southwest, southeast, pct\_median, pct\_disabled\_lt62, and pct\_female\_head. See the appendix for the summary of our model. We used best subsets to maximize the and minimize Mallow's CP for our model. The original best subsets did not include pct\_female\_head. However, it also did not include two of the region indicator variables, southeast and northeast. With southeast and northeast included in the model, we tried adding pct\_female\_head and found that the model's had substantially increased without compromising coefficient p-values. This is because we split region into indicator variables, and best subsets read each indicator variable separately, rather than reading them as different levels of a single variable. If regions were read as a single variable, best subsets would include pct\_female\_head.

The variables that are associated with an increase in logmonths are rent\_per\_month, northeast, west, pct\_disabled\_lt62, and pct\_female\_head. The variables that are associated with a decrease in the logmonths are pct\_movein, person\_income, pct\_median, pct\_white\_nothsp, southwest, and southeast.

From our results, we can see that each of the regions is significant except for southeast. If the region were in the northeast, months would increase by a factor of (p = 6.36e-06). If the region was in the west, months would increase by a factor of (p = 1.53e-09). If the region was in the southwest, months would change by a factor of (p = 1.45e-07). If the region was in the southeast, months would change by a factor of (p = 0.641580). It is worth noting that coefficient on the southeast indicator variable is not significant, while the other region indicator variables are highly significant.

Our next largest coefficient besides the region variables is pct\_median = -3.211e-02. For every percentage point increase in the average household income of participants divided by the median household income, there is a factor change in months waiting, controlling for all other variables. One possible explanation is that a county with a higher average household income of HCV participants has less income inequality, and thus the people participating in the program are in an economically similar situation to all other people in the county. Therefore, as the gap between the two groups decreases (which corresponds to a higher pct\_median), the average waiting time also decreases.

Our most interesting finding involves the pct\_white\_nothsp variable. For a percentage point increase in a county's percentage of white people, our model finds that months waiting changes by a factor of . This is an alarming result. If a county's white population increases by percentage points, it average waiting time is about shorter.

Our research indicates that region and ethnicity significantly impact the average time that participants in Housing Choice Vouchers spend on waiting lists. It is concerning that counties with higher percentages of white people are associated with lower average months on the HCV waiting list. Since local public housing agencies' ability to grant vouchers depends on funding from HUD, it appears that counties with more white people may be receiving more HUD funding, making waiting lists move along more quickly in those areas. Since the average waiting time for white participants is significantly shorter, we would expect a larger concentration of HCV-eligible housing in white neighborhoods. However, a 2001 study found that of census tracts in the U.S. with less than a population of African Americans received family public housing, a ratio of . In contrast, of tracts with greater than a African American population received family public housing, a ratio of . In the study, a ratio over 1 indicated a greater than expected percentage of assisted housing developments. (Rohe, W. M., & Freeman, L., 2001). This means that even though more HCV-eligible housing units are more often created in areas with large African American populations, the average waiting time for African Americans is still significantly longer.

### Multicollinearity

During the modeling process, we frequently found multicollinearity between explanatory variables. The data set contained two variables that measured the amount of single parent households in a county program: pct\_1adult and pct\_female\_head. While we wanted to test both variables, we were suspicious that they may have multicollinearity because every house with a single mom is also a house with one adult parent. Our suspicion was confirmed, as the variables had a correlation coefficient. Thus, even though we included both in the best subsets test, we were wary of any model that included both. Our final model only includes pct\_female\_head.

During our modeling process, we tried including interaction terms between region and pct\_white\_nothsp. We wanted to know if the effect of ethnicity on months waiting varied between regions. However, we found unacceptably high VIF values when we modeled with interaction terms, which is why our final model does not include any.

In our final model, the highest VIF is for person\_income, with a value . While is well below the acceptability threshold of , it still indicates that about of the variable's variation can be predicted by other variables in the model. We knew that multicollinearity would be an issue in our research since race, income, and family structure are often correlated with each other. Of all the models that we tested, our final model is one of the best in terms of low multicollinearity.

### Confounding Variables

One reason that northeast, southwest, and west are associated with such large increases in months waiting is possibly because they populous cities with large populations of low-income people and massive demand for subsidized housing. The southeast region does not have as many big cities. However, the midwest region includes big cities such as Chicago, Minneapolis, and Milwaukee and still has substantially lower waiting times than northeast or west. Thus, the existence of big cities might not be the only reason for variation in waiting times.

### Future Research

A weakness of our project is that the data only allowed us to examine waiting times at the county level. As a result, we can only make inferences about variation between counties, not individuals. A future research project could analyze the factors that determine the amount of time that individuals spend on HCV waiting lists. This could reveal some insights on how public housing agencies determine who to take off their waiting list first. There are official reasons that certain applicants are prioritized over others, such as domestic abuse or mental illness. However, there may be other factors that influence an applicant's time on a waiting list that housing agencies are not aware of.

As housing choice vouchers become available, eligible families are called from the waiting list, which opens periodically. Thus, the more quickly vouchers become available, the less time people spend on the waiting list. However, the availability of vouchers depends on funding. If HUD provides more funding to a public housing agency, it will administer more vouchers and the average time spent on the waiting list will be shorter. A future project could study the funding that HUD provides to local public housing agencies. It would be interesting to examine the factors that explain the variation in funding between agencies. How does the HUD determine how much funding to give to each PHA? Is there racial bias in the distribution of HUD funding to public housing agencies? Because our research found that predominantly white counties are associated with lower waiting times, and lower waiting times are likely due to more HUD funding, public housing agencies in white counties may be receiving more HUD funding.

## Sources

"Housing Choice Vouchers Fact Sheet." HUD.gov / U.S. Department of Housing and Urban Development (HUD), www.hud.gov/topics/housing\_choice\_voucher\_program\_section\_8.

Painter, G (1997). Does Variation in Public Housing Waiting Lists Induce Intra-Urban Mobility? Journal of Housing Economics, 6(3), 248-276.

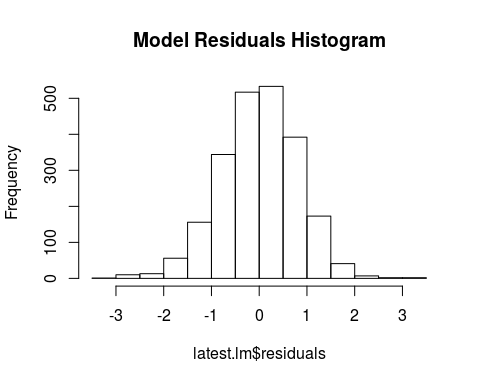
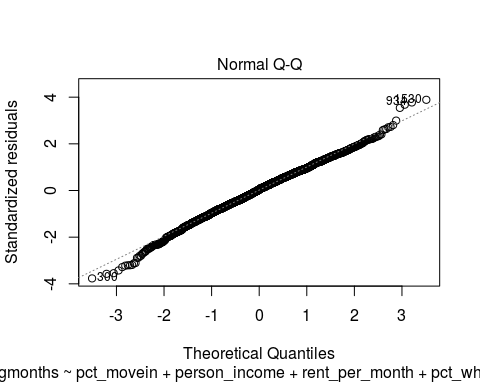
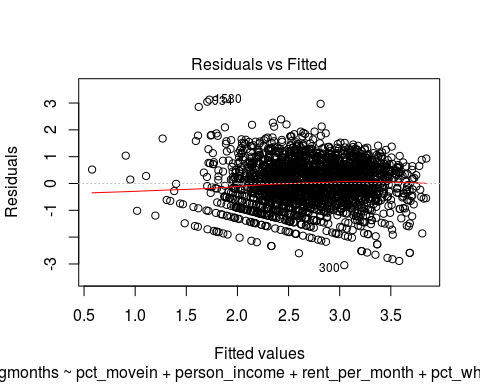
Rohe, W. M., & Freeman, L. (2001). Assisted housing and residential segregation: The role of race and ethnicity in the siting of assisted housing developments. American Planning Association. Journal of the American Planning Association, 67(3), 279-292.

## Appendix

#### Model Summary

Call:  
lm(formula = logmonths ~ pct\_movein + person\_income + rent\_per\_month +   
 pct\_white\_nothsp + west + southwest + northeast + southeast +   
 pct\_median + pct\_disabled\_lt62 + pct\_female\_head, data = hcv2016)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.04505 -0.52512 0.03303 0.55373 3.12646   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.842e+00 2.462e-01 11.544 < 2e-16 \*\*\*  
pct\_movein -1.458e-02 1.896e-03 -7.688 2.22e-14 \*\*\*  
person\_income -1.150e-04 1.864e-05 -6.171 8.04e-10 \*\*\*  
rent\_per\_month 2.684e-03 3.570e-04 7.519 7.95e-14 \*\*\*  
pct\_white\_nothsp -7.768e-03 7.730e-04 -10.049 < 2e-16 \*\*\*  
west 3.939e-01 6.493e-02 6.066 1.53e-09 \*\*\*  
southwest -3.459e-01 6.557e-02 -5.276 1.45e-07 \*\*\*  
northeast 3.037e-01 6.713e-02 4.525 6.36e-06 \*\*\*  
southeast -2.140e-02 4.598e-02 -0.466 0.641580   
pct\_median -3.211e-02 4.969e-03 -6.463 1.26e-10 \*\*\*  
pct\_disabled\_lt62 1.055e-02 1.708e-03 6.176 7.79e-10 \*\*\*  
pct\_female\_head 9.094e-03 2.470e-03 3.681 0.000238 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.8106 on 2235 degrees of freedom  
Multiple R-squared: 0.242, Adjusted R-squared: 0.2383   
F-statistic: 64.87 on 11 and 2235 DF, p-value: < 2.2e-16

Below are the diagnostic plots for our multiple regression model. Our residuals vs. fitted plot shows that our fitted values are centered around 0, which three points having leverage close to +/-3. The normal quantile plot shows that most points fall upon the line, such that the fit is linear. The histogram shows that the model has normally distributed residuals. Overall, the model reasonably meets the assumptions for multiple linear regression.



#### Confidence Intervals

2.5 % 97.5 %  
(Intercept) 2.3589911025 3.324427e+00  
pct\_movein -0.0182945445 -1.085837e-02  
person\_income -0.0001515449 -7.845457e-05  
rent\_per\_month 0.0019843306 3.384624e-03  
pct\_white\_nothsp -0.0092838002 -6.251934e-03  
west 0.2665440171 5.212045e-01  
southwest -0.4745368154 -2.173500e-01  
northeast 0.1721053459 4.353746e-01  
southeast -0.1115673633 6.875763e-02  
pct\_median -0.0418588119 -2.236992e-02  
pct\_disabled\_lt62 0.0071969556 1.389388e-02  
pct\_female\_head 0.0042494523 1.393812e-02

#### VIFs

FALSE pct\_movein person\_income rent\_per\_month pct\_white\_nothsp   
FALSE 1.214831 2.933726 2.002775 2.513416   
FALSE west southwest northeast southeast   
FALSE 1.304274 1.314571 1.305515 1.660220   
FALSE pct\_median pct\_disabled\_lt62 pct\_female\_head   
FALSE 1.406257 2.785086 1.842581

#### Region Variables

Northeast: CT, MA, ME, NH, NJ, NY, PA, RI, VT

West: AK, CA, CO, HI, ID, MT, NV, OR, UT, WA, WY

Midwest: IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI

Southwest: AZ, NM, OK, TX

Southeast: AL, AR, DC, DE, FL, GA, KY, LA, MD, MS, NC, SC, TN, VA, WV