**Regression Analysis and AR Calculations Supporting Documentation**

This document acts as a reference explaining the code files and supporting spreadsheets which were used to estimate housing costs and to calculate the affordability ratio. The code files contain the commands used within the statistical package R to manipulate the data and perform the necessary calculations. These files themselves contain sufficient comments to follow the sequence of specific operations that were performed on the datasets, even if the reader does not have an in-depth knowledge of how to use R. Combined with this document, the reader should be able to follow the analysis at whatever level of detail is desired.

In addition to using R to develop a regression model for estimating housing costs and to calculate the actual AR values, GIS was used to identify the intersections between service territory and census boundaries, as well as develop the weights that would be used for the aggregation process. Those weights were incorporated into the calculations that were performed in R.

**Pre-Analysis: Identifying income levels at various points on the income distribution**

*1\_PUMA level income and housing calculations.R* – this code file takes the raw PUMS household-level data on household income and housing costs (contained in the spreadsheet titled “PUMS\_curated\_with\_errors.xlsx”) and calculates the income levels that correspond to various points of the income distribution for each PUMA. These calculations use the sample weights that were included in the PUMS dataset, which are needed to account for sample design effects. Specifically, the income levels of various percentiles are calculated in lines 40 through 45 of the code file. The weighted mean income and housing costs are also calculated in lines 55 and 56. These values are needed to calculate the corresponding housing costs in the subsequent code files and are output to the file “1\_Weighted\_income\_and\_housing.csv”.

**Testing of Functional Forms and Model Specifications**

*2\_PUMA level regression of housing cost on income.R* – this code file uses the same raw PUMS household-level data as the previous code file (“PUMS\_curated\_with\_errors.xlsx”) and tests three different regression models with linear functional forms: 1) housing cost as a linear function of household income, 2) housing cost as a linear function of household income and household size, and 3) housing cost as a linear function of household income, household size, and an interaction variable of household income and size (household income x household size).

1. Housing Cost = a + b\*Income + error
2. Housing Cost = a + b\*Income + c\*Household Size + error
3. Housing Cost = a + b\*Income + c\*Household Size + d\*(Income\*Household Size) + error

For each model within each PUMA, the code file fits the model coefficients, calculates the r-squared value for the model in that PUMA (this is defined as the percentage of the variance of housing cost that is explained by the specification of the regression model), and determines the residuals for each individual data point (the residual is defined as the actual housing cost known from the PUMS dataset minus the predicted housing cost for that household; a positive value is an underprediction while a negative value is an overprediction).

These residuals are then normalized so that they can be compared across the 265 PUMAs. Normalization is achieved by dividing the residuals by the weighted mean housing cost of each PUMA, while the corresponding household income is normalized by dividing by the weighted mean household income of the PUMA. The normalized residuals across all PUMAs are then plotted for each model to see if there is any systematic bias in the regression model, and particularly whether that bias has a clear relationship with the independent variable of interest (household income). A regression model that shows such a bias would consistently over- or underpredict housing cost for certain income ranges. A regression model with such a pattern of bias in the residuals would do a poor job at predicting housing costs.

There are several outputs from the code file. “2\_Regression\_outputs.csv” provides the PUMA-level regression coefficients for the three linear regression models tested, the p-values of each coefficient, and the r-squared values for each model. Initially, it was thought that the r-squared values would be the most useful indicator of which model performed best at predicting housing costs. However, all three models produced similar r-squared values across the various PUMAs. It is worth noting that, although the r-squared values are generally fairly low, this is a result of the housing cost data showing a large degree of variance that is attributable to factors that are outside the scope of this analysis (value of property lot, age of house/apartment, interest rates at the time of property purchase, etc.).

Normalized residual values for each of household in the PUMS dataset are also output for each of the three models in the files “2\_Residuals\_no\_household\_size.csv”, “2\_Residuals\_with\_household\_size.csv”, and “2\_Residuals\_with\_interaction.csv”. The code file also outputs graphs of the normalized residuals, which are presented below.



Figure : Normalized Residuals for Linear Model with No Household Size Predictor



Figure : Normalized Residuals for Linear Model with Household Size Predictor



Figure : Normalized Residuals for Linear Model with Income/Household Size Interaction

The residual plots produced by the linear regression models show a pattern of systematic underprediction at lower income levels and systematic overprediction at higher income levels. This indicates that housing cost is not linearly proportional to the selected explanatory variables.

*3\_Nonlinear regression of housing cost on income.R* – this code file is very similar to the previous one in structure, but it tests two regression models with non-linear functional forms: 1) housing cost as a function of the square root of household income, and 2) housing cost as a function of the square root of household income and linearly proportional to household size. As with the previous code file, coefficients for each model were calculated for each PUMA individually based on household-level data.

1. Housing Cost = a + b\*$\sqrt{Income}$ + error
2. Housing Cost = a + b\*$\sqrt{Income}$ + c\*Household Size + error

This code file produces a similar set of outputs as the previous code file: “3\_Regression\_outputs\_nonlinear.csv” contains the model coefficients, p-values, and r-square values for each PUMA, while “3\_Residuals\_no\_household\_size\_nonlinear.csv” and “3\_Residuals\_with\_household\_size\_nonlinear.csv” contain the individual household normalized residuals for the respective models. Both models produced r-squared values that were similar to those produced by the previously discussed linear regression models, indicating that there is still a considerable amount of variance in housing cost that is not explained by household income and household size.

The normalized residual plots are presented below. The model still shows a large range of error values with considerable overprediction at extremely high income levels.



Figure : Normalized Residuals for Square Root Model with No Household Size Predictor



Figure : Normalized Residuals for Square Root Model with Household Size Predictor

At this point, it was decided that the model should exclude these extremely high income households since they would not be of interest for the study anyways.

*4\_PUMA level regression of housing cost on income\_no high income.R* and *5\_Nonlinear regression of housing cost on income\_no\_high\_income.R* – these are identical to the previous two code files discussed, except they remove from the PUMS dataset households that earn greater than five times the average household income for the given PUMA. They produce similar output files as those previously discussed for the prior two code files and also generate normalized residual plots, which are presented below.



Figure : Normalized Residuals for Linear Model with No Household Size Predictor and No High Income Households



Figure : Normalized Residuals for Linear Model with Household Size Predictor and No High Income Households



Figure : Normalized Residuals for Linear Model with Income/Household Size Interaction and No High Income Households



Figure : Normalized Residuals for Square Root Model with No Household Size Predictor and No High Income Households



Figure : Normalized Residuals for Square Root Model with Household Size Predictor and No High Income Households

Although the residual plots look similar (and much more reasonable than the models that included high income households), the linear models still showed a slight tendency towards overprediction at higher income levels as evidenced by the overall downward slope of the cluster of datapoints. The residual plot produced by the square root model with household size included as a predictor is closest to the ideal: errors scattered around zero with no systematic change in error magnitude or direction as a function of household income. This was the model selected for the final methodology.

**Geography**

The first step in calculating AR was collecting the necessary data for each industry and geography. Census blocks (blocks) and Public Use Microdata Area (PUMA) shapefiles were collected from the Census Bureau website through the Tiger/Line Files. Additional data was collected for the block shapefiles to include housing units, population, and household through the California Department of Financing (CDF) and estimated from 2010 to 2019 based on growth factors and growth keys year-to-year as of January 1 of the current year.

When there are discrepancies between the CDF and US Census Bureau (CB), such as boundary issues between cities, the CDF population estimates were modified each year to reflect CB geometry. New growth rates were calculated and incorporated into the shapefile’s attribute table, so that the CDF table and the CB shapefile attribute table match.

Public Use Microdata Survey (PUMS) data was collected from the CB which collects housing expenses and associated income levels on the PUMA level. The data was used to estimate the housing and income levels by using the Regression Analysis and AR calculations. Cross tabulation tables were used

**Utility Shapefiles and Essential service charges**

Utility shapefiles were collected differently between industries. Water collected the shapefiles from a publicly sourced data called Tracking California, where they provide the boundaries of Investor Owned Utilities (IOUs) and Municipal Own Utilities. This data was combined with information collected from the State Water Resource Control Board which collects water utility essential service charges and rates on 6, 12, 24 hundred cubic feet (CCF). Both data from the SWRCB and shapefiles were vetted to include community water systems and essential service charges for 6 CCF were used to allow for household level usage.

Communications boundaries were collected based on the carrier of last resort. The provided shapefiles were used as the carriers of last resort along with the essential service charges for each utility.

Gas and electricity IOU shapefiles were collected from the utilities. The public owned utility shapefiles were collected from the California Energy Commission’s website. Due to conflicting information such as overlapping territories, staff decided that if a utility that served one type of industry (only gas or electric) overlapped with a multi-service utility (both gas and electric), then the single service utility took precedence. In circumstances where two multi-service utilities overlapped, then the smaller territory utility took precedence over the larger utility.

The essential service charges for gas and electricity utilities were calculated using tariff sheets where available and staff was not able to obtain some publicly owned utilities essential service charges. The territories for the missing essential service charges were removed from the shapefiles and not included in the AR calculations.

**Aggregation**

*All\_Uti\_block\_20191216 –* The intersect tool was used to estimate a household’s essential service charge. This tool works by entering the shapefiles needed for intersecting and creates a new shapefile and attribute table based on the geographies. GIS removes the other data that is not overlapping. To prevent the loss of data, the areas missing essential service charges were treated as a “$0” essential service charge and merged with the industry utility essential service charges. The industry essential service charging and block shapefiles were intersected to create there are combinations of utilities that contain zero to four industry essential service charges along with the number of housing units, original block area, and the new intersected block area. Areas that contain 0 households were removed from the dataset to reduce the size of the shapefile.

**AR Calculation**

AR values were calculated for households at the 20th and 50th percentiles of the income distribution of each PUMA. Details on the overall process are provided in the methodology section of the revised staff proposal, and individual steps of the process are captured in the code files *14\_Arithmetic mean AR20\_new zones and top coding.R* and *15\_Arithmetic mean AR50\_new zones and top coding.R*. Both code files contain sufficient comments to follow along step by step.