Understanding the Factors that Affect Quality of Life in the Bay Area

Chris Chow

Economics 413

Spring 2019

I. Introduction

The Bay Area is a region in Northern California located around the San Francisco Bay. It is defined by nine counties: Alameda, Contra Costa, Marin, Napa, San Mateo, Santa Clara, Solano, Sonoma, and San Francisco. In recent years, the cost of living in the Bay Area has drastically increased. Households that make less than $105,350 are still considered low income in San Francisco (Sciacca 2017). In that area, the median price for a two-bedroom apartment is $4,500. The issue stems from multiple factors, but a glaring problem is the lack of resources for an ever-growing population.

At the heart of the Bay Area lies the Silicon Valley, an area rich in technology start-ups and industry giants. Companies such as Google, Apple, Facebook, and Tesla all have headquarters in the Bay Area. This dense conglomeration of companies has led to an influx of tech workers at a rate disproportional to employees in other industries to the point where "one in five Bay Area jobs are now in tech, and the region leads the nation in the growth of new technology jobs" (Treuhaft et. al 2018). The influx of workers has stimulated the Bay Area's economy and lowered unemployment rates, but it also has also created an uneven distribution in wealth. Blue collar workers and individuals making less money cannot compete for resources such as housing. A combination of increasing construction cost and increasing demand for housing has led to a staggering increase in housing prices (SFCED 2015).

I wanted to understand what factors have affected the quality of life in the Bay Area. I decided to define quality of life in economic terms by analyzing median income as my dependent variable. Using data from the United States Census Bureau's American Community Survey, I examined the effects of diversity, employment rate, education level, household size, gender, median age, population size, and percentage of population in the technology sector on the Bay Area's median income.

II. Prior Theory and Research

Research on the quality of life in the Bay Area has indicated a variety of reasons for the change in quality of life. To focus my discussion, I looked at economic and demographic factors. I was interested in how population variation has shaped the different counties in the Bay Area.

Data as far back as 1995 has suggested that San Francisco ranks lowest on the affordability index (Corley and Kroll 1995). This growth has been altered through the past two decades "in several ways, including the location of growth, age composition of the population, and ethnic makeup" (Corley and Kroll 1995). While these factors may not directly affect the median income in the Bay Area, the fluctuations in these factors through the past decade may provide strong indicators that are correlated to changes in income.

Treuhaft et. al highlights the additional factors of race, gender, and occupations in creating wage disparities. Their research found that "jobs in middle-wage industries have grown at a slower pace than those in low-wage and high-wage industries, further polarizing job growth" (Treuhaft et. al 2018). This San Francisco Center for Economic Development corroborated this finding, stating that "The region's challenges continue to be related to the

interplay of employment change, population shifts, and housing supply" (SFCED 2015). Essentially, competition for resources like housing have arisen from demographic and economic changes.

III. Data and Variables

I used cross section data that spans roughly 10 years. I had 103 observations from the United States Census Bureau's American Community Survey looking at the variation in factors from individual Bay Area counties between 2006-2017. I looked at 8 independent variables and their effects on my dependent variable, median income.

In addition to the Bay Area's nine counties, I added data from Santa Cruz County, a region south of the Bay Area. With the recent issues in housing and resource competition, many Bay Area residents are moving south into the Santa Cruz county. I thought adding data from Santa Cruz into the regression analysis would give a more holistic view of the Bay Area.

Table 1: Variable Descriptions and Predicted Signs

MedIncome	Dependent Variable, Median Income (dollars)	
Diversity	Percentage of Population that is white	-
Employed	Percentage Employed	+
Education	Percentage of Population with college education	+
Household	Size of Average Household	+
Male	Percent of Population that is Male	+
MedAge	Median Age	?
Population	Size of Population	?
Tech	Percentage of Population working in technology	+

MedIncome

MedIncome is my dependent variable. While I ultimately want to understand the effects on quality of life, I believe that median income is the most fitting indicator for economic and demographic factors.

Diversity

Diversity is actually a misnomer. The variable tracks the percentage of population that is white, non-Hispanic. I predict that the growing job market will bring in more diversity and replace the native residents, many of which are white. The growing diversity would mirror the growing median income and thus be a positive predictor of the dependent variable. However, much of the technology industry is dominated by white men, so this factor may have conflicting results.

Employed

I include the Employed variable to measure employment rates in each county. Data from the San Francisco Center for Economic Development have shown that unemployment has dropped in the

area. The growth in the tech sector has left a need for workers as the companies expand, opening many potential positions for Bay Area residents. I expect that the employment rate will positively correlate with median income as residents in the counties fill up the high paying jobs.

Education

The Education variable highlights the percentage of individuals in each county that has obtained a bachelor's degree or above. Most high paying jobs are specialized careers that require a college education. For example, software engineers in the tech sector typically have at least a bachelor's in computer science. I expect that the education variable should positively correlate with median income.

Household

The Household variable measures the average household size in each county. I expected that larger households must make more money to support the number of individuals so I believed that household size would positively correlate with median income.

Male

I include the Male variable to document the percentage of population that is male. I believe that the tech industry is male-dominated, so growth in that sector would mean bringing more men into the area. From this, I believe that male would be positively correlated with median income.

MedAge

The variable MedAge tracks the median age of the county. I do not know how the variable will correlate with median income because there are many factors that may affect the relationship. While older individuals may be paid more because they have had time to work up the corporate ladder, younger individuals are more heavily sought out to work for their energy.

Population

The Population variable measures the population in each county in a given year. Given the many reasons people decide to move into or out of a region, I could not properly assess the correlation between population and median income.

Tech

I included the Tech variable because I wanted to see how the Bay Area has been affected by the influx of workers in the tech sector. However, the variable from the data actually includes other professions. Specifically, the census data lists out the percentage of individuals in "professional, scientific, and management, and administrative and waste management services" (United States Census Bureau). This may skew the results because it does not only measure the tech sector.

Table 2: Descriptive Statistics

	Mean	Median	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis	Jarque-Bera	Probability	Sum	Sum Sq. Dev.	Observations
DIVERSITY	48.28252	44.30000	75.00000	31.30000	12.60047	0.604016	2.166132	9.247167	0.009818	4973.100	16194.73	103
EDUCATION	41.41748	41.20000	57.80000	23.90000	9.024829	-0.13545	2.249350	2.733209	0.254971	4266.000	8307.649	103
EMPLOYED	61.29417	61.10000	68.70000	55.10000	2.665900	0.068280	3.195423	0.243932	0.885178	6313.300	724.9165	103
HOUSEHOLD	2.706505	2.760000	3.000000	2.310000	0.196038	-0.67713	2.283493	10.07422	0.006492	278.7700	3.919942	103
MALE	49.68738	49.60000	51.20000	48.70000	0.681791	0.592949	2.326025	7.985049	0.018453	5117.800	47.41359	103
MEDAGE	38.85049	38.50000	46.10000	35.70000	2.358126	1.339842	4.471211	40.10634	0.000000	4001.600	567.1975	103
MEDINCOME	79609.10	74609.00	119035.0	63274.00	13331.03	0.934733	3.156621	15.10422	0.000525	8199737.	1.81E+10	103
POPULATION	824235.4	747373.0	1938153.	132173.0	560674.4	0.578674	2.064137	9.507298	0.008620	84896251	3.21E+13	103
TECH	15.50680	16.10000	25.40000	8.600000	4.232583	-0.03357	2.018951	4.149888	0.125563	1597.200	1827.305	103

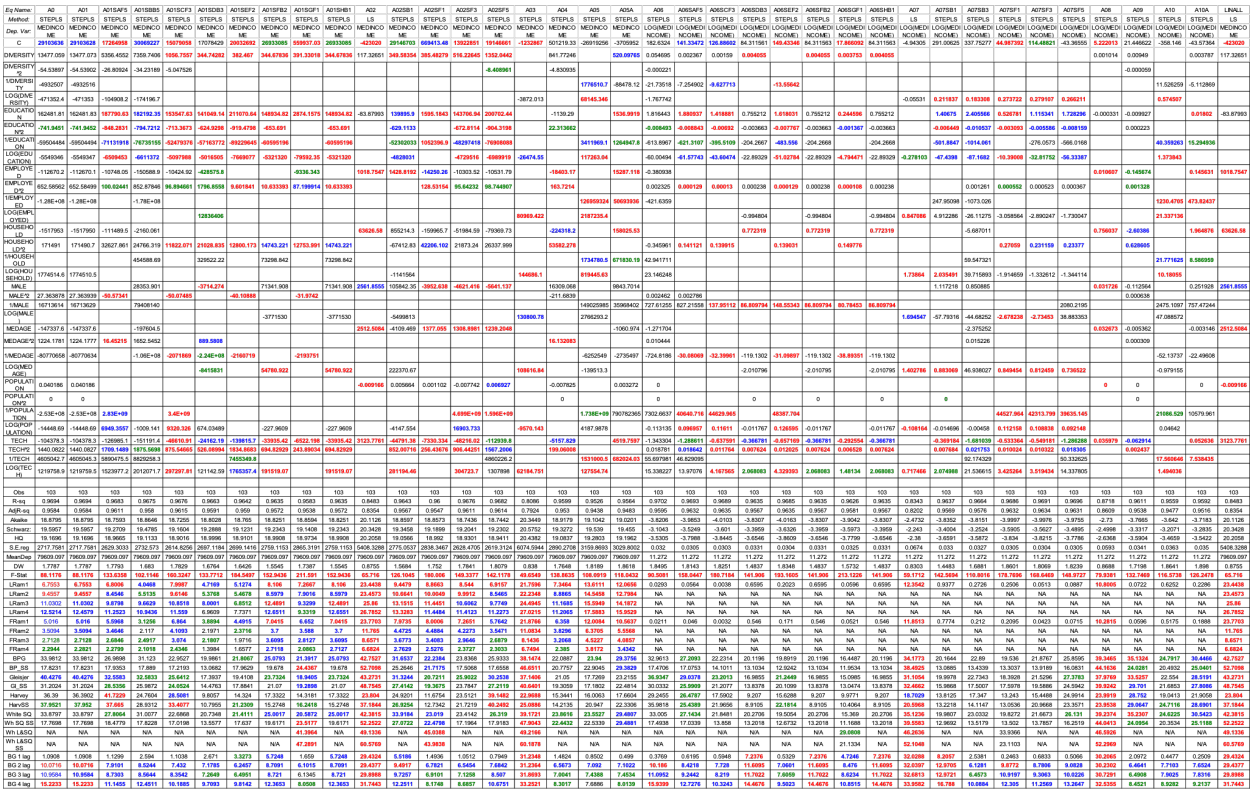
Discussion of Variables

From the table above, one apparent issue is the low standard deviation in the MALE and HOUSEHOLD variables. The relative uniformity among each observation's MALE and HOUSEHOLD obscure their effects on median income. There is not enough variation among observations to draw strong correlations because every county has roughly 50% male and 2.7 individuals in each household.

IV. Regression Analysis

I began with 38 regression equations, including the equation with all linear variables. I disregard the linear regressions because the superset, A01, fails the Ramsey test. All following linear regressions are low powered passes. My logarithmic superset, A06, passes the Ramsey test on all four terms, so I analyze the logarithmic regression equations. An issue is that the Ramsey test for the 2,3, and 4 term Ramsey test are NA, which means there is a non-singular matrix error. Some of the logarithmic regressions pass the Ramsey so I continue my analysis on these equations rather than selecting equations that do not pass the Ramsey.

Table 3: Regression Table



Analyzing the logarithmic equations, I find that A06 has the largest R-squared value and A06SCF3 has the largest adjusted R-squared value. Both equations pass the Ramsey at the first term and are NA for the rest.

I continue to analyze the Akaike, Schwarz, and HQ values. I find that A06SEF2 has the lowest values for all three criteria. The regression equation also has high R-squared and adjusted R-squared values similar to A06 and A06SCF3, respectively.

Final Equation

I choose to continue with A06SEF2 as my final equation. It has a high R-squared and adjusted R-squared value and it has the lowest Akaike, Schwarz, and HQ values. In addition, it passes the Ramsey 1 term test. This equation also includes many of the independent variables and most were statistically significant.

Looking at heteroskedasticity, I find that A06SEF2 passes all the tests at a 5% significance level, but not a 10% significance level. I assume that this is not significant enough to invalidate the classical assumption.

After accepting the 7 classical assumptions, I analyzed multicollinearity. A06SEF2 had a VIF value of 2084008 which is indicative of multicollinearity. However, all the variables are statistically significant and large t values, so I do not believe there is a multicollinearity problem.

Most of the signs from the regression equation are as predicted. My diversity variable, 1/Diversity, showed a negative correlation. This corroborates my idea that more diversity should

increase median income as workers move into the area for high paying positions. My diversity variable tracks the percentage of white, non-Hispanic individuals, so the variable proves that diversity is positively correlated with median income. Opening positions in these roles may attract more white men and thus drive down diversity. Education was confusing because all four variables for education were statistically significant. The Employed^2 variable was positive and agrees that higher percentages of employed individuals would increase median wage. While this connection may not be direct, the rapid growth of high paying jobs may explain the increase in median wage. The Household variable, Household^2 was also positive, showing that bigger average households is correlated with higher median incomes. However, the coefficient is small relative to the other variables so it may not be as relevant in the analysis. Returning to my earlier concern, this may have occurred because there was little variation in the household size data so there may not be a clear conclusion regarding the correlation. The Male variable, 1/Male, showed a negative correlation with median income. This disproves my earlier prediction most likely due to industries making a concerted effort to increase gender equality in the workplace. My median age variable showed a negative correlation to median income. This may indicate that younger populations, especially those who have high salaries, may be pulling the median income up when entering the work force. Older individuals may not have the education necessary to work in some high paying fields like tech. My technology variable was statistically significant for Tech, Tech^2, and log(Tech) so it was hard to analyze.

I move on to analyze functional form. All variables had monotonic graphs except for education and tech. Education showed a sigmoidal shape that ends higher than the beginning. The general shape moves upwards as education increases so I conclude that there is a positive correlation between education and median income. This would agree with my earlier assessment that higher educated populations may work higher paying jobs. The nonlinear shape may show that populations with too many college educated individuals results in a saturated job market for high paying positions, with some individuals having to resort to lower paying jobs. For my tech variable, I have a nonlinear graph that has a general increase in median income as tech increases. The nonlinearity may be explained by the census data. The data grouped tech jobs with other occupations like waste management and administrative services so the variable does not perfectly represent the tech sector. Still, the general positive correlation agrees with my earlier prediction that increases to jobs in the tech sector will raise median income.

V. Conclusion

Using A06SEF2, the regression equation has an adjusted R-squared value of 0.9635, which means the economic and demographic factors explain over 96% of the variation in median income. I like the equation for it's low Akaike, Schwarz, and HQ values and because it passes the heteroskedasticity tests.

All the variables are statistically significant, but the household variable has a small coefficient, hinting that it is not magnitudinally significant when predicting effects on median income. Still, the high R-squared and adjusted R-squared values make me confident that my regression factors in many of the variables related to median income.

In the future, I would like to create a more holistic model to track quality of life. One change would be using more specific factors for my independent variables. To better track the growth in the technology sector, I would prefer to use a variable that solely measures jobs in tech, possibly even specific areas like software engineering. I think it would give a better measurement of how tech has affected the quality of life in the Bay Area. I would also want to track how socio-economic stratification impacts quality of life. With the growth in high paying jobs pushing out individuals with lower paying positions, I think it would be interesting to see how that changes the median income or other measures of quality of life.

I would also like to find other measures of quality of life. Median income may be accurate for economic and demographic issues, but quality of life is a multifaceted issue and can be observed from many different perspectives. For example, I would like to see how general happiness has been affected. I chose to ignore this for my project because there is no strong dataset that measures happiness in the Bay Area, but I think it is important for understanding changes in quality of life.

Understanding what factors affect a region's quality of life is important for policymakers and individuals in positions of power to improve the community. Increasing median income may help the community's economy, but it can also negatively affect poorer individuals who cannot compete. The Bay Area is a unique region that has been strongly impacted by the growth of various technologies in the recent years. This topic will be relevant in understanding how to people have been affected by the tech boom and how we can remedy the issues in the economic and demographic changes.

VI. References

American Community Survey Office. “Data Tables & Tools.” *Data Tables & Tools | American Community Survey | U.S. Census Bureau*, 15 Jan. 2015 www.census.gov/acs/www/data/data-tables-and-tools/.

Corley, Mary & Kroll, Cynthia. (1995). The Bay Area Housing Market - Is it Ready for New Growth?.

“Nine Bay Area Counties Profile.” *San Francisco Center for Economic Development*, 22 Aug. 2015 sfced.org/wp-content/uploads/2016/10/Nine-Bay-Area-County-Profiles-Aug-2016Update.pdf

Sciacca, Annie. “In Costly Bay Area, Even Six-Figure Salaries Are Considered 'Low Income'.” The Mercury News, 25 Apr. 2017

www.mercurynews.com/2017/04/22/in-costly-bay-area-even-six-figure-salaries-are-considered-low-income/.

Soursourian, Matthew, 2012. "Suburbanization of poverty in the Bay Area," Community Development Research Brief, Federal Reserve Bank of San Francisco, issue January, pages 1-17.

Treuhaft, Sarah. “Solving the Housing Crisis Is Key to Inclusive Prosperity in the Bay Area.” *PolicyLink*, Apr. 2018

www.policylink.org/resources-tools/solving-housing-crisis-bay-area.

Appendix

Table 3: Final Regression Equation

Eq Name:	A06SEF2
Method:	STEPLS
Dep. Var:	LOG(MEDINCOME)
C	149.43346
DIVERSITY	
DIVERSITY^2	
1/DIVERSITY	-13.55642
LOG(DIVERSITY)	
EDUCATION	1.618031
EDUCATION^2	-0.007767
1/EDUCATION	-483.556
LOG(EDUCATION)	-51.02784
EMPLOYED	
EMPLOYED^2	0.000129
1/EMPLOYED	
LOG(EMPLOYED)	
HOUSEHOLD	
HOUSEHOLD^2	0.139031
1/HOUSEHOLD	
LOG(HOUSEHOLD)	
MALE	
MALE^2	
1/MALE	148.55343
LOG(MALE)	
MEDAGE	
MEDAGE^2	
1/MEDAGE	-31.09897
LOG(MEDAGE)	
POPULATION	
POPULATION^2	
1/POPULATION	48387.704
LOG(POPULATION)	0.126595
TECH	-0.657169
TECH^2	0.012025
1/TECH	
LOG(TECH)	4.329393
Obs	103
R-sq	0.9685
AdjR-sq	0.9635
Akaike	-4.0163
Schwarz:	-3.6326
HQ	-3.8609
S.E.reg	0.0304
MeanDep	11.272
DW	1.8348
F-Stat	193.1605
LRam1	0.2023
LRam2	NA
LRam3	NA
LRam4	NA
FRam1	0.171
FRam2	NA
FRam3	NA
FRam4	NA
BPG	19.8919
BP_SS	12.9242
Gleisjer	21.2449
GI_SS	20.1099
Harvey	15.6288
HarvSS	22.1814
White SQ	19.5054
Wh SQ SS	12.6732
Wh L&SQ	N/A
Wh L&SQ SS	N/A
BG 1 lag	0.5329
BG 2 lag	7.0601
BG 3 lag	7.6059
BG 4 lag	9.5023
MaxVIF	2084008

A06SEF2 Regression

Dependent Variable: LOG(MEDINCOME)

Method: Stepwise Regression

Date: 04/24/19 Time: 15:06

Sample: 1 103

Included observations: 103

Number of always included regressors: 1

Number of search regressors: 32

Selection method: Stepwise forwards

Stopping criterion: p-value forwards/backwards = 0.2/0.2

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
C	149.4335	55.47491	2.693713	0.0085
TECH	-0.657169	0.131511	-4.997061	0.0000
HOUSEHOLD^2	0.139031	0.006189	22.46265	0.0000
1/MEDAGE	-31.09897	6.555609	-4.743873	0.0000
EDUCATION	1.618031	0.481154	3.362815	0.0011
EMPLOYED^2	0.000129	1.92E-05	6.712233	0.0000
LOG(EDUCATION)	-51.02784	17.92388	-2.846919	0.0055
TECH^2	0.012025	0.002060	5.838431	0.0000
LOG(POPULATION)	0.126595	0.030520	4.147929	0.0001
1/MALE	148.5534	23.74251	6.256855	0.0000
EDUCATION^2	-0.007767	0.002112	-3.677068	0.0004
LOG(TECH)	4.329393	0.999326	4.332313	0.0000
1/EDUCATION	-483.5560	219.0467	-2.207548	0.0299
1/DIVERSITY	-13.55642	2.239013	-6.054643	0.0000
1/POPULATION	48387.70	11111.36	4.354796	0.0000

R-squared	0.968484	Mean dependent var	11.27195
Adjusted R-squared	0.963470	S.D. dependent var	0.158946

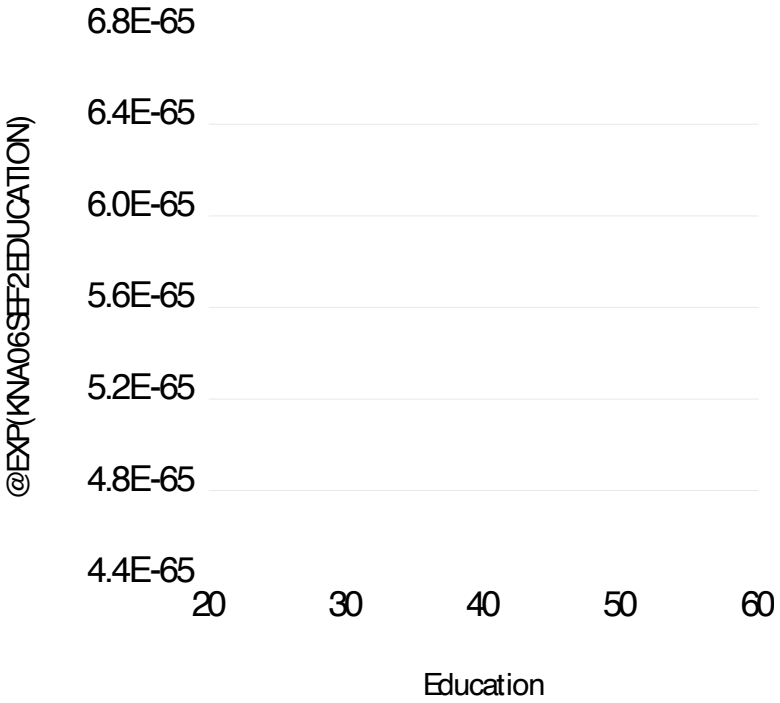
S.E. of regression	0.030379	Akaike info criterion	-4.016262
Sum squared resid	0.081214	Schwarz criterion	-3.632563
Log likelihood	221.8375	Hannan-Quinn criter.	-3.860851
F-statistic	193.1605	Durbin-Watson stat	1.834789
Prob(F-statistic)	0.000000		

Selection Summary

Added EDUCATION^2
Added HOUSEHOLD^2
Added 1/MEDAGE
Added EDUCATION
Added EMPLOYED^2
Added LOG(EDUCATION)
Added TECH^2
Added DIVERSITY
Added TECH
Removed EDUCATION^2
Added 1/MALE
Added EDUCATION^2
Added LOG(TECH)
Added 1/EDUCATION
Added 1/TECH
Added 1/POPULATION
Added LOG(POPULATION)
Added 1/DIVERSITY
Removed 1/TECH
Removed DIVERSITY

*Note: p-values and subsequent tests do not account for stepwise selection.

Slopes Program for Education Variable



Slopes Program for Tech Variable