Relationship between Income and Cost of Living in US Cities

Aamir Surani & John Michael Young

ECON 3161, Fall 2022

Georgia Institute of Technology

Dr. Shatakshee Dhongde

November 18, 2022

Abstract: This paper examines the relationship between the cost of living and median household income. The hypothesis is that there is a positive correlation between the two variables, where the median income is higher in areas with a higher cost of living. Variables such as water, electricity, energy, and rent were utilized in order to obtain a well-rounded cost of living factor and then compared to the median household income. Evidence suggests this hypothesis to a certain extent. While there is a positive correlation, the increases are not simultaneous.

I. Introduction:

The aim of this paper is to examine the correlation between the cost of living and the median household income by city in the United States. This topic is important because the United States has a history of fluctuating housing prices that respond to the state of the economy. From the 2007-2009 housing market crash to today's interest rates at an all time high, examining the correlation between household income and the cost of living in cities will give us a clearer understanding of how household incomes respond to changes in the cost of living.

The study of household income, which is a component of consumer spending power, can paint a better picture of how the economy reacts to changes in prices of goods. This study bears significance as current costs of living are at great heights. This is due to in large part the high inflation rates influencing costs of energy, groceries, and other goods and services. With the impact of COVID-19, the economy saw an influx of money through multiple stimulus checks. As the relative prices of goods have increased, consumers have needed more money in order to be able to afford basic necessities. Such a drastic change in the costs can have significant impacts on the overall economy.

Through our data analysis, we believe that the household incomes will depend on the cost of living in U.S. cities and that they will be closely and positively correlated. The median household income will be the dependent variable and the cost of living in U.S. cities will be the primary independent variable. As prices of goods are increasing, the income for households should also increase in order to keep real earnings constant. With higher prices of goods and services, ordinary consumers require higher incomes to support the same needs. For example, if city A has a cost of living index that is 10% higher than city B, the occupants of city A are less flexible to increased prices of goods, which would result in a higher wages in city A than city B.

II. Literature Review

Campbell Jr. (2021) conducted an empirical study on the impact of economic inequality of the cost of living in U.S. metropolitan areas. He found an association between higher costs of living and increasing economic inequality, particularly in the distribution of metropolitan income. Household poverty effects have significance but there was less consistency. He concluded that a reduction in economic inequality would result in shared benefits from decreased living costs by all metropolitan inhabitants. These benefits should be larger in areas that are rapidly growing with greater income disparities. Some of the data Campbell gathered was the highest and lowest cost of living in metropolitan areas in 2018 and inequality measures by metro area in 2015. He compiled his findings into a base model and inequality models, interaction models, and regional models. His base model revealed that the cost of living is somewhat inelastic with respect to changes in household income. Income growth is often faster than any changes in cost of living. His results suggested that area-wide cost-of-living would fall by 1.25-2.50 percent for each percentage point reduction in household poverty. Living costs that are related to wage growth in bigger cities like San Francisco, Boston, and Seattle are likely to be bundled by less mobile, lower income households, particularly if it is strong enough to raise the income per capita. He suggested that "rather than inducing top wage earners to leave, metropolitan areas can pursue development strategies aimed at building the wealth of those in the bottom half of the wage distribution by targeting and nurturing mid-level occupations and industries to fill production gaps in economic structure, lessen wage and income disparities while providing a pathway for lower and moderately skilled members of the workforce to enhance earnings and lower the incidence of poverty." (Campbell 2021).

(Bauer, Breitwieser, Nunn, Shambaugh 2018) collaborated on a paper which attempts to compare the significance of location on individual incomes. The study utilizes the cost of living index in various locations and the median annual earnings for those locations. Bauer et al. find that the actual median earnings change in relation to cost of living. The researchers utilized data from the BEA 2018 American Community survey which included the variables of median annual earnings, the cost of living index, and also tax rates in those regions. After plotting median annual income and the cost of living index, the data shows a positive correlation between the two variables. The findings suggest that for each $1,000 increase in the median earnings, the cost of living is about 1% greater on average. Therefore, a salary increase of $10,000 from $40,000 would correlate with a 10% greater cost of living index, signifying a 44% offset of the increased salary. The study also takes into account income taxes,

which are relatively similar across the regions observed. When adjusting the median earnings to cost of living and tax rates, the actual earnings are still on a positive relationship. However, this relation is not one-for-one - meaning that the areas with the higher median annual earnings are in fact higher in actual salary accounting for cost of living index as well as taxes, but there is some erosion of the earnings.

Handbury (2021) shows through her paper that products and prices offered in markets are correlated with local income-specific tastes. She calculates the local price indexes micro-founded by a model of non-homothetic demand over thousands of grocery products to quantify the welfare impact of this variation. The indexes revealed large differences in how wealthy and poor households perceive the choice sets available in wealthy and poor cities. Relative to low-income households, high-income households enjoy 40 percent higher utility per dollar expenditure in wealthy cities, relative to poor cities. Similar patterns were observed across stores in different neighborhoods. Most of the variation is explained by differences in product assortment offered, instead of the relative prices charged, by chains that operate in different markets. She found that stores favor high-income consumers more in wealthy locations than in poor ones through both product offerings and pricing. The differences in availability and pricing are shown to matter for consumers through spatial price indexes that reveal large differences in how high and low-income households perceive the prices and variety available in different U.S. cities. She also shows how the differences in relative grocery costs across cities are driven more by cross-city variation in product variety than by variation in prices. Handbury found that higher-income households face relatively lower price indexes in stores located in higher-income neighborhoods. Through her paper, Handbury contributes the first direct evidence of income-specific taste for local consumption amenities. This aids the hypothesis that tastes explain spatial disparities in income and skill observed across U.S. cities where high-skill, high-income workers co-locate because they enjoy more utility from certain endogenous local amenities than low-skill, low-income consumers. Through the data it is shown that high-income households face much lower grocery costs in wealthy cities, than in poor cities, while low-income households face slightly higher grocery costs in these locations.

This paper will follow up the existing literature and further develop it by evaluating income and cost of living in depth. In doing so, the research will provide a data supported conclusion on whether or not higher incomes lead to higher costs of living. Through our research, we are able to provide recommendations for government and corporation policy makers. Local government officials will be able to utilize this data to inform policy decisions in their city such as minimum wage and tax subsidization which eases the financial burdens of the city's occupants.

III. Data

Other variables that play a role in determining the cost of living are rent and utility costs. The utilities cost defines the amount that a household spends on electricity, energy, and water. The rent cost is how much a household spends on rent annually. The year for the control variables is 2017 in order to avoid significant inconsistencies in the data that result from the COVID-19 pandemic. The control variables include the number of weeks worked per year, household size, and education level. These variables are all utilized as controls to help determine an accurate representation of average household income.

The year for the control variables is 2017 in order to avoid significant inconsistencies in the data that result from the COVID-19 pandemic. The source of the data is from IPUMS USA. This data source is based on the Current Population Survey which gathers data from households across various cities in the United States. In order to get the most effective data and keep the dataset manageable, the data filters out observations that do not have a city in order to relate the cost values to income reliably by city.

Table 1: Description of Variables

Variable	Description	Units	Year	Source
loghhincome	Log of Annual Household Income	U.S. Dollars	2017	IPUMS USA
logcostliving	Log of Annual Cost of Living (Annual Rent + Utilities)	U.S. Dollars	2017	IPUMS USA
wkswork	Weeks Worked Annually	Weeks Intervals	2017	IPUMS USA
famsize	Household Family Size	Persons	2017	IPUMS USA
educ	Years/Grades of Education Completed*	Grade level completed[1]	2017	IPUMS USA

There are 5 variables being considered in the analysis. They include 1 dependent variable and 4 independent variables. The dependent variable of the analysis is log*hhincome*, which is the log annual total household income. The primary independent variable is *logcostliving* which is the log of annual cost of living including annual rent, cost of water, electricity and gas. *Famsize* is the number of persons in

[1] Grouped according to grade levels. Up to grade level 4 = 01, Grades 5-8 = 02. Increases by 1 for each grade level completed afterwards (including college) up to 4 years of college. 5+ years of college = 11.

each household. *Wkswork* is a measure of how many weeks worked annually. Each number 1-6 is an interval of 13 weeks. *Educ* is a categorical and numerical measure of years of schooling completed. For *educ*, 0 means n/a or no schooling, 1 includes nursery school to grade 4, 2, includes middle school, 3-6 are freshman through senior year of highschool, and 7-11 includes 1 to 5+ years of college.

Table 2: Summary Statistics

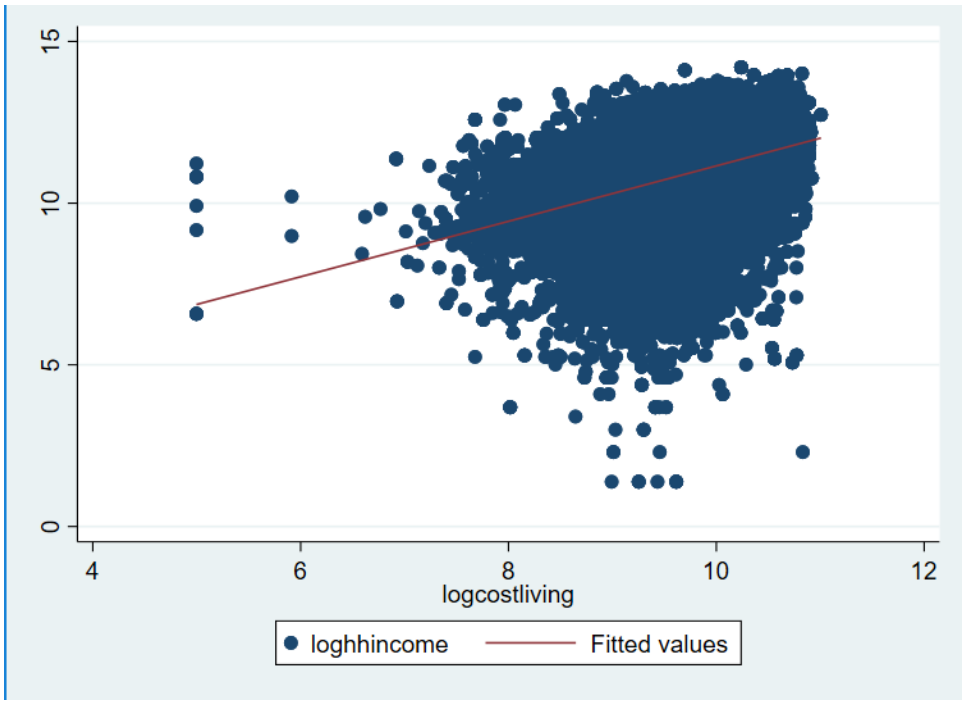
Variable	Obs	Mean	Std. Dev.	Min	Max
loghhincome	178,047	10.81101	.9343301	1.386294	14.20077
logcostliving	178,047	9.600467	.5013683	4.997212	11.00642
wkswork	178,047	2.797638	2.821121	0	6
famsize	178,047	3.47984	1.92072	1	16
educ	178,047	5.470893	3.284108	0	11

The variables, measured annually, produced relatively expected results in regards to the values generated. As seen in Table 2 above, all variables in the data consist of 178,047 observations, which are all included from the original dataset. The variables loghhincome and logcostliving are the logged results of the original variables - hhincome and costliving. Both of those variables are in US Dollars in the dataset.

Table 3: Correlation of Regressors

Variables	logcostliving	wkswork	famsize	educ
logcostliving	1.000			
wkswork	0.0694	1.0000		
famsize	0.1556	-.02587	1.0000	
educ	0.1307	0.5700	-0.3843	1.0000

Figure 1 Scatter Plot of loghhincome & logcostliving



Before running the regression analysis, the data was examined to ensure it met all Classical Linear Model (CLM) Assumptions:

1. **The regression model is linear** with respect to its coefficients and error term.

Our model satisfies the first Gauss Markov Assumption of linearity since all coefficients in the model are constants to be multiplied with an explanatory variable. (logcostlivingβ_1+ wksworkβ_2 + famsizeβ_3 + educβ_4 + u)

2. **Random Sampling**

Our data was collected by the Current Population Survey, so we can assume that it meets the condition of being randomly sampled from the population.

3. **Non-Collinearity**

As shown in Table 3, none of our explanatory variables are perfectly correlated to another, thus our data meets the third assumption of non-collinearity.

4. **Zero Conditional Mean**

This means that the expected value of the error term, u will be 0 for any independent variable values. There are likely other factors not included in the dataset impacting the correlation between income and cost of living. As a result, we will assume that this is true, even though it is a difficult assumption to make.

5. **Homoscedasticity**

This assumption is that our error term has a constant variance. While there may be outliers for the data which impact the error term, we will assume that it has a constant variance.

III. Results

Simple Regression Model:

Model 1: loghhincome = $\beta_0 + \beta_1$(logcostliving) + u

Based on the STATA output from Appendix B, the estimated equation for this model is:

loghhincome = 2.58 + 0.857(logcostliving) + u

N = 178,047 R^2 = 0.21

In this simple regression model, the log of cost of living is the only independent variable being tested against the main dependent variable of the log of annual median household income. The R^2 value of 0.21 is not that high, so it is hoped that adding other control variables might yield a higher value. The t-statistic is quite high at 218.55 so this shows that *logcostliving* is significant even at the 1% level. The R^2 value reveals that up to 21% of variance in median household income can be explained by the cost of living. This makes sense upon interpretation because there are many other factors that can affect household income, such as size of the household, how many weeks are spent working out of the year, and education level of the household. These factors will be explored in the multiple regression models to help understand further the effect of cost of living on household income. The table below summarizes the results of this model.

Table IV: Model 1 Estimation Results

Dependent Variable: loghhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
logcostliving	.857	.004	218.55

Multiple Regression Models:

Model 2: loghhincome = $\beta_0 + \beta_1$(logcostliving) + β_2(wkswork) + β_3(famsize) + β_4(educ) + u

Based on the STATA output from Appendix C, the estimated equation for this model is:

loghhincome = 2.82 + 0.781(logcostliving) + 0.079(wkswork) + 0.059(famsize) + 0.012(educ) + u

n = 178,047 R^2 = 0.28

The first multiple regression model includes all 4 independent variables in the analysis to provide a basis for further model construction. The R^2 of 0.28 is the highest value out of all the regressions in the study and proves to be the best model for predicting median household income in the study. This makes sense because the other variables added all help explain household income. However, because the best model only explains up to 28% of the variance in household income, there must be other explanatory factors not included in this study that affect household income on a more substantial level. All variables had high t-statistic values: *logcostliving* was 200.72, *wkswork* was 97.14, *famsize* was 53.81, *educ* was 16.25, meaning that these are all significant at the 1% level. The primary independent variable *logcostliving* maintained the highest coefficient as predicted. Its coefficient did drop by 0.07, but this is an acceptable amount and makes sense as the other variables also help explain *loghhincome*. As we move to the next model, the primary independent variable *logcostliving* will be dropped to examine how strongly the control variables explain the dependent variable on their own.

Table V: Model 2 Estimation Results

Dependent Variable: loghhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
logcostliving	0.781	0.004	200.72
wkswork	0.079	0.001	97.14
famsize	0.059	0.001	53.81
educ	0.012	0.001	16.25

Model 3: loghhincome = $\beta_0 + \beta_1$(wkswork) + β_2(famsize) + β_3(educ) + u

Based on the STATA output from Appendix D, the estimated equation for this model is:

loghhincome = 10.00 + 0.080(wkswork) + 0.108(famsize) + 0.038(educ) + u

n = 178,047 R^2 = 0.11

In this multiple regression model, the primary independent variable *logcostliving* was removed to see how well the other control variables explained the variance in household income. The R^2 value dropped by more than half, revealing that *logcostliving* does indeed play a big role in this model for explaining variance in household income. This multiple regression model only explains up to 11% of the variance in household income up to the 1% level. The t-statistics for *famsize* and *educ* both increased, with *educ*'s value increasing by almost 3 times, while *famsize* increased by almost 2 times. Education level would be thought to have a large effect on household income which is shown in this model, but for some reason it lost its significance when the cost of living is incorporated. Interestingly, t-statistic for *wkswork* decreased slightly. Because of the significant increase in education's coefficient and t-statistic and the decrease in the relevance of weeks worked, we will drop *wkswork* in our final model. Overall, this model provides evidence of the explanatory power of cost of living on household income.

Table VI: Model 3 Estimation Results

Dependent Variable: loghhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
wkswork	0.080	0.001	88.74
famsize	0.108	0.001	91.68
educ	0.038	0.001	47.25

Model 4: loghhincome = $\beta_0 + \beta_1$(famsize) + β_2(educ) + u

Based on the STATA output from Appendix E, the estimated equation for this model is:

loghhincome = 10.04 + 0.103(famsize) + 0.076(educ) + u

n = 178,047 R^2 = 0.07

In this regression model, *logcostliving* and *wkswork* were both removed to test the effect that family size and education levels had on household income. As shown in Table VII, education level becomes significantly more relevant to the model as its coefficient is 6 times higher than in the original MLR (Model 2) with all the variables accounted for. The R^2 value, however, is the lowest among the MLRs at 0.07. This tells us that this model only explains about 7% of the variance in household income, signifying that there are other factors that must be included to gain a better understanding of household income. The t-statistic values for both variables are higher in this model when compared to the first MLR model, with the t-statistic for *educ* having increased by over 6 times as well. While the *famsize* variable is higher in coefficient and t-statistic values than Model 2, removing *wkswork* decreased the values compared to Model 3. In Model 5, we will take a look at how big of an impact education level and cost of living have together on household income.

Table VII: Model 4 Estimation Results

Dependent Variable: loghhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
famsize	0.103	0.001	85.29
educ	0.076	0.001	108.38

Model 5: loghhincome = $\beta_0 + \beta_1$(logcostliving) + β_2(educ) + u

Based on the STATA output from Appendix F, the estimated equation for this model is:

loghhincome = 2.68 + 0.826(logcostliving) + 0.037(educ) + u

n = 178,047 R^2 = 0.23

In this regression model, we add in our primary independent variable, *logcostliving* and remove all others except for *educ* to examine how household income is explained by just these two variables. Based on the results in table VIII, we can see that the *logcostliving* variable has significant effects and reduces the relevance of education levels. Both variable's coefficients and t-statistics are higher than Model 1, as expected. The model also better explains the variance in household income with an R^2 of 23, which is slightly lower than Model 2 with all the variables included and slightly higher than Model 1 with just the primary independent variable. This shows that these two variables bear significant importance to the study.

Table VIII: Model 5 Estimation Results

Dependent Variable: loghhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
logcostliving	0.826	0.004	210.93
educ	0.037	0.001	61.40

Table IX: Estimation Results Summary

Dependent Variable: loghhincome					
Independent Variable:	SLR	MLR1	MLR2	MLR3	MLR4
logcostliving	0.857*** (0.004)	0.781*** (0.004)			0.826*** (0.004)
wkswork		0.079*** (0.001)	0.080*** (0.001)		
famsize		0.059*** (0.001)	0.108*** (0.001)	0.103*** (0.001)	
educ		0.012*** (0.036)	0.038*** (0.001)	0.076*** (0.001)	0.037*** (0.001)
No. of obs.	178,047	178,047	178,047	178,047	178,047
R^2	0.21	0.28	0.11	0.07	0.22

Extensions:

From the correlation table, the highest correlation value was 0.57 between *wkswork* and *educ*. This may be because those with higher levels of education are more likely to have full-time jobs. Although this is not an alarmingly high value, it is worth testing for joint significance through an F-test. The null hypothesis is the following:

$H_0: \beta_2 = \beta_4 = 0$ $H_A: \beta_2 \neq \beta_4 \neq 0$

The SSR value for the unrestricted model from the STATA output in Appendix C is 112541.135 and the SSR value for the restricted model from Appendix H is 122341.968. These values are used in the f-value calculation below:

$$F = \frac{(SSRr - SSRur)\ /\ q}{SSRur\ /\ (n-k-1)} = \frac{(122341.968 - 112541.135)\ /\ 2}{112541.135\ /\ (178{,}044)} = 0.86$$

The degrees of freedom in the model are $df_1 = 2$ and $df_2 = 178{,}044$, and at 10% level of significance the critical value is 2.30. Because the f-value calculated does not fall in the rejection region, we fail to reject the null hypothesis and conclude that *wkswork* and *educ* are not jointly significant.

Because the value is not within the rejection region at 10% level of significance, it will also not be in the rejection region at 5% and 1% levels of significance.

Because our dependent and primary independent variables *hhincome* and *costliving* are measured in dollars, we used the log values of them to minimize the difference in absolute values. To show the difficulties of interpreting the regression without the log values, we created a model with no log estimations. The result is that the standard error becomes very high for the control variables and the primary independent variable *costliving* looks as if it has a much smaller effect on *hhincome*. This STATA output is found in Appendix I. Overall, taking the log values of *hhincome and costliving* allowed the interpretation of the model to become clearer, without the confusion of large differences in coefficient values as seen below.

Table X: Model NoLog Estimation Results

Dependent Variable: hhincome			
Independent Variable:	Coefficient	Standard Error	t-statistic
costliving	3.68	0.017	215.03
wkswork	2990.24	61.21	48.85
famsize	2332.79	81.60	28.59
educ	1468.98	55.97	26.24

n = 178,047 , R^2 = 0.25

90% & 99% Confidence Interval for and T-Test for β_1

H_0: $\beta_1 = 0$ H_A: $\beta_1 \neq 0$ $t\beta_1 = 200.72$ $\beta_1 = 0.781$ s.e. $\beta_1 = 0.004$

Critical Value at 10%: 1.65 Critical Value at 1%: 2.58

|200.72| > |1.65| & |200.72| > |2.58|

90% Confidence Interval: 0.781 ± 1.65(0.004) = (0.774, 0.788)

99% Confidence Interval: 0.781 ± 2.58(0.004) = (0.771, 0.791)

Conducting a t-test of our primary independent variable determines whether the difference in the means of *logcostliving* are significant. Because the t-value of 200.72 is incredibly high, it is immediately clear that the results are significant and that the null hypothesis is rejected up to the 1% level. Constructing a confidence interval for a primary independent variable *logcostliving* allows us to define which interval will contain our population mean. Using the critical values at 10% and 1% yield the 90% and 99% confidence intervals for our primary independent's mean shown above. It makes sense that the interval for the 99% confidence level would be bigger than the 90% because we can be more certain that the mean will fall within a larger range than we would in a smaller interval.

V. Conclusions

Throughout our analysis of household income and cost of living, we can see that there is a slight positive correlation among the two variables of interest. This supports our original hypothesis stating that household income will be correlated with the cost of living of the areas of the households. The cost of living by itself, however, does not accurately predict the household income levels completely. Therefore, other variables such as education level, weeks worked in a year, and family size are added in to have a more complete model of analysis.

The cost of living for households, which takes into account rent and basic utilities such as gas, water, and electricity, is consistently a large component of the accuracy of the model. The other variables (education, weeks worked, and family size) are unable to have the same significance of cost of living on household income. As seen in Model 2, the model is most accurate when all of the aforementioned variables are included. The observations of these variables display a positive coefficient, while the cost of living remains significant throughout the models.

The observations conducted in our research support the idea that cost of living is significant to a small extent in determining the income levels of a household. The research suggests that in areas with a higher cost of living, incomes will likely be higher for the individuals from the respective area. Overall, as a result of the research conducted, we can infer that there may be more significant factors that influence income level of households. The observations stated in this paper may be of use to persons hoping to evaluate a starting salary for a job, or even to employers hoping to properly adjust the salaries of their employees.

References

Bauer, L., Breitwieser, A., Nunn, R., & Shambaugh, J. (2018, July). Where Work Pays- How Does Where You Live Matter for Your Earnings? Retrieved October 13, 2022, from https://www.brookings.edu/wp-content/uploads/2018/07/ES_THP_071018_where_work_pays_tech_appendix.pdf

Campbell, Harrison S. 2021. Income and cost of living: Are less equal places more costly? *Social Science Quarterly*. 102: 2689– 2705. https://doi.org/10.1111/ssqu.13017

Handbury, J. (2021), Are Poor Cities Cheap for Everyone? Non-Homotheticity and the Cost of Living Across U.S. Cities. Econometrica, 89: 2679-2715. https://doi.org/10.3982/ECTA11738

Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 12.0 [dataset]. Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D010.V12.0

Appendix:

Appendix A: Data Summary

```
. sum loghhincome logcostliving wkswork2 famsize educ
```

Variable	Obs	Mean	Std. dev.	Min	Max
loghhincome	178,047	10.81101	.9343301	1.386294	14.20077
logcostliv~g	178,047	9.600467	.5013683	4.997212	11.00642
wkswork2	178,047	2.797638	2.821121	0	6
famsize	178,047	3.47984	1.92072	1	16
educ	178,047	5.470893	3.284108	0	11

Appendix B: Simple Linear Regression STATA Output

```
. reg loghhincome logcostliving
```

Source	SS	df	MS		
				Number of obs	= 178,047
				F(1, 178045)	= 47763.79
Model	32876.8965	1	32876.8965	Prob > F	= 0.0000
Residual	122552.403	178,045	.688322631	R-squared	= 0.2115
				Adj R-squared	= 0.2115
Total	155429.299	178,046	.872972712	Root MSE	= .82965

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
logcostliving	.8570821	.0039217	218.55	0.000	.8493957	.8647685
_cons	2.582623	.0377013	68.50	0.000	2.50873	2.656517

Appendix C: Multiple Linear Regression #1 (Model 2) STATA Output

```
. reg loghhincome logcostliving wkswork2 famsize educ
```

Source	SS	df	MS	Number of obs	=	178,047
				F(4, 178042)	=	16962.45
Model	42888.1644	4	10722.0411	Prob > F	=	0.0000
Residual	112541.135	178,042	.63210442	R-squared	=	0.2759
				Adj R-squared	=	0.2759
Total	155429.299	178,046	.872972712	Root MSE	=	.79505

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
logcostliving	.7808661	.0038903	200.72	0.000	.7732413	.788491
wkswork2	.0790701	.000814	97.14	0.000	.0774747	.0806655
famsize	.0587566	.001092	53.81	0.000	.0566164	.0608969
educ	.0120782	.0007431	16.25	0.000	.0106217	.0135347
_cons	2.822581	.0363025	77.75	0.000	2.751429	2.893733

Appendix D: Multiple Linear Regression #2 (Model 3) STATA Output

```
. reg loghhincome wkswork2 famsize educ
```

Source	SS	df	MS	Number of obs	=	178,047
				F(3, 178043)	=	7491.54
Model	17420.9906	3	5806.99686	Prob > F	=	0.0000
Residual	138008.309	178,043	.775140325	R-squared	=	0.1121
				Adj R-squared	=	0.1121
Total	155429.299	178,046	.872972712	Root MSE	=	.88042

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
wkswork2	.0799901	.0009014	88.74	0.000	.0782235	.0817568
famsize	.1080272	.0011783	91.68	0.000	.1057179	.1103366
educ	.03828	.0008101	47.25	0.000	.0366922	.0398678
_cons	10.00189	.0068774	1454.31	0.000	9.988406	10.01536

Appendix E: Multiple Linear Regression #3 (Model 4) STATA Output

```
. reg loghhincome famsize educ
```

Source	SS	df	MS		
Model	11316.4096	2	5658.20481	Number of obs	= 178,047
Residual	144112.89	178,044	.809422894	F(2, 178044)	= 6990.42
				Prob > F	= 0.0000
				R-squared	= 0.0728
				Adj R-squared	= 0.0728
Total	155429.299	178,046	.872972712	Root MSE	= .89968

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
famsize	.102555	.0012024	85.29	0.000	.1001984	.1049117
educ	.0762139	.0007032	108.38	0.000	.0748356	.0775922
_cons	10.03718	.0070161	1430.59	0.000	10.02343	10.05093

Appendix F: Multiple Linear Regression #4 (Model 5) STATA Output

```
. reg loghhincome logcostliving edu
```

Source	SS	df	MS		
Model	35418.1843	2	17709.0921	Number of obs	= 178,047
Residual	120011.115	178,044	.674053128	F(2, 178044)	= 26272.55
				Prob > F	= 0.0000
				R-squared	= 0.2279
				Adj R-squared	= 0.2279
Total	155429.299	178,046	.872972712	Root MSE	= .82101

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
logcostliving	.8256709	.0039144	210.93	0.000	.8179987	.833343
educ	.0366931	.0005976	61.40	0.000	.0355218	.0378644
_cons	2.683442	.0373446	71.86	0.000	2.610248	2.756637

Appendix G: Correlation Table for Independent Variables

```
. corr logcostliving wkswork2 famsize educ
(obs=178,047)
```

	logcos~g	wkswork2	famsize	educ
logcostliv~g	1.0000			
wkswork2	0.0694	1.0000		
famsize	0.1556	-0.2587	1.0000	
educ	0.1307	0.5700	-0.3843	1.0000

Appendix H:

```
. reg loghhincome logcostliving famsize
```

Source	SS	df	MS		
Model	33087.3315	2	16543.6658	Number of obs	= 178,047
Residual	122341.968	178,044	.68714457	F(2, 178044)	= 24075.96
				Prob > F	= 0.0000
				R-squared	= 0.2129
				Adj R-squared	= 0.2129
Total	155429.299	178,046	.872972712	Root MSE	= .82894

loghhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
logcostliving	.8462844	.0039666	213.35	0.000	.8385099	.8540589
famsize	.0181196	.0010354	17.50	0.000	.0160902	.0201489
_cons	2.623233	.0377405	69.51	0.000	2.549263	2.697204

Appendix I: Model "NoLog", A Version of Model 2 Without Log Variables, STATA Output

```
. reg hhincome costliving wkswork2 famsize educ
```

Source	SS	df	MS	Number of obs	=	178,047
				F(4, 178042)	=	15189.52
Model	2.1715e+14	4	5.4287e+13	Prob > F	=	0.0000
Residual	6.3632e+14	178,042	3.5740e+09	R-squared	=	0.2544
				Adj R-squared	=	0.2544
Total	8.5346e+14	178,046	4.7935e+09	Root MSE	=	59783

hhincome	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
costliving	3.676764	.0170985	215.03	0.000	3.643251	3.710276
wkswork2	2990.236	61.21287	48.85	0.000	2870.26	3110.211
famsize	2332.793	81.60262	28.59	0.000	2172.854	2492.732
educ	1468.977	55.9738	26.24	0.000	1359.27	1578.684
_cons	-14639.52	499.0216	-29.34	0.000	-15617.59	-13661.45