Project -- Second Deliverable

Introduction

After reviewing the comments of first deliverable, we learned several things and fixed these problems in the second deliverable. First of all, we did not give sufficient and thorough introductions of the database's background, which made readers have difficulties understanding our analysis based on the data. Second, we gave too many unnecessary details, such as data names in database, meanings of data values, which were confusing because readers cannot see outputs from Stata, therefore they do not know what we were referring to. So, in this deliverable, we will pay more attention to clarify each variable's representation and relations between dependent variable and independent variables. Moreover, one label in our table was misleading. "family member number" was supposed to represent family size, the amount of people in each household, but readers may interpret that in different ways. We also concluded that, based on small t statistics and large p values, a control variable, employment status does not have much explanatory power with relation to the dependent variable. So in the second deliverable, we will replace it by region, which affects the cost of water much more significantly.

William Even
title should indicate content [...]

William Even
this is unnecessary. [...]

William Even
you have said what the dependent [...]

Regression Table:

Table 1

DETERMINANTS OF WATER COST AT HOUSEHOLD LEVEL

	Model 1 Simple linear regression model		Model 2 Log regression model		Model 3 Quadratic regression model		Model 4 Robust error regression model	
	Determinant effect	*t*-statistic	Determinant effect	*t*-statistic	Determinant effect	*t*-statistic	Determinant effect	T-statistic
Household income, thousand $	0.216	30.86*	0.0002	18.15*	0.0006	50.11*	0.216	22.51
Family size	51.434	119.90*	0.930	99.30*	54.470	128.84*	51.434	105.82
Farm	-101.598	-13.47*	-0.275	-16.70*	-87.217	-11.52*	-101.509	-13.61
Value of house, thousand $	0.176	100.85*	0.0002	61.06*	0.0000	60.05*	0.176	0.003
Region								-
New England	0	-	0	-	0	-	0	
Middle Atlantic	35.725	9.76	0.0282	3.53	25.047	6.82	35.725	10.12
E. N. central	26.508	7.44	-0.263	-3.39	-4.973	-1.39	26.508	7.88
W.N. central	-3.925	-1.01	-0.136	-16.06	-36.627	-9.40	-3.925	-1.08
S. Atlantic	-12.257	-3.56	-0.170	-22.66	35.225	-10.20	-12.257	-3.72
E.S. central	-71.73	-18.64	-0.305	-36.31	-106.027	-27.52	-71.730	-20.12
W.S. central	44.418	12.45	-0.111	-14.35	11.420	3.20	44.418	12.73
Mountain	44.577	11.62	-0.086	-10.27	24.784	6.44	44.577	11.90
Pacific	155.594	43.90	0.076	9.78	162.263	45.58	155.594	42.92
Constant	384.775	112.88	5.673	762.16	445.054	3.373	384.775	115.43
R-squared	0.0645		0.0359		0.547		0.0646	
Adjusted R-squared	0.0646		0.0359		0.547			-
Breusch-Pagan test	2635.22*		603.28*		2144.64*			-
White test	13492.05*		509.50*		10337.06*			-

William Even
quadratic in x must include both x a [...]

William Even
?? [...]

William Even
t–statistics should be in (.) below [...]

Adjusted R-squared	0.0646	0.0359	0.547
Breusch-Pagan test	2635.22*	603.28*	2144.64*
White test	13492.05*	509.50*	10337.06*

Corr (exp((cost water), cost water)=0.2305

* $p < 0.05$

Note: The sample is based on American Community Survey Census, with participants. The sample includes only the data for 2016, and the water cost is investigated at household level. The sample included 659,503 participants.

William Even

need to square this for comparison [...]

Discussion of results from new regression analysis

a. different specification considered

Besides the linear regression model, we generated two more alternative specifications, log regression and quadratic regression, and determined the preferred one based on a comparison of their R squares. Because all three regressions have exactly same amount of control variables on the right side of regressions, comparing R square is as unbiased as comparing the adjusted R square.

To estimate what elements affect the dependent variable, annual water cost of households measured in dollars, both Linear and log regressions have control variables: total income each household earns measured in thousands of dollars, family size, whether household is located at farm or not, value of house measured in thousands of dollars, and region of household. The quadratic regression, however, has the square of total income and house value instead of their original first order terms.

William Even

you need to include the income as [...]

After running regression models in Stata, we got a R square and an adjusted R square for all three regressions. To determine the preferred one between linear and log regressions, however, it's necessary to transform logged dependent variable to unlogged dependent variable first (generating the squared correlation between annual water cost and estimated annual water cost). An R-square comparison is meaningful only if the dependent variable is the same for both models. For log model, the R-square measures the amount of variation in ln(watercost), but not true variation in cost of water.

The squared correlation between annual water cost and estimated annual water cost equals to (0.2305)^2, which is 0.053. The R square of linear regression is 0.0646. So, independent variables in linear regression model explain higher proportion of the variation in the dependent variable and fits observations better. Then we compared the R square of linear regression

with R square of quadratic regression, which is 0.0547. The linear regression is still better. We finally chose linear regression as the best fitted model.

b. test statistics to determine between specifications

After testing the heteroskedasticity by both Breusch-Pagan and White tests, we noticed that all three regessions are statistically significant. That means all three regressions have heteroskedaticity. To solve that probelm, we generated robust standard deviation for each of them. The exitsence of heteroskedasticity causes the varaince of residual varies with changes of control variables' values. For example, in the Breusch-Pagan test of linear regression model, coefficient on control varaiable household income is 212.7275 and it is statistically significant. That means, household income causes the variance of residual to be higher; every 1000 more income rises the variance of residual by 212.7275. Negative and statistically significant coefficeints cause the variance of residual lower. Also, the variance of residual of households located in new england division is lower than households in other regions becuase the cofficient on new england is zero and others are all positive.

Statical significant dose not mean economically significant. According to the robust linear regression, the coefficent between hosehold income and annual water cost is 0.216, with t statistics of 22.51 and P value of 0.000.

William Even

you didn't estimate a quadratic. it [...]

William Even

what's this mean? [...]

William Even

errors for all estimated coefficients [...]

annual water cost is 0.216, with t statistics of 22.51 and P value of 0.000. Such statistics indicate that household income have effcet on annual water cost. However, it is not economically signifcant because every 1000 dollars increase in income just incrased the annual water cost by 0.216 dollar. Similary, effecr of variable house value is statistically significant, but it's not economically significant becuase 1000 dollars increase in house value just increase the annual water cost by 0.18 dollar.

Discussion of whether the key control variable and at least two others fit your prior expectations

We expected the key control variable, household income has positive effect on annual water costs of water. Keeping all other elements constant, the more people earn, the more they would like to spend to improve their standard of living qualities. Family size were supposed to be positively related to annual water cost, too. Because more people means more requirements of water. Last, we assumed that the coefficient on households in west south central division is positive, which means households in that region spend more on water than people in the new England division. This is because of the long-lasting warm weather there and people need to drink and wash more. Our preferred model's outputs match our expectations. Also, all effects are statistically significant at the 0.05 level except the effect of west north central division, which means that there are no big differences in annual cost of water between households in new England division and west north central division.

William Even

why? [...]

related to annual water cost, too. Because more people means more requirements of water. Last, we assumed that the coefficient on households in west south central division is positive, which means households in that region spend more on water than people in the new England division. This is because of the long-lasting warm weather there and people need to drink and wash more. Our preferred model's outputs match our expectations. Also, all effects are statistically significant at the 0.05 level except the effect of west north central division, which means that there are no big differences in annual cost of water between households in new England division and west north central division.

William Even

why? [...]

William Even

perhaps water is more scarce there [...]