Grocery expenditure Data Analysis

Jung-a Kim

Jiali Chen

Yuting Chen

Dong Lin

Abstract

Our team searched for demographic characteristics and consumption patterns of households in U.S. that contribute to its weekly grocery expenditure. We started with 43 predictors and 500 randomly selected observations from the original dataset for this project. At first, we compared the grocery expenditures of households where a reference person[1] is not married and those where a reference person is married using 2 independent sample t-test. Then we used three-way ANOVA to test main effects and interaction effects of marital status and other categorical variables. Subsequently, we selected 8 variables out of the 43 predictors using backward elimination. With the selected variables, we conducted a multiple regression analysis. In addition, we performed another multiple regression with the significant variables from the ANOVA results. The 43 predictors were reduced down to only a handful number in our final models with the correlation of determination around 0.3.

[1] A reference person is the person or one of the persons who owns or rents their house

Introduction

As grocery shopping is a daily part of people's lives, we were curious how much money people spend weekly on groceries on average and which factors contribute to their grocery expenditure. Our team asked ourselves the following questions to investigate those factors.

1. Does a reference person's marital status affect the household's weekly grocery expenditure?
2. If it does, what other factors could affect the weekly grocery expenditure?
3. What information do we need to predict the average weekly grocery expenditure?
4. Could one type of factor depend on another?

Method of data collection

Our team selected Diary survey data from Consumer Expenditure Survey(CE) conducted by the U.S. Bureau of Labor Statistics(BLS) in 2016. Diary survey data consists of daily grocery expenditures and demographic data of households in US.

Diary survey is defined as:

> The daily expense record self-reported by the sample consumer units[2](CUs) for two consecutive 1-week periods. The sample is surveyed across a 12-month period. Data collected each week are treated as statistically independent – each week's diary is separately weighted to be representative of the sample. The diary is divided by day of purchase and by four classifications of goods and services—food away from home, food at home, clothing, and all other goods and services—a breakdown designed to aid the respondent in recording the entire consumer unit's daily purchases. ("Diary survey details", Feb. 25, 2016, para. 3)

The Bureau of Labor Statistics states that:

> The weights have been adjusted so that the sum of all CU weights for one month approximates one third of the U.S. population. Consequently, the weights for three months of data approximate the total U.S. population. ("2016 Consumer Expenditure Surveys Microdata Diary Documentation", Aug. 29, 2017, pp.11)

We chose our sample data that was recorded in the 1st quarter of the year 2016[3]. We randomly selected 500 CUs

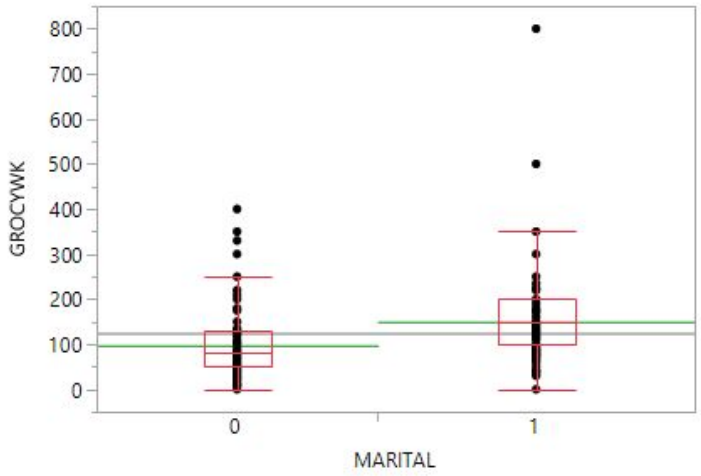
[2] A Consumer Unit(CU) is defined as 1) all members of a respondent's housing unit who are related by blood, marriage, adoption, or some other legal arrangement, such as foster children; 2) a respondent living alone or sharing a household with others, or living as a roomer in a private home, lodging house, or in permanent living quarters in a hotel or motel, but who is financially independent; or 3) two or more unrelated persons living together who pool their income to make joint expenditure decisions.

[3] \DIARY16\FMLD161 (Diary FMLD file for first quarter, 2016)

among 2951 CUs in the original sample.

Variables of interest

MARITAL	Marital status of reference person CODED 0: Not married 1: Married
GROCYWK	Weekly expenditure for grocery store purchases (in USD)
AGE_REF	Age of reference person
FAM_SIZE	Number of members in CU
HIGH_EDU	Highest level of education within the CU CODED 0: 8th Grade or lower 1: 9th-12th Grade (no high school diploma) 2: HS Graduate 3: Some college, no degree 4: AA degree 5: Bachelors degree 6: Masters degree, professional/doctorate degree
BEEF	Weekly expenditure on buying beef products
FRSHFRUT	Weekly expenditure on buying fresh fruits. Fresh fruits such as apples, oranges, bananas that are displayed on stacks or fridge (Processed fruits are ones that are canned or cupped.)
FRSHVEG	Weekly expenditure on buying fresh vegetables (Processed vegetables are usually seasoned and canned.)
OILS	Weekly expenditure on buying fats and oils such as olive oil, margarine, butter, coconut oil
EARNCOMP	Composition of earners CODED 1: Reference person only 2: Reference person and spouse 3: Reference person, spouse, and others 4: Reference person and others 5: Spouse only 6: Spouse and others 7: Others only 8: No earners
HOUSING	Housing tenure CODED 1: Owned with mortgage 2: Owned without mortgage 3: Owned(mortgage not reported) 4: Rented 5: Occupied without payment of cash rent 6: Student housing



Our first question was whether a reference person's marital status affects their weekly grocery expenditures.

First by looking at the scatter plot of each marital status, it seemed that unmarried(0)[4] families spent about $50 less on weekly groceries than married(1) families. The grocery data for each group seemed to be distributed similarly. There were a couple of married families that spent extreme amount of money on weekly groceries.

It was noticeable that one family spent extremely large amount of money on average for weekly groceries. Thus we looked into the original dataset with 2951 observations and searched for variables that are relatively highly correlated with grocery expenditure. The strongest predictors were family size, count of minors, income(in 2015), and count of earners; correlation coefficients were 0.5, 0.39, 0.37, 0.35 in order. The outlier, the family that spent $800 on groceries on average had 6 family members, 5 earners, and their annual income was $259,473.3 in 2015. In the dataset, the proportion of families with 6 or more members was 3.2% and the proportion of those with annual income greater than $200,000 was 3.1%. The maximum count of earners was 5 in the dataset and it comprised only 0.6%. There were only 2 out of 2951 households that had 6 family members with income more than $200,000 including the outlier. With insufficient amount of data to compare and considering the outlier's exceptionally large family size, number of earners and annual income, we decided to keep the outlier.

For testing two independent married and not married households, our null hypothesis was

> "The average weekly grocery expenditure of the family is the same for married families and unmarried families."

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
0	225	97.227	70.1319	4.6755	88.01	106.44
1	254	149.846	91.9417	5.7689	138.49	161.21

Sample standard deviation of the unmarried family's grocery expenditure was $70.13 and that of the married family's was $91.94 which is less than the twice of the sample standard deviation of the unmarried family's expenditure. Thus, we used the pooled variance for 'two independent samples t-test'. We concluded that the average weekly grocery expenditure is different for married families and unmarried families with p-value less than 0.0001.

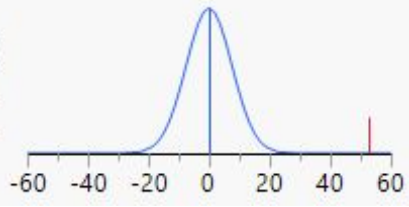
[4] For convenience purposes, we referred to a family where reference person is married as a 'married family' and a family where reference person is not married as 'unmarried family' in this paper.

t Test

1-0

Assuming equal variances

Difference	52.6198	t Ratio	6.973455
Std Err Dif	7.5457	DF	477
Upper CL Dif	67.4468	Prob > |t|	<.0001*
Lower CL Dif	37.7928	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



Additionally, the difference between the average weekly grocery expenditures of married families and unmarried families will be between $37.79 and $67.45 for 95% of the time random samples of 500 observations are measured.

When we excluded the outlier, our conclusion was the same, but with a different 95% confidence interval.

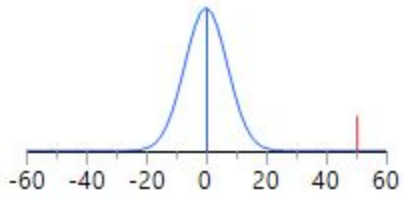
Without the outlier, the average weekly grocery expenditure difference between married and unmarried families was between $36.20 and $63.90 with 95% confidence.

t Test

1-0

Assuming equal variances

Difference	50.0500	t Ratio	7.101091
Std Err Dif	7.0482	DF	476
Upper CL Dif	63.8995	Prob > |t|	<.0001*
Lower CL Dif	36.2006	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



We were also curious if EARNCOMP[5], HOUSING[6], and MARITAL altogether affect GROCYWK[7] separately and/or interactively. With the three-way ANOVA model, we tested three main factor effects and four interaction effects; 1)EARNCOMP, 2)HOUSING, 3)MARITAL, 4)EARNCOMP*HOUSING, 5)EARNCOMP*MARITAL, 6)HOUSING*MARITAL, 7)EARNCOMP*HOUSING*MARITAL.

Tests of Between-Subjects Effects

Dependent Variable: GROCYWK

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1258116.79[a]	47	26768.442	4.989	.000
Intercept	813833.756	1	813833.756	151.671	.000
HOUSING * EARNCOMP * MARITAL	129299.288	5	25859.858	4.819	.000
HOUSING * EARNCOMP	233989.427	22	10635.883	1.982	.005
EARNCOMP * MARITAL	69317.472	3	23105.824	4.306	.005
HOUSING * MARITAL	60212.134	4	15053.034	2.805	.025
HOUSING	107047.457	5	21409.491	3.990	.002
EARNCOMP	212406.196	7	30343.742	5.655	.000
MARITAL	24721.607	1	24721.607	4.607	.032
Error	2312657.189	431	5365.794		
Total	11070657.00	479			
Corrected Total	3570773.975	478			

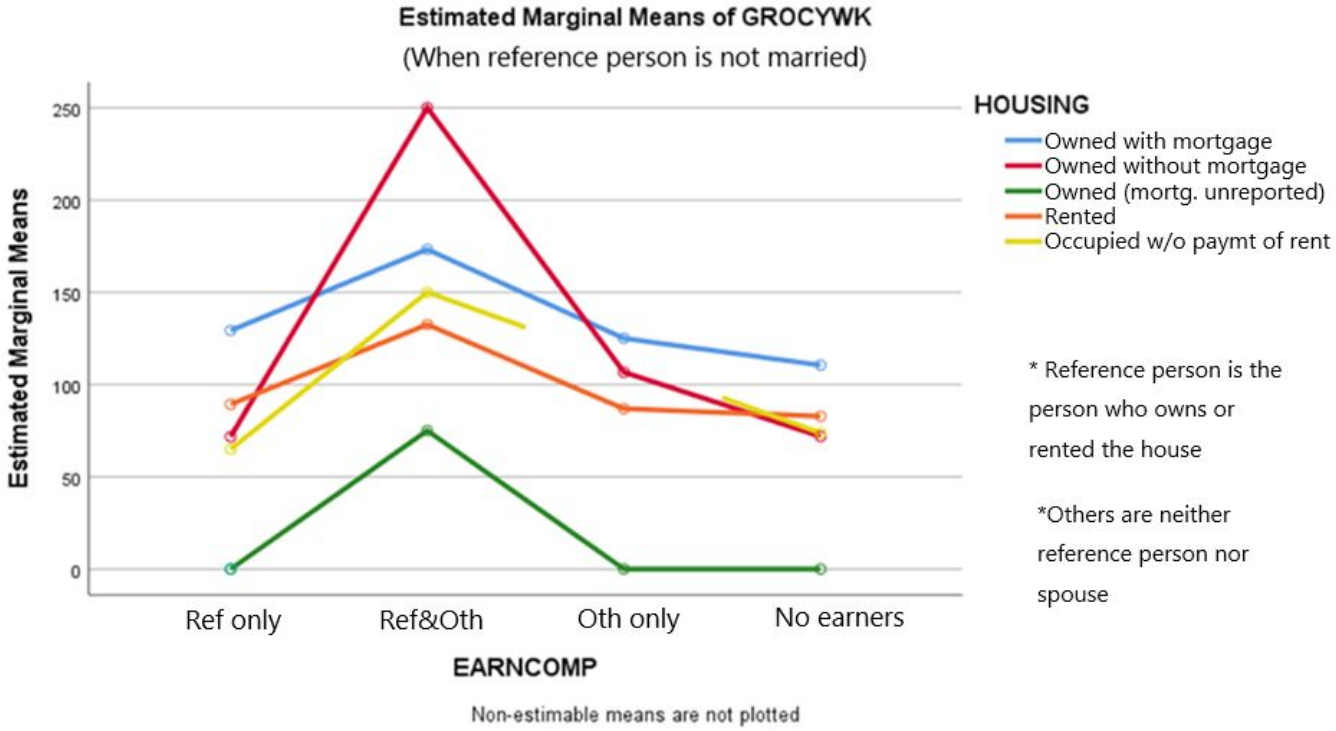
a. R Squared = .352 (Adjusted R Squared = .282)

[5] Composition of earners e.g., Reference person & his/her spouse, Spouse only, Reference person only

[6] Housing status e.g., Rented, Owned with/without mortgage

[7] Weekly grocery expenditure

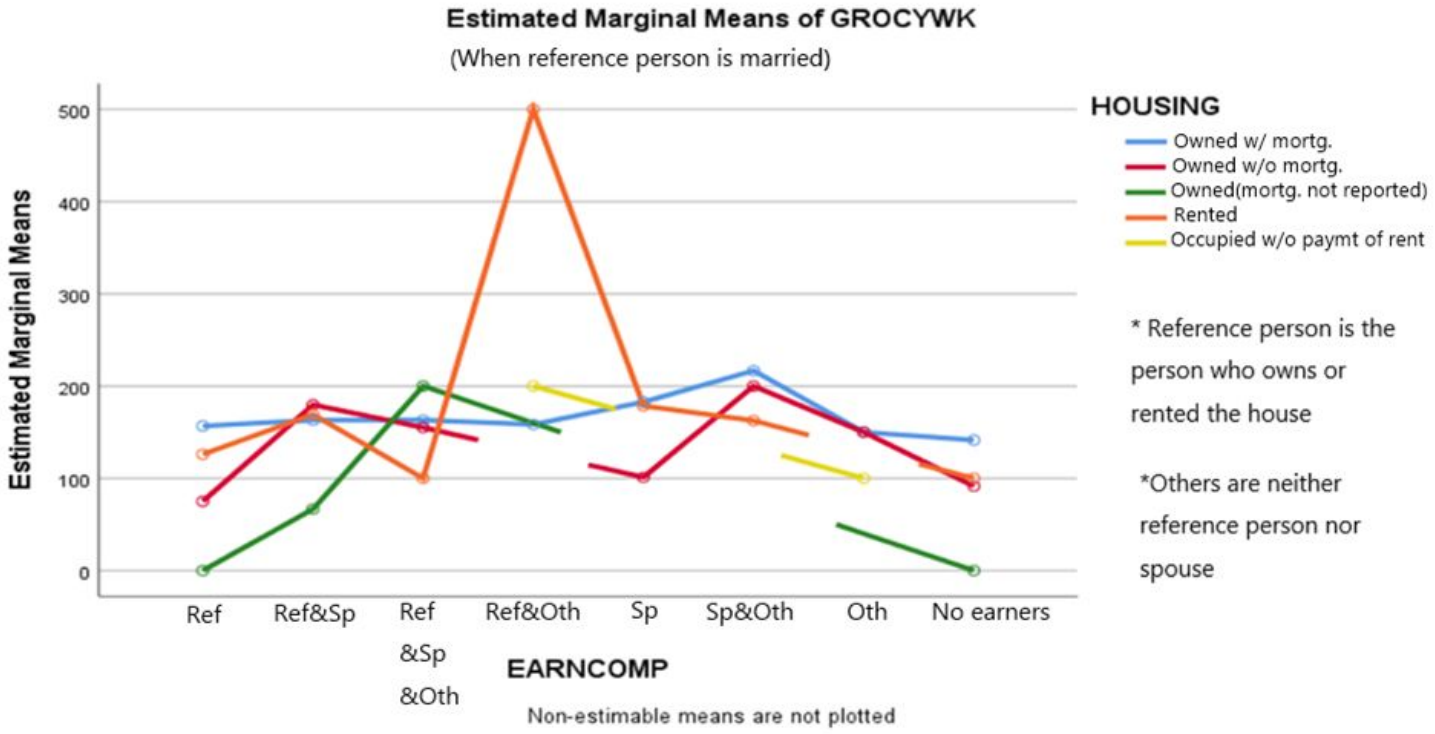
The result showed that F-test statistics of all the individual and interaction terms have significant p-values less than 0.05. Thus our conclusion was that all the main terms and all the possible interaction terms affect a weekly grocery expenditure which are 1)Housing status, 2)Composition of earners, 3)Marital status, 4)Housing & Earners, 5)Housing & Marital status, 6)Marital status & Earners, and 7)Housing status & Earners & Marital status.



The plot for the ANOVA model for the unmarried families showed the interaction effect between some of HOUSING categories and some of EARNCOMP categories.

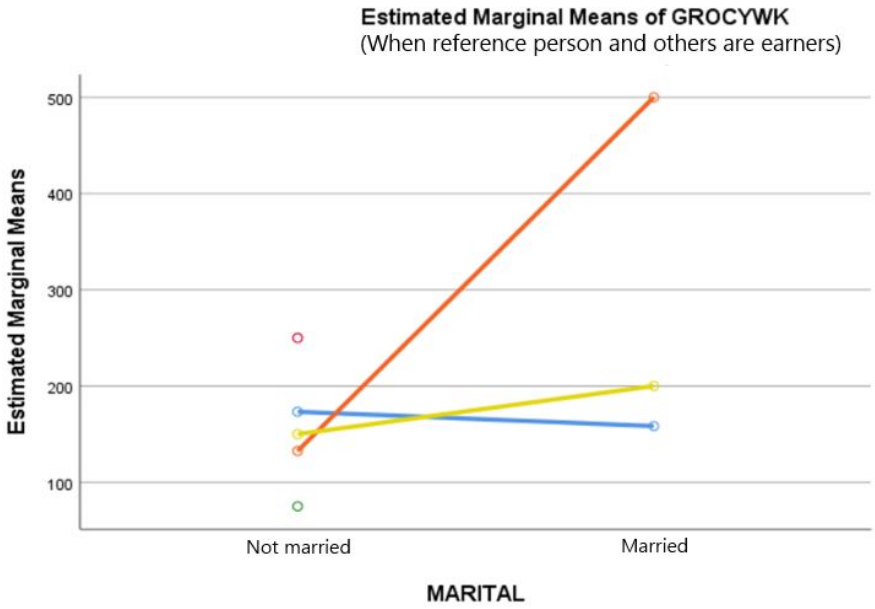
Overall, unmarried families with reference person and others as earners spent more money on groceries than families with other categories: ‘reference person only’, ‘others only’, and ‘no earners’. Their expenditure became extreme when the reference person owned the house without mortgage.

The blue line and the orange line seemed to be parallel for ‘reference person only’, ‘reference person and others only’, and ‘others only’. In other words, for unmarried families, the three earner categories affect average weekly grocery expenditure in the same way for a family with home mortgage and a family paying rent. For instance, let’s assume there is an unmarried family where the reference person is the only earner. If another family member is added as a new earner, the difference in their average weekly grocery expenditure will be the same whether the family owns their house with mortgage or pays rent.



The plot for the married families indicated even more significant interaction between HOUSING and EARNCOMP. The lines were so intertwined that there was no pair of lines that are parallel, which meant that every level of HOUSING affects GROCYWK differently for every level of EARNCOMP and vice versa. For married families that pay rent, those with reference person and others as earners spent over twice as much money on groceries as do those with any other earner compositions on average. But having reference person and others as earners does not affect the average expenditure as extreme when the house is on mortgage or occupied without payment of rent. Also, families with home mortgage is less likely to be affected by which household members are earners than those with other housing status.

From both plots, families with a house of unreported mortgage showed zero expenditure for certain earner categories. Thus we assumed the data was biased and excluded it from our interpretations of the whole data.



Replacing the horizontal axis variable with MARITAL, we could verify that the average weekly grocery expenditure of families that pay rent depends heavily on the reference person's marital status when the reference person and others(not spouse) earn money. With this type of earners, married families that pay rent spent almost four times more money on groceries on average than unmarried families that pay rent.

Without the outlier, however, our ANOVA results and conclusions became different.

Tests of Between-Subjects Effects

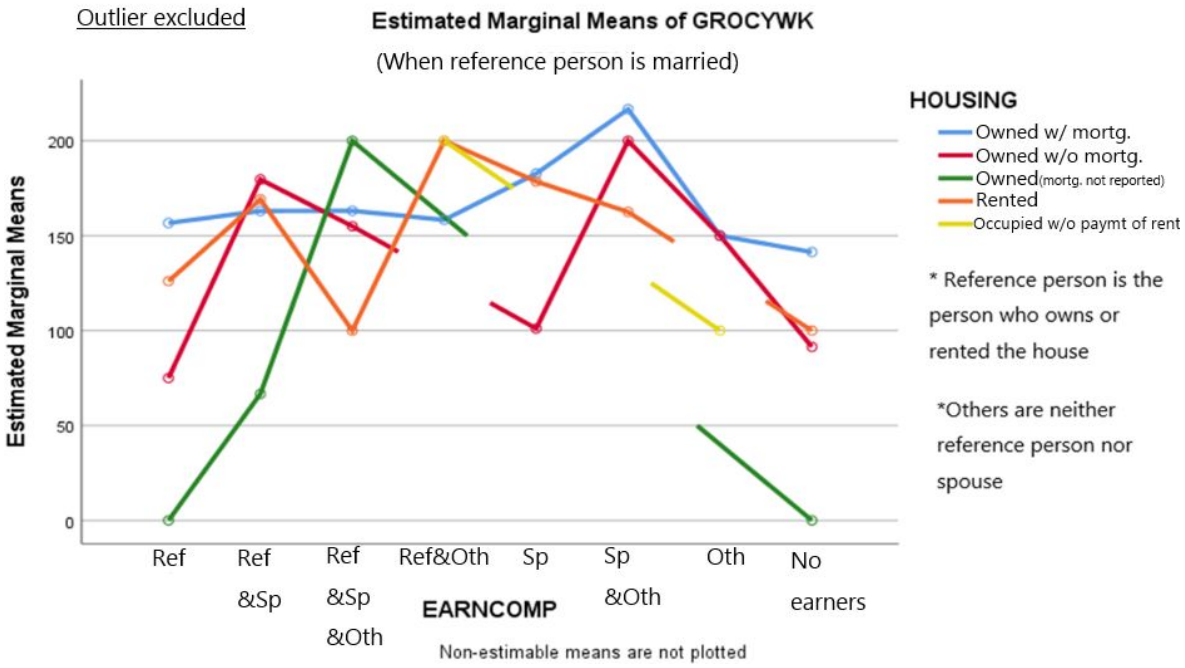
Dependent Variable: GROCYWK

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	981713.684[a]	47	20887.525	4.211	.000
Intercept	800156.594	1	800156.594	161.333	.000
EARNCOMP	121953.976	7	17421.997	3.513	.001
HOUSING	115893.578	5	23178.716	4.673	.000
MARITAL	7193.801	1	7193.801	1.450	.229
EARNCOMP * HOUSING	116968.756	22	5316.762	1.072	.375
EARNCOMP * MARITAL	624.318	3	208.106	.042	.989
HOUSING * MARITAL	3947.297	4	986.824	.199	.939
EARNCOMP * HOUSING * MARITAL	4795.418	5	959.084	.193	.965
Error	2132657.189	430	4959.668		
Total	10430657.00	478			
Corrected Total	3114370.872	477			

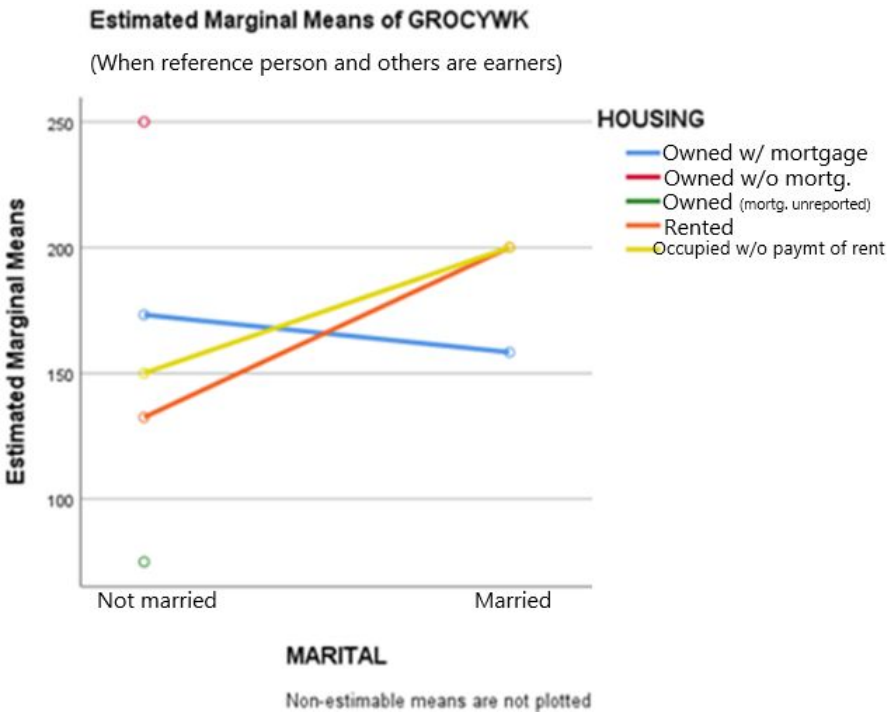
a. R Squared = .315 (Adjusted R Squared = .240)

The test result showed that only main effects of EARNCOMP and HOUSING are significant for GROCYWK.

The plot for the unmarried group looked the same with and without the outlier since the outlier belonged to the married group, but the plot for the married group looked very different with and without the outlier.



In the plot for the married group, the lines were still intertwined, but unlike the plot with the outlier, married families that pay rent did not spend extreme amount of money on groceries on average when the reference person and others are earners.



Based on the left plot, an unmarried household that owns a house without mortgage with reference person and others as earners spent at least $75 more on average than those with other housing situations with the same type of earners. A married group that rented a house(with payment of rent or without) with reference person spent about $200 on average for groceries. This outcome was completely different from the previous test result that included the outlier. Although a married group that rented a house with reference person and others as earners was the largest consumer than other married groups with the same category of earners, but it was not as extreme as it was with the outlier.

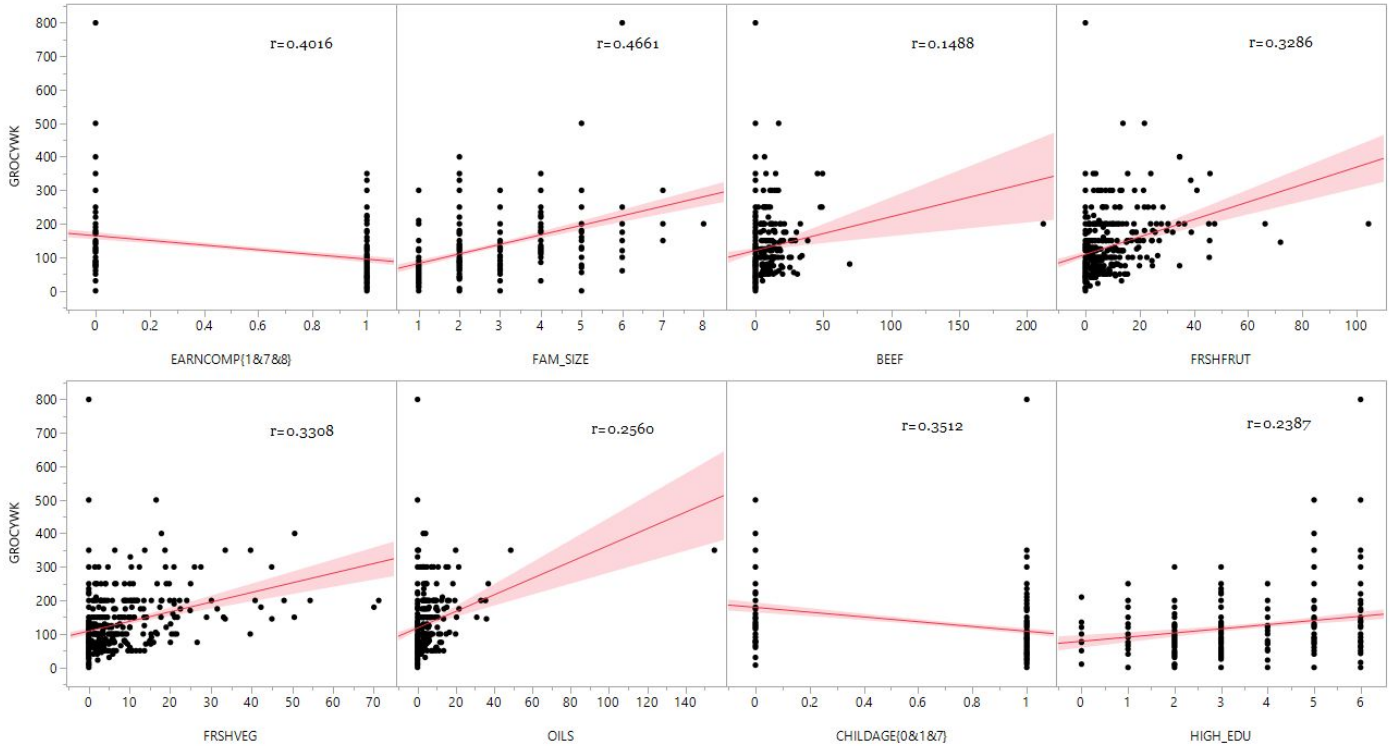
After ANOVA test, we searched for other significant predictors to produce a reliable multiple linear regression model. Using backward variable selection, we found 8 significant linear predictors out of 43 predictors[8] from our dataset which were 1)EARNCOMP{1&7&8}[9], 2)FAM_SIZE, 3)BEEF, 4)FRSHFRUT, 5)FRSHVEG, 6)OILS, 7)CHILDAGE{0&1&7}[10], and 8)HIGH_EDU.

[8] 43 predictors:

AGE_REF							
URBN_RUR	HOUSING	EARNCOMP	EDUC_REF	FAM_SIZE	MARITAL	NO_EARNR	
PERSLT18	PERSOT64	REGION	SEX_REF	VEHQ	WK_WRKD	CEREAL	
BAKEPROD	BEEF	PORK	OTHMEAT	POULTRY	SEAFOOD	EGGS	
MILKPROD	OTHDAIRY	FRSHFRUT	FRSHVEG	PROCFRUT	PROCVEG	SWEETS	NONALBEV
OILS	MISCFOOD	FOODAWAY	ALCBEV	SMOKSUPP	PET_FOOD	DRUGSUPP	
HOUSKEEP	CHILDAGE	INCOME	HIGH_EDU	BIRTHYR	REFGEN		

[9] EARNCOMP{1&7&8} = 1 if a household is one of the three cases below.
Only the reference person(who owns/rents the house) earns money. **OR**
Neither reference person or his/her spouse earn money, but there are other earners. **OR**
There is no earner in the household.
EARNCOMP{1&7&8} = 0 otherwise.

[10] CHILDAGE{0&1&7} = 1 if there is no children **OR** all children are less than 6 years old **OR** all children are greater than 17.
CHILDAGE{0&1&7} = 0 otherwise.



From the scatter plot, FAM_SIZE and EARNCOMP{1&7&8} seemed to be the strongest linear predictors with r>0.4. BEEF seemed to have extremely weak linear predictability with r=0.15. EARNCOMP{1&7&8} and CHILDAGE{0&1&7} have a negative linear relationship with GROCYWK while other predictors have positive relationship with GROCYWK. We could also tell that FRSHFRUT and FRSHVEG have similar relationship with GROCYWK, so we assumed FRSHFRUT and FRSHVEG would have a linear relationship. As we had expected, there was a moderately strong positive linear relationship between FRSHFRUT and FRSHVEG with r = 0.6755. Thus we anticipated one of them or both would eventually be eliminated due to the collinearity.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	63.56612	19.3714	3.28	0.0011*
FAM_SIZE	19.157341	3.374282	5.68	<.0001*
BEEF	-0.10854	0.462894	-0.23	0.8147
FRSHFRUT	1.4186877	0.476575	2.98	0.0031*
FRSHVEG	1.0657232	0.535962	1.99	0.0474*
OILS	1.493623	0.65854	2.27	0.0238*
HIGH_EDU	8.114671	2.031663	3.99	<.0001*
CHILDAGE{0&1&7}	-23.41724	12.31549	-1.90	0.0579
EARNCOMP{1&7&8}	-20.72625	7.759441	-2.67	0.0078*
(FAM_SIZE-2.52818)*(CHILDAGE{0&1&7}-0.762)	10.959597	7.086825	1.55	0.1227
(BEEF-4.6699)*(CHILDAGE{0&1&7}-0.762)	0.8355449	0.703594	1.19	0.2356
(FRSHFRUT-6.26786)*(CHILDAGE{0&1&7}-0.762)	1.0544303	0.9212	1.14	0.2530
(FRSHVEG-5.69619)*(CHILDAGE{0&1&7}-0.762)	0.0507416	0.94295	0.05	0.9571
(OILS-2.92499)*(CHILDAGE{0&1&7}-0.762)	1.6829943	1.107804	1.52	0.1294
(HIGH_EDU-3.77453)*(CHILDAGE{0&1&7}-0.762)	-9.723799	4.906735	-1.98	0.0481*
(FAM_SIZE-2.52818)*(EARNCOMP{1&7&8}-0.56367)	-2.900573	5.469691	-0.53	0.5962
(BEEF-4.6699)*(EARNCOMP{1&7&8}-0.56367)	0.2210394	0.925853	0.24	0.8114
(FRSHFRUT-6.26786)*(EARNCOMP{1&7&8}-0.56367)	-0.216326	0.930047	-0.23	0.8162
(FRSHVEG-5.69619)*(EARNCOMP{1&7&8}-0.56367)	0.9800287	0.967509	1.01	0.3116
(OILS-2.92499)*(EARNCOMP{1&7&8}-0.56367)	1.4735533	1.21638	1.21	0.2264
(HIGH_EDU-3.77453)*(EARNCOMP{1&7&8}-0.56367)	-7.570745	4.272472	-1.77	0.0771

The parameter t-test results showed that 1)FAM_SIZE, 2)FRSHFRUT, 3)FRSHVEG, 4)OILS, 5)HIGH_EDU, 6)EARNCOMP{1&7&8}, and 7)HIGH_EDU*CHILDAGE{0&1&7} were significant.

Parameter Estimates

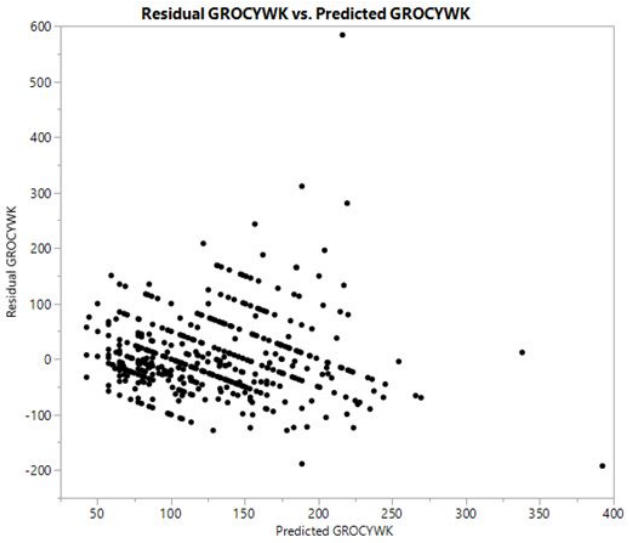
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	50.306685	12.88475	3.90	0.0001*
FAM_SIZE	19.714675	2.61117	7.55	<.0001*
FRSHFRUT	0.6900653	0.411965	1.68	0.0946
FRSHVEG	0.8927261	0.452855	1.97	0.0493*
OILS	1.0167543	0.378814	2.68	0.0075*
HIGH_EDU	7.361796	2.027815	3.63	0.0003*
EARNCOMP{1&7&8}	-27.41787	7.555177	-3.63	0.0003*
(HIGH_EDU-3.77453)*(CHILDAGE{0&1&7}-0.762)	-7.556113	4.635734	-1.63	0.1038

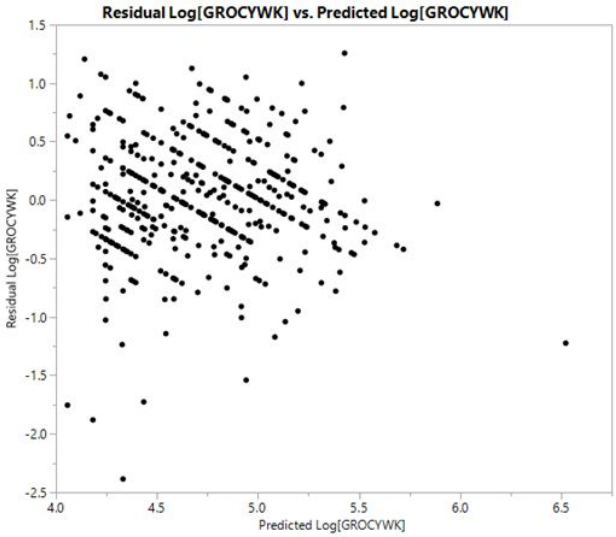
The model with the 7 terms indicated that FRSHFRUT and HIGH_EDU*CHILDAGE{0&1&7} were not significant. As we had assumed earlier, FRSHFRUT was insignificant due to the existence of the correlated FRSHVEG. After the insignificant predictors were removed, our final model had 5 predictors: FAM_SIZE, FRSHVEG, OILS, HIGH_EDU, and EARNCOMP{1&7&8}.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	51.111501	12.91615	3.96	<.0001*
FAM_SIZE	20.056326	2.602861	7.71	<.0001*
FRSHVEG	1.4022446	0.35327	3.97	<.0001*
OILS	1.0927411	0.379341	2.88	0.0041*
HIGH_EDU	7.4631306	2.034182	3.67	0.0003*
EARNCOMP{1&7&8}	-28.45909	7.561047	-3.76	0.0002*

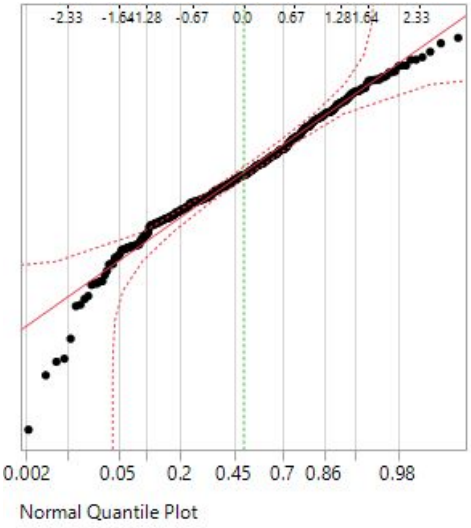
After this step, we decided to check the residual plot against the predicted values.





The residuals(on the left) seemed to have a tendency to spread out as the predicted values grow larger. Thus, we transformed GROCYWK to Log[GROCYWK] to observe a better residual plot.

The residuals of the transformed model(on the right) were more evenly spread out along the x-axis than the original residuals.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	4.1560503	0.093128	44.63	<.0001*	3.9730251	4.3390754
FAM_SIZE	0.1491275	0.018609	8.01	<.0001*	0.1125547	0.1857003
FRSHVEG	0.0087114	0.002503	3.48	0.0006*	0.0037921	0.0136307
OILS	0.0052937	0.002669	1.98	0.0479*	4.8076e-5	0.0105394
HIGH_EDU	0.0629014	0.01474	4.27	<.0001*	0.033933	0.0918697
EARNCOMP{1&7&8}	-0.248134	0.054388	-4.56	<.0001*	-0.355023	-0.141244

The qq-plot with the transformed model seemed to fit well with x=y line except for a few dots at the bottom left tail. Since data size was large enough, we could apply Central Limit Theorem and assume normality of the residuals with caution.

So our final model rounded to three decimals was:

Log[GROCYWK] = 4.156 + 0.149(FAM_SIZE) + 0.009(FRSHVEG) + 0.005(OILS) + 0.063(HIGH_EDU)

- 0.248(EARNCOMP{1&7&8})

Summary of Fit

RSquare	0.361225
RSquare Adj	0.354047
Root Mean Square Error	0.498102
Mean of Response	4.711454
Observations (or Sum Wgts)	451

The adjusted coefficient of determination was 0.354. Thus, 35.4% of variation in Log[weekly grocery expenditure] data could be predicted by our fitted model.

For example, we could say for every person added to a family, their average weekly grocery expenditure will increase by 16.1%(1.161=e^0.149) given that all the other variables remain the same.

FRSHVEG effect in this model was small since the average coefficient was between 0.004 and 0.014 with 95% confidence. It means for every dollar spent on buying fresh vegetables, the increased rate of weekly grocery expenditure will be 0.4%~1.4% on average given that all the other variables remain the same.

Similarly, for every dollar spent on buying oil products, the increased rate of weekly grocery expenditure will be at most 1% on average with 95% confidence given that all the other variables remain the same.

For HIGH_EDU, the average coefficient was between 0.034 and 0.092 with 95% confidence. For example, if a person with the highest education in the family had a bachelor's degree and earns a master's degree, the weekly grocery expenditure will increase by 3.5%~9.6% on average given that all the other variables remain the same. But if the person with a master's earns a doctoral degree, it will not affect the average grocery expenditure in this model.

For EARNCOMP{1&7&8}, suppose there are two households that have different compositions of earners in their families, but their FAM_SIZE, FRSH_VEG, OILS, and HIGH_EDU are the same. Let's also assume family A has reference person as the only earner OR has earners other than reference person or his/her spouse OR has no earner and family B is not in one of those cases. Then families of type A will spend 13.2%~30% less money on weekly groceries on average than families of type B(0.132=1-e^-0.141 0.30=1-e^-0.355) with 95% confidence.

Without the outlier, the model created after by the same process as above was a little different.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.1914996	0.094712	44.26	<.0001*
FAM_SIZE	0.1556797	0.018796	8.28	<.0001*
HIGH_EDU	0.0725699	0.014673	4.95	<.0001*
EARNCOMP{1&7&8}	-0.288412	0.054482	-5.29	<.0001*

Summary of Fit

RSquare	0.324929
RSquare Adj	0.320388
Root Mean Square Error	0.505677
Mean of Response	4.707069
Observations (or Sum Wgts)	450

The fitted model without the outlier was:

Log[GROCYWK] = 4.191 + 0.156(FAM_SIZE) + 0.073(HIGH_EDU) - 0.288(EARNCOMP{1&7&8})

32% of variation in Log[weekly grocery expenditure] data can be predicted by this regression model.

From ANOVA test results, we confirmed that HOUSING and EARNCOMP have effects on GROCYWK. In our multiple regression model, EARNCOMP{1&7&8} and FAM_SIZE were strong predictors for GROCYWK. Thus, we tried making another regression model with HOUSING, EARNCOMP, and FAM_SIZE.

The individual t-test results showed HOUSING[1], HOUSING[2], EARNCOMP[7], FAM_SIZE*HOUSING[2], FAM_SIZE*EARNCOMP[4], and FAM_SIZE*EARNCOMP[5] were significant.

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	62.175492	25.27703	2.46	0.0143*
FAM_SIZE	Biased	10.637108	13.67103	0.78	0.4369
HOUSING[1]	Biased	62.724666	23.31144	2.69	0.0074*
HOUSING[2]	Biased	51.805357	24.22771	2.14	0.0330*
HOUSING[3]	Biased	-15.51691	31.36864	-0.49	0.6211
HOUSING[4]	Biased	37.503147	22.94267	1.63	0.1028
HOUSING[5]	Biased	19.174895	28.62746	0.67	0.5033
EARNCOMP[1]		-16.03383	9.700462	-1.65	0.0990
EARNCOMP[2]		17.50522	10.15023	1.72	0.0853
EARNCOMP[3]		-13.052	22.16	-0.59	0.5562
EARNCOMP[4]		11.753907	12.63083	0.93	0.3526
EARNCOMP[5]		-0.88801	14.75228	-0.06	0.9520
EARNCOMP[6]		67.142254	42.2473	1.59	0.1127
EARNCOMP[7]		-33.42891	16.09062	-2.08	0.0383*
HOUSING[1]*(FAM_SIZE-2.52818)	Biased	-8.430989	14.51013	-0.58	0.5615
HOUSING[2]*(FAM_SIZE-2.52818)	Biased	30.934663	15.69843	1.97	0.0494*
HOUSING[3]*(FAM_SIZE-2.52818)	Biased	30.367204	22.79895	1.33	0.1835
HOUSING[4]*(FAM_SIZE-2.52818)	Biased	15.097232	14.23544	1.06	0.2895
HOUSING[5]*(FAM_SIZE-2.52818)	Zeroed	0	0	.	.
EARNCOMP[1]*(FAM_SIZE-2.52818)		3.2426565	5.591402	0.58	0.5622
EARNCOMP[2]*(FAM_SIZE-2.52818)		2.2225032	7.293909	0.30	0.7607
EARNCOMP[3]*(FAM_SIZE-2.52818)		5.0535626	11.52381	0.44	0.6612
EARNCOMP[4]*(FAM_SIZE-2.52818)		18.753029	7.326773	2.56	0.0108*
EARNCOMP[5]*(FAM_SIZE-2.52818)		19.281578	9.062901	2.13	0.0339*
EARNCOMP[6]*(FAM_SIZE-2.52818)		-29.32789	15.40732	-1.90	0.0576
EARNCOMP[7]*(FAM_SIZE-2.52818)		-8.326401	16.67343	-0.50	0.6178

Although the individual t-test showed that FAM_SIZE itself was not a significant predictor, considering its relatively strong correlation with GROCYWK and bias in its estimate, we included it in our model with other significant variables and tested it again.

Parameter Estimates

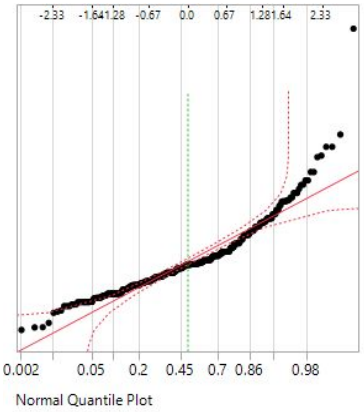
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	36.394567	8.386433	4.34	<.0001*
HOUSING_1	38.645646	7.846082	4.93	<.0001*
HOUSING_2	23.821919	9.764053	2.44	0.0151*
EARNCOMP_7	-22.76397	17.32394	-1.31	0.1895
(HOUSING_2-0.22756)*(FAM_SIZE-2.52818)	28.364122	7.596081	3.73	0.0002*
(EARNCOMP_4-0.09603)*(FAM_SIZE-2.52818)	19.646425	7.433981	2.64	0.0085*
(EARNCOMP_5-0.05637)*(FAM_SIZE-2.52818)	9.0886611	9.401523	0.97	0.3342
FAM_SIZE	28.329146	2.605842	10.87	<.0001*

FAM_SIZE was strongly significant with p-value<0.0001 while EARNCOMP[7] and EARNCOMP[5]*FAM_SIZE terms changed from significant to insignificant with bigger p-value>0.05. Thus we were left with 5 terms: HOUSING[1], HOUSING[2], HOUSING[2]*FAM_SIZE,

EARNCOMP[4]*FAM_SIZE, and FAM_SIZE.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	35.42201	8.35858	4.24	<.0001*
HOUSING_1	39.327282	7.836526	5.02	<.0001*
HOUSING_2	23.949722	9.767837	2.45	0.0146*
(HOUSING_2-0.22756)*(FAM_SIZE-2.52818)	28.082794	7.599761	3.70	0.0002*
(EARNCOMP_4-0.09603)*(FAM_SIZE-2.52818)	19.149127	7.380979	2.59	0.0098*
FAM_SIZE	28.359158	2.602954	10.89	<.0001*



Then we checked the residual plot. The residuals seemed to have spread out evenly along the x-axis, but the qq-plot had a u-shaped pattern. So we transformed GROCYWK to Log[GROCYWK] and checked the estimates again.

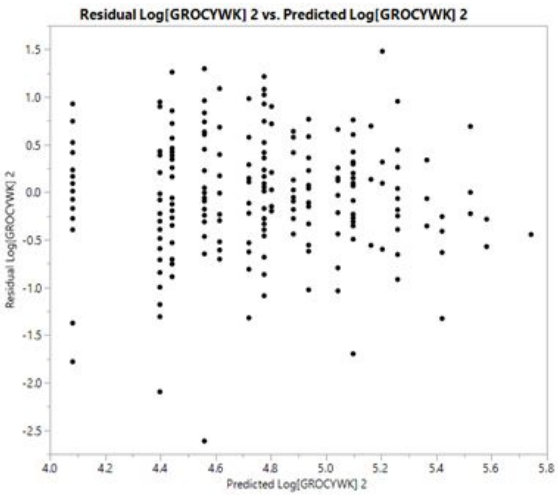
Parameter Estimates

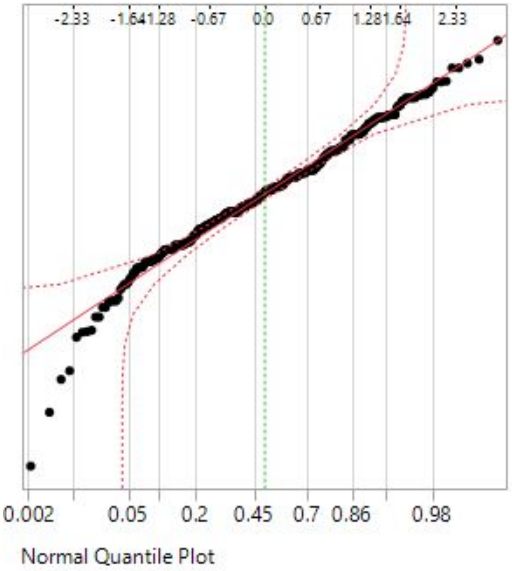
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.0822736	0.064183	63.60	<.0001*
HOUSING_1	0.2384114	0.057321	4.16	<.0001*
HOUSING_2	0.0585304	0.072141	0.81	0.4176
(HOUSING_2-0.24169)*(FAM_SIZE-2.58758)	0.2211161	0.054006	4.09	<.0001*
(EARNCOMP_4-0.09534)*(FAM_SIZE-2.58758)	0.0395486	0.053368	0.74	0.4591
FAM_SIZE	0.2123905	0.018987	11.19	<.0001*

Removing HOUSING[2] and EARNCOMP[4]*FAM_SIZE gave us following estimates of coefficients.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.1126887	0.052183	78.81	<.0001*
HOUSING_1	0.2163486	0.051842	4.17	<.0001*
(HOUSING_2-0.24169)*(FAM_SIZE-2.58758)	0.1989971	0.048845	4.07	<.0001*
FAM_SIZE	0.2092619	0.018208	11.49	<.0001*





The residuals plot seemed to spread even more nicely than the original plot and qq-plot seemed to have lost the u-shape and fit the x=y line except for a few residuals at the bottom left tail.

Thus the final model rounded to the nearest thousandth was:

Log[GROCYWK]=
4.113+0.216(HOUSING[1])+0.199(HOUSING[2]-0.242)*(FAM_SIZE-2.588)+0.209(FAM_SIZE)

There are 4(=2^2) different cases and the models for each of those cases has FAM_SIZE as the only variable. One example would be a family of 4 members that owns a house with mortgage. In that case,

Log[GROCYWK] = 4.113+0.216(1)+0.199(0-0.242)*(4-2.588)+0.209(4)=5.097

Then, GROCYWK = e^5.097 = $163.53 will be the average weekly grocery expenditure.

If the same family owns a house without mortgage,

Log[GROCYWK] = 4.113+0.216(0)+0.199(1-0.242)*(4-2.588)+0.209(4) = 5.162

Then, GROCYWK = e^5.162 = $174.51 will be the average weekly grocery expenditure.

Summary of Fit

RSquare	0.296268
RSquare Adj	0.291545
Root Mean Square Error	0.521644
Mean of Response	4.711454
Observations (or Sum Wgts)	451

The adjusted coefficient of determination for this model is 0.291545. In other words, 29% of Log[weekly grocery expenditure] data can be explained by this model.

Without the outlier, the interaction term HOUSING*FAM_SIZE was excluded from the fitted model above.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.1232618	0.05281	78.08	<.0001*
FAM_SIZE	0.1860823	0.01804	10.32	<.0001*
HOUSING_1	0.2564375	0.051913	4.94	<.0001*

Summary of Fit

RSquare	0.265399
RSquare Adj	0.262113
Root Mean Square Error	0.526912
Mean of Response	4.707069
Observations (or Sum Wgts)	450

The fitted model without the outlier was:

Log[GROCYWK] = 4.123+0.186(FAM_SIZE)+0.2564375(HOUSING[1])

26.2% of variation in Log[weekly grocery expenditure] data can be explained by this model.

Bibliography

U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, Diary Survey, April 10, 2018. https://www.bls.gov/cex/pumd_data.htm#csv

U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, 2016 Consumer Expenditure Surveys Microdata Diary Documentation, August 29, 2017. https://www.bls.gov/cex/2016/csxdiary.pdf

U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, 2016 Diary Data Dictionary, Wednesday, September 13, 2017. https://www.bls.gov/cex/2016/csxdiarydata.pdf

U.S. Department of Labor. Bureau of Labor Statistics. Division of Information and Marketing Services. (February 25, 2016). *Consumer Expenditures and Income: Collections & Data Sources: Diary Survey details.* https://www.bls.gov/opub/hom/cex/data.htm

U.S. Department of Labor. Bureau of Labor Statistics. Division of Information and Marketing Services. (February 25, 2016). *Consumer Expenditures and Income: Concepts.* https://www.bls.gov/opub/hom/cex/concepts.htm

SAS Institute Inc. 2017. *Using JMP® 13*. Cary, NC: SAS Institute Inc.

IBM Corp. Released 2017. IBM SPSS Statistics Subscription for Microsoft Windows 64-bit, build 1.0.0.1012. Armonk, NY: IBM Corp.