Economics 143 Final Project

David Yu

June 15, 2021

1 Introduction

For my final project, I chose to extend Project 1 by applying the empirical analysis. I have provided a revised version of my original Project 1 submission below. Then, in the newly added Part G, I record the finalized results of the empirical work. The cleaned data, analysis code, and intermediate output are included with my submission as separate files. By intermediate output, I mean the results of the initial regressions and the results of my examination of each issue (for example, auxiliary regressions). These intermediate results are omitted for brevity, but are included in the submission for completeness (though the output can be fully replicated using the provided code, of course). Additionally, I have linked a GitHub repository in Part G, which contains the same files.

2 Part A: Question

Question: What is the effect of various factors (such as income, household size, distance to store, among others) on the size of a grocery store bill?

The size of a grocery store bill is measured by the monetary amount of each transaction (in USD) at each store.

3 Part B: Motivation

Governments and other aid providers would be interested in the results of this investigation, because it could help inform their policies for providing food-related assistance. For example, suppose that the local government wants to support low-income families in their city, and we find that income is negatively correlated with bill size (maybe because higher income individuals go to restaurants more often, or can afford to make more trips). Then the government might consider a rebate program for spending a lot of money in one trip, to supplement an existing aid program.

Stores themselves would also benefit from this information, as it would allow them to structure their pricing model and overall strategy to increase revenue. Suppose people who live far away spend a lot per trip, and the store wants to offer some type of reward for making a large purchase. Then the store might prefer giving a gas coupon over another potential reward, since this would likely benefit people who live further away more, and thus give them incentive to spend more at the store.

4 Part C: Model and Data

4.1 Model

The basic OLS model is as follows:

$$bill_i = \beta_1 + \beta_2 inc_i + \beta_3 hhsize_i + \beta_4 dist_i + \beta_5 cash_i + \beta_6 items_i + e_i$$

where $bill_i$ is the amount of money spent in one transaction, inc_i is the total monthly income of the household making the purchase, $hhsize_i$ the the number of individuals in the household, $dist_i$ is the distance (in miles)

from the store to the household's living place, $cash_i$ is an indicator variable that equals 1 if the transaction used cash and 0 otherwise, and $items_i$ is a discrete variable measuring the number of items purchased.

4.2 Data

The Economic Research Service of the US Department of Agriculture (USDA) provides a large collection of data at this link. All of the variables discussed in this project have been examined and recorded in this dataset.

5 Part D: Issues

5.1 Issue 1

We may have nonlinear effects in some variables. For example, it is plausible that as income increases, so does bill size, until eventually income is large enough that households begin to spend more money at restaurants and less money on groceries. Then bill size might go down as income gets really high.

5.2 Issue 2

We might suffer from omitted variable bias because the age of the person making the purchase is not included. Since age could be correlated with household size, this could cause the estimate β_2 to become biased.

However, upon further thought, I decided this probably wouldn't be much of a problem. Age should only be predictive of grocery bill size by way of household size. As an example, suppose there were individuals of varying age, all of whom lived by themselves. Then age probably wouldn't say very much about the size of each individual's grocery store bill.

5.3 Issue 3

Some of the explanatory variables might be correlated, leading to collinearity. This might happen with household size and number of items purchased, or between income and distance from store.

5.4 Issue 4

As income increases from low to high, it is plausible that the variance in bill size will increase as well. This makes intuitive sense since low-income individuals don't have a lot of choice (people need to eat) but high-income individuals do have choice. They might enjoy eating foods that are low-cost, have expensive taste, or prefer to eat in restaurants (therefore incurring low costs at the store). In this case there would be heteroskedasticity in the model.

6 Part E: Issues (Mathematical)

6.1 Issue 1

Let's return to the example outlined in Part D, in which bill size depends on income in a nonlinear fashion. To spotlight the issue, suppose that we estimate the simpler model

$$bill_i = \beta_1 + \beta_2 inc_i + v_i$$

but that the true model is

$$bill_i = \beta_1 + \beta_2 inc_i + \beta_3 inc_i^2 + e_i$$

Our OLS estimator for β_2, which is b_2, would be

$$b_2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})y_i + \overline{y}\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}$$

Here I have used the notation $x = inc$, $y = bill$ in order to simplify the equations. To investigate if our estimator is biased, we take the conditional expectation with X (the vector of x_i's) fixed.

$$E[b_2|X] = E\left[\frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}\right] = E\left[\frac{\sum(x_i - \overline{x})(\beta_1 + \beta_2 x_i + \beta_3 x_i^2 + e_i)}{\sum(x_i - \overline{x})^2}\right]$$

x_i^2 is used to denote the values of inc_i^2. Since X is fixed, we have

$$= \beta_1 \frac{\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + \beta_2 \frac{\sum(x_i - \overline{x})x_i}{\sum(x_i - \overline{x})^2} + \beta_3 \frac{\sum(x_i - \overline{x})x_i^2}{\sum(x_i - \overline{x})^2} + \frac{\sum(x_i - \overline{x})E[e_i|X]}{\sum(x_i - \overline{x})^2}$$

In particular, the expectation passes through the x_i^2 observations because those are entirely determined by the x_i's, which are of course fixed. Since we have assumed the form of the true model, we have $E[e_i|X] = 0$. The sum $\sum(x_i - \overline{x}) = n\overline{x} - n\overline{x} = 0$, so the first term disappears. The second term is just β_2, since $\sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$ and $\sum(x_i - \overline{x})x_i = \sum x_i^2 - n\overline{x}^2$. We are left with

$$E[b_2|X] = \beta_2 + \beta_3 \frac{\sum(x_i - \overline{x})x_i^2}{\sum(x_i - \overline{x})^2}$$

Therefore b_2 is not a consistent estimator of β_2.

6.2 Issue 2

As in Part D, suppose that the customer's age is an important predictor of bill size, but is not included in our regression. Then our estimates may be biased. Consider a simplified model that looks like

$$bill_i = \beta_1 + \beta_2 hhsize_i + \beta_3 age_i + e_i$$

and suppose it is correct. However, we estimate

$$bill_i = \beta_1 + \beta_2 hhsize_i + v_i$$

Recall that b_2, our estimate for β_2, is

$$b_2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})y_i + \overline{y}\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}$$

where we have substituted $x = hhsize$ and $y = bill$ for brevity. Then we substitute $y_i = \beta_1 + \beta_2 x_i + v_i$ to obtain

$$= \frac{\sum(x_i - \overline{x})(\beta_1 + \beta_2 x_i + v_i)}{\sum(x_i - \overline{x})^2} = \beta_1 \frac{\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} + \beta_2 \frac{\sum(x_i - \overline{x})x_i}{\sum(x_i - \overline{x})^2} + \frac{\sum(x_i - \overline{x})v_i}{\sum(x_i - \overline{x})^2}$$

Then the first term disappears since $\sum(x_i - \overline{x}) = n\overline{x} - n\overline{x} = 0$. In the second term we have $\sum(x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$ in the denominator and $\sum(x_i - \overline{x})x_i = \sum x_i^2 - n\overline{x}^2$ in the numerator, so that

$$b_2 = \beta_2 + \frac{\sum(x_i - \overline{x})v_i}{\sum(x_i - \overline{x})^2}$$

But since we have omitted the age variable (abbreviated a_i), the error term $v_i = \beta_3 a_i + e_i$, so that

$$b_2 = \beta_2 + \beta_3 \frac{\sum(x_i - \overline{x})a_i}{\sum(x_i - \overline{x})^2} + \frac{\sum(x_i - \overline{x})e_i}{\sum(x_i - \overline{x})^2}$$

Then $E[b_2|X, A]$ is

$$= \beta_2 + \beta_3 \frac{\sum(x_i - \overline{x})(a_i - \overline{a})}{\sum(x_i - \overline{x})^2} + \frac{\sum(x_i - \overline{x})E[e_i|X, A]}{\sum(x_i - \overline{x})^2}$$

where we have used the fact that adding in $\overline{a}$ is an identity. By assumption the expected error $E[e_i|X, A]$ is 0, and dividing the numerator and denominator of the second term by N (sample size) shows that

$$= \beta_2 + \beta_3 \frac{\widehat{\text{cov}}(x, a)}{\widehat{\text{var}}(a)}$$

So if age and household size are correlated (which is quite possible), our estimate for β_2 is biased.

6.3 Issue 3

Suppose that we have nonexact collinearity between household size ($hhsize$, which we will denote h_i) and the number of items purchased ($items$, which we will denote t_i). As before, suppose that we have a simpler model, that is, we estimate

$$y_i = \beta_1 + \beta_2 h_i + \beta_3 t_i + e_i$$

where y_i denotes the size of the bill. Then, depending on the degree of correlation between the two, we may have high variances for our estimates of β_2 and β_3. Since both cases are similar we examine β_2. Let b_2 be the estimator for β_2, then let X be the matrix of observations, i.e.

$$X = \begin{pmatrix} 1 & h_1 & t_1 \\ 1 & h_2 & t_2 \\ \vdots & \vdots & \vdots \\ 1 & h_n & t_n \end{pmatrix}$$

Similarly let b^T denote the matrix (b_1, b_2, b_3) (so that b is a column vector). Then, we can compute the variance as

$$var(b|X) = var((X^TX)^{-1}X^Te|X) = (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

assuming that $var(e|X) = \sigma^2$. From the textbook (and Problem Set 2), we have that the entry in position $(2,2)$ (which is just the variance of b_2), is

$$var(b_2|X) = \frac{\sigma^2}{(1 - r_{ht}^2)(\sum(h_i - \overline{h})^2)}$$

Therefore if the correlation between household size and items purchased is large, then the r_{ht}^2 term will be large. Then $1 - r_{ht}^2$ is small so that variance of our estimator will be large. It is then possible that we will fail to recognize the significance of some of the coefficients, if they actually are significant.

The preceding discussion is general and can easily be applied to β_3 by substituting b_2 for b_3. It extends to collinearity between arbitrary variables as well. In a later section I discuss how to test for this more thoroughly and what potential solutions are.

6.4 Issue 4

In this section we show why heteroskedasticity would be a problem. Suppose that the true model is

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

where y_i is bill size and x_i is income. We estimate this simpler model for illustrative purposes. Suppose that high income individuals have higher variance in their bill size (as discussed in Part E), so that $var(e_i|X) = \sigma_i^2$, instead of $var(e_i|X) = \sigma^2$. Then the variance of the estimator b_2 for β_2 is as follows.

$$var(b_2|X) = var\left(\frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}\middle|X\right) = var\left(\frac{\sum(x_i - \overline{x})(\beta_1 + \beta_2 x_i + e_i)}{\sum(x_i - \overline{x})^2}\middle|X\right)$$

$$= var\left(\beta_1\frac{\sum(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}\middle|X\right) + var\left(\beta_2\frac{\sum(x_i - \overline{x})x_i}{\sum(x_i - \overline{x})^2}\middle|X\right) + var\left(\frac{\sum(x_i - \overline{x})e_i}{\sum(x_i - \overline{x})^2}\middle|X\right)$$

Then the first term is 0 since the numerator is 0 (as shown earlier). The second term becomes β_2 because the fraction is 1 (as shown in Issue 2), so the variance is 0. The variance passes through the sums of the third term (since X is fixed) so that we get

$$var(b_2|X) = \frac{\sum(x_i - \overline{x})^2\sigma_i^2}{(\sum(x_i - \overline{x})^2)^2}$$

This is then not the same as the variance of b_2 under homoskedasticity. Therefore we will obtain inaccurate standard errors if we continue with OLS regression. This will affect t-tests and other test statistics, so that we might come to the wrong conclusion in certain cases.

7 Part F: Issues (Solutions)

7.1 Issue 1

The possibility of nonlinear explanatory factors is likely (in my personal belief) but it should not be difficult to overcome. I can include many polynomial terms and interactive terms between variables, then use a RESET to find out whether any of the terms have significant nonzero coefficients.

Based on my intuition, I would include inc_i^2 (for the reasons discussed in 4.1). This would give an augmented model of the form

$$bill_i = \beta_1 + \beta_2 inc_i + \beta_3 hhsize_i + \beta_4 dist_i + \beta_5 cash_i + \beta_6 items_i + \beta_7 inc_i^2 + e_i$$

I would then run a t-test with the null hypothesis $H_0 : \beta_7 = 0$ to see if this term should be kept or dropped. To be thorough, I would also try a RESET by estimating

$$bill_i = \beta_1 + \beta_2 inc_i + \beta_3 hhsize_i + \beta_4 dist_i + \beta_5 cash_i + \beta_6 items_i + \beta_7 inc_i^2 + \alpha_1 \widehat{bill_i}^2 + \alpha_2 \widehat{bill_i}^3 e_i$$

to check for other possible nonlinear factors. In this regression the $\widehat{bill_i}$ terms are the results from the first model. I would use an F-test with $H_0 : \alpha_1 = \alpha_2 = 0$ to check for significance of any polynomial terms. If the null hypothesis was rejected, then I would have to go back through the model one term at a time with t-tests to find out which specific terms should be included.

7.2 Issue 2

The simplest solution would be to incorporate the age variable into the data. The source of data discussed in Part C has several tables with a wide variety of variables, which may have data on the age of the customer making the purchase. If so, including that would ensure that the error term is not correlated with household size. Then, instead of estimating the incorrect model

$$bill_i = \beta_1 + \beta_2 hhsize_i + v_i$$

and getting a biased estimator for β_2 because v_i is correlated with $hhsize_i$, we could estimate

$$bill_i = \beta_1 + \beta_2 hhsize_i + \beta_3 age_i + e_i$$

and have errors e_i that are uncorrelated with the explanatory variables. I would need to confirm this by first plotting the residuals against each of the explanatory variables, and then regressing the residuals on them as well to check for correlation.

Another way to deal with this would be to introduce an instrumental variable for household size. A couple of candidates are: amount spent on childcare and number of vehicles in the household. The first is probably highly correlated with household size, and doesn't have a direct impact on bill size, but is probably correlated with age. The second is probably correlated with household size, wouldn't directly affect bill size, and may not be correlated with age. After choosing an instrumental variable, I would then use two-stage least squares (2SLS) with the IV to account for the endogenous regressor. I would be sure to check the strength of the instrument by applying a t-test and asking for a high value (> 3.16) in order to confirm that the IV is a good one.

However, I am ultimately not convinced that age is a good explanatory variable anyways, given that household size is already included in the model. I would need to come up with a convincing story first, which explains the intuition for why age would affect bill size, before inserting new variables.

7.3 Issue 3

First, I would test whether or not collinearity is actually present. This could be done in two ways: by inspecting the estimated standard errors of the regression; and by running auxiliary regressions of the form

$$inc_i = \alpha_1 + \alpha_2 hhsize_i + \ldots + \alpha_5 items_i + v_i$$

The example above is only for concreteness. I would of course run separate auxiliary regressions for each of the explanatory variables. Obtaining a high R^2 value from any of these regressions would give cause for concern.

Second, to deal with collinearity, I would try to get additional data, with the hope of creating more separation between the correlated variables. The Economic Research Service (where I obtained the original data from) has many more data sets relating to the supply and consumption of food, so that I could find relevant data there. An additional strategy would be to play with the model a bit. For example, if household size and number of items purchased have high collinearity, I might remove one of them since the added explanatory power is lower. I might also replace both of them with a new variable, such as items purchased divided by household size.

7.4 Issue 4

I would test for heteroskedasticity first to confirm whether it exists. Initially, I would try a White test:

$$\hat{e}_i^2 = \gamma_1 + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \ldots + \gamma_n z_{ni} + \eta_i$$

The various z's represent the original explanatory variables, their powers, and interactive terms between them. I would run an F-test with $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_n = 0$. If we reject the null hypothesis, I would not yet be convinced of heteroskedasticity, since there could be other reasons for the rejection. To confirm, I would use a Breusch-Pagan test to check whether or not heteroskedasticity is actually the root cause of the rejection, by estimating

$$\hat{e}_i^2 = \gamma_1 + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \ldots + \gamma_n z_{ni} + \eta_i$$

except now, the z's are just the explanatory variables. I could test the null hypothesis that all of the γ's are 0 using the test statistic obtained from an F-test. Alternatively, I could compute NR^2, where N is the sample size and R^2 is the R^2 from running this regression. Then I would compare this to a χ^2 distribution to see if $NR^2 \geq \chi^2_{(1-\alpha),(S-1)}$, where $1-\alpha$ is the desired confidence level (for example, $\alpha = 0.05$ for 95% confidence) and $S-1$ is the degrees of freedom. Combining these two tests with visual examination of the residual plots would likely reveal any serious issues.

Then, depending on the residual plots, I might try different things to mitigate the problem. If the residual plots were obviously following a clear pattern, then I would try to compute the variance as a function of the x_i's and then use generalized least squares (GLS). More likely, however, I won't be able to find a clear functional form, so that I will need to use feasible GLS, which would involve estimating the model using OLS, then using the general form

$$\sigma_i^2 = \exp(\gamma_1 + \gamma_2 z_{2i} + \ldots + \gamma_n z_{ni} + \eta_i)$$

to help find heteroskedasticity. Specifically, I would estimate the regression

$$\log(e_i^2) = \gamma_1 + \gamma_2 z_{2i} + \ldots + \gamma_n z_{ni} + \eta_i$$

This would give me an estimate for w_i, defined to be $w_i = \exp(g_1 + g_2 z_{2i} + \ldots + g_n z_{ni})$ (here g_i is the estimator of γ_i). Then I could estimate the original OLS again but divide all of the terms by $\sqrt{w_i}$. I would need large sample sizes to be confident in this process, but this is something we have, as the data I obtained have tens of thousands of observations.

8 Part G: Empirical Analysis

8.1 Data and Code

As mentioned in Part C, the raw data was acquired from the Food Acquisition and Purchase Survey (FoodAPS), conducted by the USDA. This survey contains over 10,000 cross-sectional observations on a nationally representative sample of U.S. households. I used the CSV data files listed there and manually extracted the variables of interest.

This extraction, as well as the analysis, was conducted in Stata. The cleaned dataset and the full code used to perform the analysis, can be found at the following GitHub repository.

8.2 Analysis

I preface this section with a note. The output of each of the steps I describe below is given in the log file that was submitted along with this paper (and on GitHub). I describe the analysis and results below, and the actual Stata output can be referenced in the log file or reproduced using the do-file.

I began by running the initial OLS model given in Part C, making all of the standard assumptions for the multiple regression model.

1. Linear Model
2. Exogeneity of Error
3. Conditional Homoskedasticity
4. Conditionally Uncorrelated Errors
5. No Exact Collinearity

I assume (1) but then address the possibility that it does not hold by checking Issue 1 from Part D, which examines nonlinear terms. I justify (2) and (4) since there is a large sample and because I believe I have included all relevant variables. I assume (3) and (5) for now, and study them more closely later.

Then, I considered each of the issues I described previously. In particular, I considered nonlinear terms such as squared terms and interactions between variables, and chose to include the ones that made sense theoretically (i.e. I asked myself: would it make sense qualitatively to include this term?). Then I applied OLS to the updated model and removed some of the insignificant variables.

To address the possibility of collinearity, I ran auxiliary regressions on each of the independent variables. As an example, I regressed income against all of the other independent variables. Then, if I observed a high R^2 value (the textbook suggests 0.8 as a cutoff), I would be concerned that the independent variables were too good at explaining the variation in income. However, each of the regressions produced a low value for the R^2, each around 0.03 or less. So I concluded that there did not appear to be any issues.

Finally, I checked for heteroskedasticity through a variety of methods. I created residual plots for each independent variable, and saw a concerning pattern for the distance-to-store variable. Specifically, it appeared that the variance of the residual was higher at lower values of distance (i.e. for people who lived closer to the store). I then formally checked for heteroskedasticity using the White test and the Breusch-Pagan test, both of which supported the idea that heteroskedasticity was present. To combat this, I used feasible generalized least squares (FGLS), and these results were my final output.

As a note, the Stata command I used (hetregress) technically executes something called Harvey's two-step GLS method. However, I manually performed FGLS as well to confirm that Harvey's estimator delivered the same results as the FGLS method.

8.3 Results

Below I give the finalized results, obtained after conducting FGLS and all of the other techniques described above.

Variable	Estimate	Standard Error	t	$P > \|t\|$
Income	.0009434	.000148	6.37	0.000
Household Size	.4620454	.110149	4.19	0.000
Distance	.3504998	.0343416	10.21	0.000
Items	2.462082	.0283513	86.84	0.000
Cash	-4.522312	.4001154	-11.30	0.000
Income^2	-1.26e-08	9.05e-09	-1.39	0.163
Income $\times$ Distance	-.0000179	2.47e-06	-7.27	0.000
Constant	3.610921	.5581717	6.47	0.000

Each coefficient estimate was highly significant, except for the one for income squared. Many results are as expected, for example, being farther away from the store correlates with spending more on each trip, as does

household size and the number of items purchased. Using cash is associated with a large downward shift in amount spent.

I found it interesting that none of the income-related terms had as much of an effect as I was expecting. I thought that wealthier households would indeed have more expensive tastes, but the results suggest this might not be the case. Of course, it is more than possible that these households do spend more money on food, and they just do it by going to restaurants, which are not covered by the data.

With regard to the questions stated in my original motivation, I would recommend that the government does not provide the aid plan that I had envisioned, but that businesses should consider the gas-based reward program that I thought of.

Overall, there was a lot of data, which I believe contributed to smaller standard errors and stronger hypothesis tests. I am confident in the results shown because of this and because of the methods that we learned in class and that I applied here.

9 Conclusion

In this project, we study some of the determinants of grocery store bill size, using cross-sectional data at the household level. We examine the effects of monthly household income, household size, distance from store to home, number of items purchased, and whether or not cash was used in the transaction. To conduct this analysis, we used a standard OLS regression, then augmented it using nonlinear terms and FGLS. These additional techniques were used to combat model misspecification and heteroskedasticity, respectively. We find that almost all variables included have a significant effect on the size of a grocery store bill.