Karlsruher Institut für Technologie

Problemset 2 - Marketing Analytics

Institute of Information Systems and Marketing (IISM)

Julius Korch, Marco Schneider, Stefan Stumpf, Zhaotai Liu

Last compiled on December 01, 2023

Contents

Preparations

For this task we interpreted that it is first necessary to fully reconstruct the dataset and model used in the paper. Even though in the FAQ we later found out that it was not necessary to rebuild the model 1-to-1 the work was done already. In the following we will describe the approach we took prior to the knowledge of the FAQ.

1. We read through the paper and noted important information that we would later need for the code reconstruction.
2. We converted the Stata code that was provided by the authors to R code. This took the most time but with the evaluation method below we eventually got the right configuration for the dataset and model.
3. To evaluate whether the dataset and the log-log model was correctly reconstructed we used the α_s coefficient and the p-value that is provided in table 3 of the paper. Based on that, the coefficient value for frequency should be $\alpha_s = -0.001$ and the p-value should be 0.006. In the recreation of the log-log model we did in Task 1: Price and Shopping Frequency the summary output of the model gives us a coefficient for frequency of -0.009 and a p value of 0.006 which basically equals the values provided by the authors (the coefficient need to be rounded up). Furthermore the authors provided the α_s values and the p-values for the log-log models including the instruments. Applying our reconstructed model on the data in Task 2: Endogeneity provided us with the same values using the instruments which further assured us that we reconstructed the model.

Using control variables

Even though in the FAQ it was said that the model is expected to only include the frequency: $ln(P) \sim ln(freq)$, because we recreated the original model which uses control variables, we also included those in our models:

$ln(P) \sim ln(freq) + ln(N) + ln(Nmod) + ln(Q)$

where $ln(N)$ describes the log number of UPC codes purchased per month per household, $ln(Nmod)$ describes the log number of product categories purchased per month per household and $ln(Q)$ describes the log number of the quantity index.

Task 1: Price and Shopping Frequency

The authors of the paper want to investigate the relationship between a price index and shopping intensity. They assume a logarithmic relationship between the two variables. Using a log-log model, they aim to capture the diminishing returns of shopping intensity (as more time spent shopping should lead to smaller and smaller price reductions).

In our approach to this task, we recreated the price index and the frequency variable from the paper. We then used the same log-log model as the authors of the paper. This allowed us to confirm that we had accurately replicated the price index and frequency. We then considered additional possible relationships between the two variables and formulated two models to reflect these possibilities. Finally, we compared each of the models with each other to arrive at our final result.

Creating the price index and frequency

For the price index and shopping intensity, we have closely followed the path mentioned in the paper. Shopping intensity is calculated as the average number of shopping trips per month per household (named as shopping frequency). The formula for the price index is given by

$\tilde{p}_m^j \equiv \frac{X_m^j}{Q_m^j}$

where X_m^j is the total expenditure for household j in month m and Q_m^j is the average price a household paid for the same basket of goods in month m. Additionally the price index is normalized to be centered around 1 by taking the average:

$\tilde{p}_m^j \equiv \frac{\tilde{p}_m^j}{\frac{1}{J}\sum_{j'} \tilde{p}_m^{j'}}$

A normalised price index greater than 1 means that a household paid more for its basket than it would have paid at the average prices paid by households in that month for the products in the basket. On the other hand, a normalised price index less than 1 means that a household paid less for its basket than it would have paid at the average prices paid by households in that month for the products in the basket.

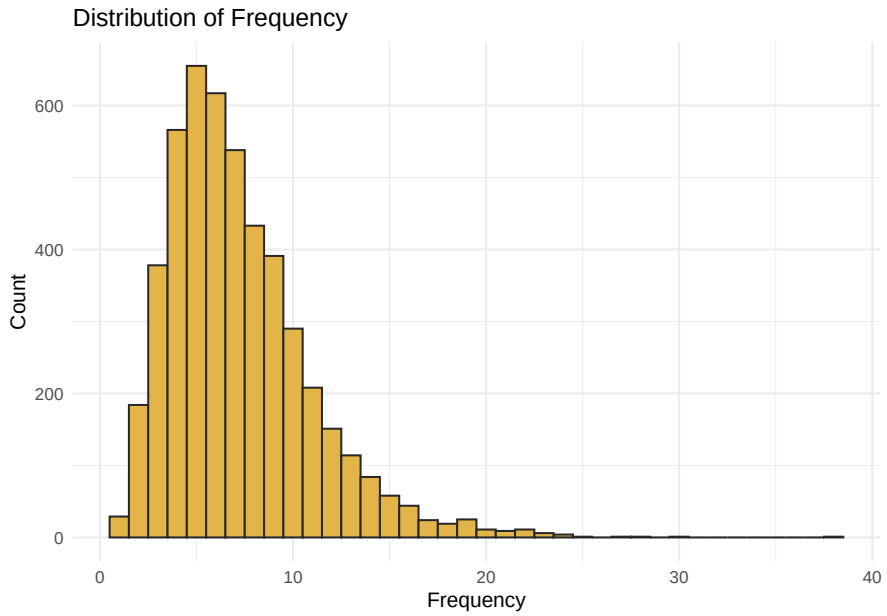
Recreating the log-log model from the paper

After the recreation of the price index and frequency variable, we used the same log-log model as the authors of the paper to verify that our recreation of the price index and frequency variable was correct. This can be seen in the code below.

```
##
## Call:
## lm(formula = lnP ~ ln_freq + lnN + ln_Nmod + lnQ, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69315 -0.03384  0.01601  0.05179  0.49433
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002839   0.006480  -0.438 0.661301
## ln_freq     -0.009072   0.003304  -2.746 0.006064 **
## lnN         -0.047214   0.007413  -6.369 2.08e-10 ***
## ln_Nmod      0.032840   0.008767   3.746 0.000182 ***
## lnQ          0.017252   0.003122   5.526 3.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08299 on 4849 degrees of freedom
## Multiple R-squared:  0.0155, Adjusted R-squared:  0.01469
## F-statistic: 19.09 on 4 and 4849 DF,  p-value: 1.371e-15

## [1] "The alpha_s value should be -0.01 with a p-value of 0.006."

## [1] "The alpha value for our model is  -0.00907"

## [1] "and the p-value is  0.00606 . One can see the"

## [1] "values are almost identical. This results means that a doubling"

## [1] "of shopping frequency lowers prices paid by 1 percent. "
```

Further approaches to investigate the relationship between price and shopping frequency

In order to derive possible relationships between the variables, we first want to have a look at the distribution of the variable shopping frequency:



We can see that most of the observations are around 4-7, indicating that most households shop 4-7 times a month. The distribution also looks normal with a right skew. As the majority of households go less shopping than the mean and because of the long tail of observations with higher shopping frequencies, pulling the mean to a higher value, it is particularly interesting to see whether a high frequency results in a higher or lower price index.

What relationship do we expect?

Approach 1 (linear model):
We assume that households with low shopping frequency pay more for the products in their basket than other households. This is because they bundle their purchases into fewer trips and do not split their purchases as much to buy products that are cheaper elsewhere or at another time in an additional trip. Accordingly, low-frequency households have a higher normalised price index. We expect this price index to fall at a roughly constant rate as frequency increases. This is because households that shop more frequently bundle their purchases less and thus perceive lower prices in another shop or at another time (spread over the month). Since in this approach we assume a constant rate of change of the price index over the purchase reference, we will first use a linear model to examine the relationship between the two variables.

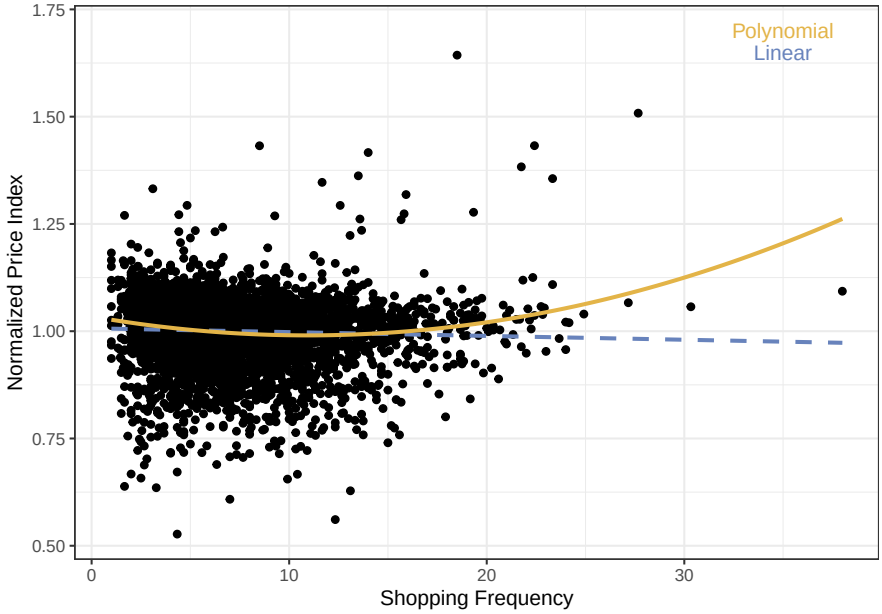
Aguiar and Hurst used a logarithmic model in their paper. However, the logarithmisation was probably done because returns are expected to decrease as purchasing intensity increases, and because the percentage change is more important than the absolute difference (logarithmisation can help to represent proportionally equal percentage changes on the logarithmic scale as constant differences). This results in a linearisation of the data (as described in the lecture). Considering the non-logarithmic data and inserting the linear function further motivated us to analyse approach 1, as the linear function has a negative slope (see blue line in the graph below). This supports our approach 1. It would be a simpler and more straightforward approach than logarithmisation.

Approach 2 (polynomial (quadratic) model):
Another approach is to assume that the relationship between price index and shopping frequency is non-linear. Here it is assumed that households with a lower shopping frequency still have a higher price index. This does not mean that these households consume less, but rather that their purchases are spread over fewer trips. Products that are cheaper elsewhere or at another time are still bought in bundles, without taking advantage of discounts. However, we also assume that households with a very high frequency of shopping tend to buy spontaneously or emotionally in order to satisfy their short-term needs. Accordingly, these households do not pay as much attention to the price of the products and therefore have a higher price index. Households that find a good middle ground in terms of shopping frequency, take advantage of discount opportunities

that require additional trips and do not tend to make emotional purchases, have lower price indices. This means that households with an average frequency of purchase represent a kind of minimum in terms of the price index. Since we expect a non-linear relationship here, we assume a slightly U-shaped price-frequency function. Accordingly, our second approach is based on a polynomial (quadratic) model.

As explained above, we expect the data to follow a slightly U-shaped price-frequency function, which would correspond to reverse supersaturation. By inserting the quadratic function into the plot (see the orange line in the plot below), we can also see this. This encourages us to continue with this approach.

```
## `geom_smooth()` using formula = 'y ~ x'
```



Both of our approaches (linear: constant rate of change; polynomial: reverse supersaturation) would contradict the assumption of Aguiar and Hurst that there are diminishing returns as shopping intensity/frequency increases. By simply looking at the data points, we are of the opinion that the authors' assumption does not necessarily have to be present in the underlying data set. Thus, it is even more interesting, if we can find significant effects with our models.

Results:

By running the linear model, we receive the following results:

```
##
## Call:
## lm(formula = p_index_norm ~ freq + N + Nmod + Q, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47835 -0.03567  0.01242  0.04853  0.63707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.007e+00  3.642e-03 276.585  < 2e-16 ***
## freq        -1.664e-04  4.033e-04  -0.413  0.67997
## N           -2.398e-03  3.544e-04  -6.767 1.47e-11 ***
## Nmod         2.439e-03  7.738e-04   3.153  0.00163 **
## Q            1.671e-04  3.698e-05   4.518 6.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.07766 on 4849 degrees of freedom
## Multiple R-squared:  0.01584,    Adjusted R-squared:  0.01503
## F-statistic: 19.52 on 4 and 4849 DF,  p-value: 6.013e-16
```

The $alpha_s$ value of the linear model is -0.00017 (rounded) with an according $p - value$ of 0.67997. This means that it cannot be determined that the coefficient is not equal to 0. In other words, by using the linear model, we cannot find a significant relationship between the price index and frequency. This contradicts our hypothesis from Approach 1 that the price index decreases on a constant rate with increasing shopping frequency. Interesting is now to analyse whether evidence can be found that the price index increases again for comparably higher shopping frequencies.

When running the polynomial model, we receive following results:

```
##
## Call:
## lm(formula = p_index_norm ~ poly(freq, 2) + N + Nmod + Q, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47860 -0.03570  0.01225  0.04776  0.62499
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.985e-01  3.902e-03 255.917  < 2e-16 ***
## poly(freq, 2)1 -1.198e-01  1.037e-01  -1.156 0.247727
## poly(freq, 2)2  6.143e-01  8.084e-02   7.598 3.58e-14 ***
## N              -2.443e-03  3.523e-04  -6.933 4.66e-12 ***
## Nmod            3.472e-03  7.812e-04   4.444 9.03e-06 ***
## Q               1.399e-04  3.694e-05   3.788 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07721 on 4848 degrees of freedom
## Multiple R-squared:  0.02743,    Adjusted R-squared:  0.02642
## F-statistic: 27.34 on 5 and 4848 DF,  p-value: < 2.2e-16
```

The $alpha_s$ value of the polynomial (quadratic) model is 0.6143 (rounded) with an according $p - value$ of $3.58e - 14$. By using this model, a highly significant relationship can be found between the price index and frequency variables. This supports our assumption of the second approach that households with medium frequency have the lowest price index and households with comparatively low/high frequency have to pay more (high price index). Interestingly, however, this contradicts the observations of Aguiar and Hurst. Using our approach suggests that the price index for increasing frequencies initially falls and then rises again. Aguiar and Hurst's approach suggested that the price index falls by one percent when the frequency doubles.

Conclusion

Finally, we can conclude from our analysis that we were unable to find a significant linear relationship between the variables price index and shopping frequency. However, we were able to find evidence of a polynomial relationship, which indicates that the price index initially decreases for increasing shopping frequency and increases again after a certain shopping frequency (with a minimum price index). This result is not consistent with the results of Aguiar and Hurst, who concluded from their analysis that the price index decreases with increasing purchase frequency. A possible reason for this could be that the relationship between the two variables is more complex than represented by the models and may be influenced by various factors (e.g. the used control variables). Based on the R-squared of the authors' log-log model and our polynomial model, we can also see that the models explain only a small part of the variance in the price index and are therefore not very powerful. Why this is so should be investigated further.

Task 2: Endogeneity

Endogeneity in the shopping frequency-price function

Endogeneity refers to a situation in which one or more explanatory variables in a statistical model are correlated with the error term (residuals) of the model.

In their paper, Aguiar and Hurst argue that the shopping frequency-price function suffers from endogeneity because the shopping productivity or the shopping skill is omitted in the model. Basically that means that the influence of shopping frequency on the price paid is biased by the fact that different people have different shopping skills and for some it might be easier to find the lowest price for a product and therefore those people need less shopping trips and still pay lower prices. So in this case the OLS estimates of the effect of shopping frequency on price may be wrong, directly speaking biased downwards because the shopping productivity is not included in the model, but a higher shopping productivity leads to fewer shopping trips but also lower prices paid.

Another reason for endogeneity in this case could be a measurement error in the shopping time, as the authors point out. This is because shopping frequency does not really capture the time spend in the store. For example, there could be households which shop more frequently but spend less time on each trip and some households which shop less frequently but spend more time doing it. When we assume that on average the time spent shopping is the same for these two groups when calculating the shopping frequency-price function, including only the shopping frequency could lead to an endogeneity problem because the actual time spent is omitted.

To isolate the effect of shopping frequency on price, the authors use 3 different instrumental variables (age, income, family size) which predict the shopping frequency but are not influenced by the shopping productivity. Since we only have to choose 2 instruments for our analysis, we decided to use income and family size because we would expect age to be influenced by shopping productivity as well, since older people have more experience in shopping and therefore they are more "skilled" at it.

Instrumental variables

Model without accounting for endogeneity (from Task 1)

First, we compute the results for the model without taking endogeneity into account, in order to be able to compare the result of this model with the models using instrumental variables.

```
##
## Call:
## ivreg(formula = lnP ~ ln_freq + lnN + ln_Nmod + lnQ, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69315 -0.03384  0.01601  0.05179  0.49433
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002839   0.006480  -0.438 0.661301
## ln_freq     -0.009072   0.003304  -2.746 0.006064 **
## lnN         -0.047214   0.007413  -6.369 2.08e-10 ***
## ln_Nmod      0.032840   0.008767   3.746 0.000182 ***
## lnQ          0.017252   0.003122   5.526 3.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08299 on 4849 degrees of freedom
## Multiple R-Squared: 0.0155,  Adjusted R-squared: 0.01469
## Wald test: 19.09 on 4 and 4849 DF,  p-value: 1.371e-15
```

Income

Arguments for using income as instrumental variable

One of the instruments the authors use is income or wages. They argue that a higher wage leads to fewer shopping trips, which they have already stated earlier in the paper. First, we calculated the correlation between the shopping frequency and the categorical income variable to test whether income is indeed able to predict the shopping frequency of the household. The results of the correlation are shown below:

```
## [1] -0.0208446
```

In this case, the correlation is quite small but has a negative sign, confirming the negative relationship between shopping frequency and income, so that a household in a higher income category makes fewer shopping trips than a similar household in a lower income category. A possible explanation for this effect could be that people with higher incomes tend to work more and longer to earn that income and therefore have less time to shop and make fewer trips. So, at first sight, income is a sufficiently relevant instrument for predicting shopping frequency.

Another requirement for the instrumental variable is that it is independent of the omitted variable, in this case shopping productivity or shopping skill. Since income and shopping productivity are unrelated, we would not expect shopping productivity to influence income, and therefore income should isolate the exogenous variation in the indigenous variable (shopping frequency).

Step 1 - Regress shopping frequency on the instrument

First, we regress the frequency of shopping on the instrument, in this case the income categories of the households. The results of the regression are shown below:

```
##
## Call:
## lm(formula = ln_freq ~ inc_cat + lnN + ln_Nmod + lnQ, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68824 -0.21527  0.01682  0.23605  1.17859
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.060092   0.028651  -2.097    0.036 *
## inc_cat     -0.051878   0.005356  -9.686  < 2e-16 ***
## lnN          0.493209   0.031142  15.838  < 2e-16 ***
## ln_Nmod     -0.266988   0.037551  -7.110 1.33e-12 ***
## lnQ          0.318001   0.012721  24.999  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3572 on 4849 degrees of freedom
## Multiple R-squared:  0.5467, Adjusted R-squared:  0.5463
## F-statistic:  1462 on 4 and 4849 DF,  p-value: < 2.2e-16
```

We can conclude from the p-values that each regressor was significant and the R-squared value is 54.63%.

Step 2 - Compute the predicted shopping frequency

With the results of the regression analysis of the first step, we are able to compute the predicted shopping frequency with the following formula: $predictedfrequency = -0.060092 - 0.051878 * inccat + 0.493209 * lnN - 0.266988 * lnNmod + 0.318001 * lnQ$

Step 3 - Replace shopping frequency with the predicted shopping frequency

Now we replace the shopping frequency with the results from step 2 in the regression. The result of this regression analysis is shown below:

```
##
## Call:
## lm(formula = lnP ~ frequencyhatincome + lnN + ln_Nmod + lnQ,
##     data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69502 -0.03359  0.01653  0.05195  0.48401
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.010705   0.007135  -1.500  0.13359
## frequencyhatincome -0.071580   0.023981  -2.985  0.00285 **
## lnN                -0.016578   0.013801  -1.201  0.22971
## ln_Nmod             0.016305   0.010785   1.512  0.13065
## lnQ                 0.036634   0.007999   4.580 4.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08298 on 4849 degrees of freedom
## Multiple R-squared:  0.01578,    Adjusted R-squared:  0.01497
## F-statistic: 19.43 on 4 and 4849 DF,  p-value: 7.035e-16
```

The regression coefficient of shopping frequency using instrumental variables is -0.07, which is the same result the authors found in their analysis. This means that doubling the frequency of shopping reduces the price paid by 7% instead of only 1% when using real shopping frequency. In other words, the effect of shopping frequency on price increases when income is used as an instrumental variable for shopping frequency.

With these 3 steps, the standard errors are too small because the uncertainty in predicting 'predicted shopping frequency' is ignored. Therefore, we also performed a "two stage least squares" (2SLS) estimation. The result of this estimation is shown below:

```
##
## Call:
## ivreg(formula = lnP ~ ln_freq + lnN + ln_Nmod + lnQ | inc_cat +
##     lnN + ln_Nmod + lnQ, data = df_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68625 -0.03686  0.01296  0.05165  0.56414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.010705   0.007395  -1.448  0.14778
## ln_freq     -0.071580   0.024854  -2.880  0.00399 **
## lnN         -0.016578   0.014303  -1.159  0.24649
## ln_Nmod      0.016304   0.011177   1.459  0.14471
## lnQ          0.036634   0.008290   4.419 1.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.086 on 4849 degrees of freedom
## Multiple R-Squared: -0.05716,    Adjusted R-squared: -0.05803
## Wald test: 18.09 on 4 and 4849 DF,  p-value: 9.213e-15
```

The results for the regression coefficients are the same but we can see that the standard errors of those estimates increased a bit.

Hausman Test

To be able to state whether the basic model suffered from endogeneity we conduct a Hausman test. This is used to investigate whether the difference of the models arise from endogeneity. Therefore, we used the following hypotheses: $H_0 : \beta_{IV} = \beta_{OLS}$
This basically means that the regression coefficients of the two models are the same and there is no endogeneity or it is not solved by the instrument used. The alternative hypotheses than looks like this: $H_1 : \beta_{IV} \neq \beta_{OLS}$
This means that the estimates are different from each other and therefore the basic model suffered from endogeneity (when a suitable instrumental variable is used).

If H_0 is true than the test statistic is distributed according to the Chi-Squared-distribution with $df = 1$.

So now we can compute the test statistic using the estimated regression coefficients and their standard errors. By inserting them in the formula we get the following results:

```
## [1] 6.439055
```

Now we compute the p-value of the test to determine whether we can reject the H_0 hypothesis.

```
## [1] 0.0111638
```

The p-value of the Hausman test is 1,12% so assuming a significance level of 5% we can reject the H_0 hypothesis and therefore conclude that the regression estimates are significantly different from each other. That also means that our basic model probably suffered from endogeneity (when we assume income as exegenous predictor for shopping frequency).

Problems when using income categories as instrumental variable

We have already argued that income is unlikely to be influenced by the omitted variable "shopping frequency". But using income leads to a different problem because we would assume that income itself has an influence on the price paid and therefore is not only a predictor of frequency but also contains more information. For example, people with higher incomes are likely to be less price sensitive, so they may not care if they pay higher prices, or they may just put less effort into finding the lowest price. So we would expect income itself to be correlated with the error term in the model.

This means that income is likely to be an endogenous instrumental variable and therefore the effect of frequency on price is still biased.

Another reason for using income as an instrumental variable is given by the authors themselves. They argue that people with a higher income are more likely to own a car and therefore may shop more frequently as it is easier for them to get to the market. In this case, the omitted variable could be "shopping technology", which makes it easier for people with high income to shop more frequently. If this effect is present, it would also bias the result of the regression analysis.

In conclusion, we believe that income is not a good predictor of shopping frequency because it contains much more information and is therefore likely to be an instrumental variable that itself suffers from endogeneity.

Family Size

Arguments for using household size as instrumental variable

Another instrument the authors use to predict shopping frequency is family size. They create a dummy variable "corresponding to households with 1, 2, 3, 4, 5 or 6+ individuals". Their reason for choosing this instrument is that "a shopper with more children faces a higher opportunity cost of time". In other words, people with more children have to spend more time looking after them. For example, they have to do more

laundry, spend more time driving them to leisure activities or simply helping them with their homework. This reduces the time available for other activities, including shopping. Controlling for shopping needs (which are likely to be higher in larger households), the authors find that "living in a larger household significantly reduces shopping frequency". So, conditional on shopping needs, which are likely to be higher for larger households, these households make fewer shopping trips. By running a more specific analysis the authors state that for example "households with three individuals shop approximately 14 percent less than singles and 10 percent less than households with two individuals". All those arguments would imply a negative relationship between the household size dummies and the shopping frequency, when also controlling for the needs and therefore household size could be a potential instrumental variable for shopping frequency.

The other requirement for the instrumental variable is that it is independent of the omitted variable, in this case shopping productivity or shopping skill. Here one could argue both ways. At first sight, household size does not seem to be affected by shopping productivity because the decision to have another child is probably not based on one's own shopping efficiency. However, it could be argued that people with higher shopping productivity have more free time, which could be used to raise a child. However, as having and raising a child is a very time-consuming and complex task, the effect of shopping productivity on this is likely to be very small, if not zero.

Regression with household size as instrument

Since we have already carried out the 3 steps required for income as instrumental variable in detail, we decided to directly run the "two stage least squares" (2SLS) estimation, as it gives the same results but with correct standard errors.

```
##
## Call:
## ivreg(formula = lnP ~ ln_freq + lnN + ln_Nmod + lnQ | hhsize_cat +
##     lnN + ln_Nmod + lnQ, data = df_filtered, robust = TRUE, cluster = ~panid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68660 -0.03648  0.01343  0.05138  0.56065
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.010311   0.007105  -1.451 0.146779
## ln_freq     -0.068450   0.019276  -3.551 0.000387 ***
## lnN         -0.018112   0.012045  -1.504 0.132716
## ln_Nmod      0.017132   0.010352   1.655 0.097989 .
## lnQ          0.035664   0.006708   5.316 1.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08571 on 4849 degrees of freedom
## Multiple R-Squared: -0.05006,    Adjusted R-squared: -0.05093
## Wald test: 19.28 on 4 and 4849 DF,  p-value: 9.451e-16
```

Now the regression coefficient for shopping frequency when using household size as instrumental variable is -0.68 with a standard error of 0.019.

Hausman Test

In order to determine whether the basic model suffers from endogeneity, we again perform a Hausman test with the following hypotheses:

$H_0 : \beta_{IV} = \beta_{OLS}$

$H_1 : \beta_{IV} \neq \beta_{OLS}$

If H_0 is true than the test statistic is distributed according to the Chi-Squared-distribution with $df = 1$.

We can now calculate the test statistic using the estimated regression coefficients and their standard errors. Substituting them into the formula we obtain the following results:

```
## [1] 9.776151
```

Again with this result we compute the p-value of the test to determine whether we can reject the H_0 hypothesis.

```
## [1] 0.0017679
```

The p-value of the Hausman test is 0,18%. So, assuming a significance level of 5%, we can reject the H_0 hypothesis and therefore conclude that the regression estimates are significantly different from each other. This also means that our basic model probably suffered from endogeneity (when we assume household size as exegenous predictor for shopping frequency).

Problems when using household size categories as instrumental variable

Earlier we said that household size is unlikely to be affected by shopping productivity. But if we look again at this assumption, we are not sure that this is really the case. People with more children are likely to be more efficient in general, as they have more needs to meet and have adapted their lifestyle to meet them all. This could also lead to higher shopping productivity. In other words, a larger household size could lead to higher shopping productivity and therefore lower prices paid, even though they make fewer shopping trips. If this is true the results of the regression analysis using instrumental variables would still be biased downwards and suffer from endogeneity.

Another problem with using household size categories could be that any household with 6 or more people living in it is put in the same category for the analysis. However, according to the authors' assumption, households with 7 or 8 members would actually make fewer shopping trips than households with 6 members, controlling for household needs. By grouping them together, we argue that this could lead to an overestimation of the effect of shopping frequency on price paid.

To conclude, we would again argue that household size is an endogenous variable because it affects shopping productivity, which in turn affects the price paid. Therefore, we would not recommend using household size as an instrument for shopping productivity.

Alternative instrumental variables or ideas to solve the endogeneity problem

Although both Hausman tests indicate that we have solved the endogeneity problem by omitting the shopping productivity variable, we argued that both (income & household size) themselves suffer from endogeneity. Therefore, we thought about different instrumental variables that could be used to predict shopping frequency.

One idea was to use the working hours instead of the income as an instrument, if it was availabe. One reason for this idea would be that working hours are not always correlated with income, but offer more information about how much time people have left for shopping and therefore also about the opportunity cost of time. But then income should also be used in the model because we still would expect an effect of income on the price paid (as stated in earlier section, people with a higher income probably care less about the price paid for some products).

Another possible instrumental variable, if the data were available, to predict shopping frequency could be how much time the household members spend out of the house, for example at school, at work or just for leisure activities. This is because we would expect that if members spend more time away from home, their shopping needs will decrease, for example because the children get a meal at school, so the household needs less food at home and therefore needs to make fewer shopping trips to satisfy their needs at home.

Other endogeneity problems

We would expect another endogeneity problem when calculating the effect of shopping frequency on price, because the time spent preparing is not taken into account. This is because we think that if people spend more time reading store magazines and looking for the best offers and discounts before they actually go shopping, this would lead to fewer trips and lower prices. We think this would reduce the number of shopping trips because people would be able to plan their trips more carefully and buy products not because of an urgent need but because of a low price at the moment. For example, a family goes to buy some groceries, but when they look in the store magazine, they see that there is a discount on toilet paper. So instead of buying the toilet paper later, they decide to buy it now at the lower price. This could then replace the trip to the store later because they have enough of it at home. In this example, the frequency of shopping and also the prices paid would go down, which implies another endogeneity problem in the regression.

Another idea could be to use the preparation time for shopping trips as a control variable. We would argue that preparation time is likely to be highly correlated with shopping productivity, because if people prepare for their trips (for example, by reading the store magazines and planning their trips according to the discounts of the marketplaces in a particular time frame), they will be more productive when they actually go shopping. The authors mentioned that they had this data but only at an aggregate level and therefore decided to exclude it from their analysis, but we think that if this data were available at a more specific level, it might be worth including it in the model and capturing shopping productivity through it.

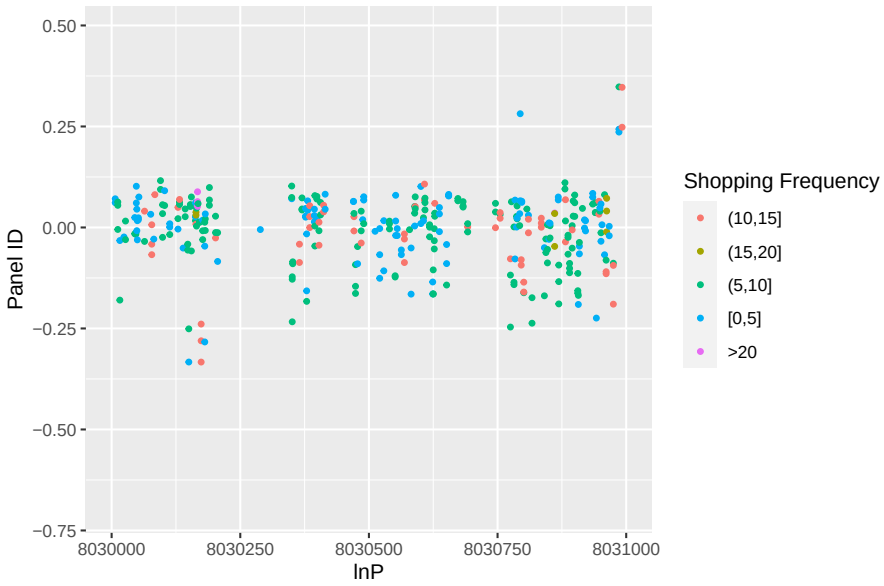
Task 3: Clustered data

The authors of the paper used clustering at the household and product levels. But clustering can become a problem in regression analysis due to the violation of the assumption that errors in the regression model are uncorrelated. Which basically means that the basic assumption, that the error (the differences between the predicted values and the actual values) of one observation is not related to the error of another observation. In this particular clustering example however, if you are looking at household H_a, the errors in predicting the price index on a particular product level P_1 might be related to the errors of another particular product level P_2 because both observations are from the same household H_a. In simple terms this basically means that errors in the same household might be similar, which can mess up the accuracy of the regression model's predictions.

Fixed effects regression can help to address this issue. This method is particularly useful if you're interested in analysing unique characteristics of each cluster that don't change over time.

To model the data at the household level, we decided to cluster the data using the panel ID, which remains constant in the dataset. We also considered the approach of using OLS with a dummy variable, but this is only suitable for a small number of groups, but if we use the panel ID for clustering we have 2056 different groups (because there are 2056 households in the panel). So we decided to use a fixed effects regression. We would expect the errors within a household to be more correlated because the omitted variable in task 2 (shopping productivity) is likely to be constant for a household over the observed period. Therefore, we believe that by grouping at the household level we are able to reduce the omitted variable bias.

To get an idea of how the fixed effects regression affects the results, we used a sample of households and looked at the price they paid conditional on the number of shopping trips for each year:



Looking at the graph, the first thing we can conclude is that in this sample there does not seem to be a relationship between the frequency of shopping and the price paid. It also appears that for a household the frequency does not seem to change much between the different time periods. Because of this, and the argument that the errors within households are correlated due to the omitted variable (shopping productivity), we would expect the results of the fixed effects regression analysis to be less significant, or perhaps there is no significant effect of shopping frequency on the price paid.

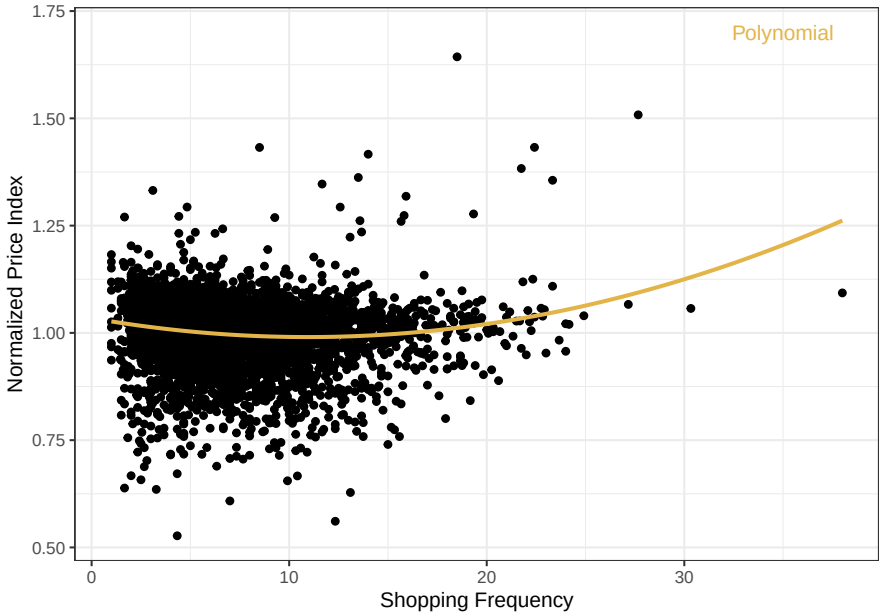
When we run the fixed effects regression model we are obtaining the following results:

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lnP ~ ln_freq + lnN + ln_Nmod + lnQ, data = panel_data,
##     model = "within")
##
## Unbalanced Panel: n = 2056, T = 1-3, N = 4854
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.362725 -0.015239  0.000000  0.015704  0.247099
##
## Coefficients:
##           Estimate Std. Error t-value Pr(>|t|)
## ln_freq  0.0164232  0.0050142  3.2754 0.001068 **
## lnN     -0.0094243  0.0091770 -1.0269 0.304533
## ln_Nmod  0.0014758  0.0098344  0.1501 0.880723
## lnQ     -0.0124200  0.0052307 -2.3744 0.017643 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.3289
## Residual Sum of Squares: 5.2882
## R-Squared:      0.0076344
## Adj. R-Squared: -0.72368
## F-statistic: 5.37364 on 4 and 2794 DF, p-value: 0.00026138
```

Interestingly the sign of the coefficient for the frequency changed and now has a value of +0.016. But by looking at the p-value the frequency still has a significant influence on the price paid. However in this model only one of the control variable is still significant, compared to the initial model where all of them had a

significant influence on the price paid.

Contrary to the authors' log-linear model, here we have a positive influence of shopping frequency on the price paid. In other words, households tend to pay a higher price when they shop more frequently. This is also supported by our analysis of the polynomial model in task 1, where we could see that shopping frequency increased when a certain threshold of shopping frequency was reached:



The result of this analysis and the fixed effect regression model are contradictory to the analysis of the authors. So we thought about possible reasons why there could be also a positive relationship between the shopping frequency and the price paid:

1. **Convenience Shopping**: More frequent shopping may indicate a preference for convenience. These consumers may be more likely to shop at convenient stores close to home or on the road, where prices may be higher than in large supermarkets or wholesale stores.
2. **Impulse Buying**: Frequent shopping opportunities may increase the odds of impulse purchases, which often occur without price consideration and may result in paying a higher price overall.
3. **Income level**: Families with better economic conditions may shop more frequently and may also pay less attention to prices, thus paying higher prices when shopping.

Conclusion

Taking into account all the different analyses, we found evidence of a negative relationship between shopping frequency and price paid, but also of a positive relationship. Therefore, we would conclude that the effect of frequency on price paid depends on other variables. These are on the one hand the control variables that seem to drive the results of the regression, but on the other hand also omitted variables such as shopping productivity that are not part of the data set and the analyses. Overall, therefore, we cannot identify a clear effect of shopping frequency on the price paid and the relationship seems to be more complex. We would therefore suggest including other variables in further analyses to investigate which factors really influence the price paid.