HW 3

David Gao

2022-10-18

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from  
##   format.tbl pillar
##   print.tbl  pillar
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.7
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)

oj<- read.csv("C:\\UW\\AUT 2022\\ECON 487\\HW 2\\oj.csv")
```

Theoretical Questions:

1. Go back to lecture 2's slides on Value Based Pricing. List each type of value-based pricing (e.g., 2-part tariffs, bundling, etc.).
    a. Give a one sentence definition of each.
        - Perfect price discrimination: Charge each consumer her willing to pay.
        - Indirect price discrimination: Offer multiple versions to all and allow consumers to self select.
        - Two Part Tariff: A flat fee and a unit price.
        - Direct price discrimination: Different prices based on customer characteristics.
    b. Try to think of one practical problem with implementing each type of value based pricing. This could be either competitive (e.g., competing firms) or information deficiencies.

        - Perfect price discrimination: car salesman who try to maximize car price for each consumer based on their willingness to pay.
        - Indirect price discrimination: mobile phone with different storage, such as 64GB, 128GB and 256GB.
        - Two Part Tariff: credit cards which charge an annual fee plus a per-transaction fee.
        - Direct price discrimination: gas station charge based on how much gasoline is charged for each consumer.

2. Assume that in addition to orange juice, you also observe demand for bananas.
    a. What regression would you run to determine if bananas and orange juice are compliments or substitutes? What is the coefficient of interest (i.e. on what variable) that would inform you?

        I might run a regression on quantity of orange juice and price of bananas ($Q_{organejuice}$ = $\beta \cdot P_{bananas}$ + a), and a positive β appears, then they are substitute.

    b. Assume you find they are substitutes. What would the sign of the coefficient be? Would you be more or less likely to bundle these products if they are substitutes?

        i. Explain why with an equation, figure or a sentence or two.

            For equition $Q_{organejuice}$ = $\beta \cdot P_{bananas}$ + a, if a positive β appears, then they are substitute, as more sales in bananas leads to less sales in orange juice. And we are less likely to use bundle, as if both of them are in bundle, we might lost both consumer groups that prefer either orange juice or bananas.

        ii. Would the price of the bundle be less than or more than the sum of the two independent prices? (Not a trick question; verifying you understand bundles.)

            If they are substitute, we might want to make the bundel price less than the sum of independent price.

    c. During a sale for orange juice, should you continue to offer the bundle? Why or why not? HINT: who is price sensitive for orange juice? Who comes into market? Would you want to offer the bundle at a lower price than before?

        Sale in orange juice might be more attactive to people who likes orange and does not prefer bananas. Then, we should not use bundle as this might have negative influence on sales of orange juice.

Empirical Section

1. Let’s return to the orange juice assignment and investigate how store demographics are related to demand.

    a. Take one of the final models from HW2 (logmove ~ log(price)*brand*feat) and add in the store demographics as linear features (e.g. + demo1 + demo2). Report your output (past into your answer document).

```
model1 <- lm(logmove ~ log(price)*brand*feat+store, data = oj)
summary(model1)
```

```
##
## Call:
## lm(formula = logmove ~ log(price) * brand * feat + store, data = oj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9211 -0.4280 -0.0112  0.4149  3.2076
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      10.3276122  0.0251152 411.209  < 2e-16 ***
## log(price)                       -2.7735297  0.0387819 -71.516  < 2e-16 ***
## brandminute.maid                  0.0460163  0.0465721   0.988    0.323
## brandtropicana                    0.7083248  0.0507346  13.961  < 2e-16 ***
## feat                              1.0952023  0.0380578  28.777  < 2e-16 ***
## store                             0.0009716  0.0001147   8.473  < 2e-16 ***
## log(price):brandminute.maid       0.7841704  0.0613293  12.786  < 2e-16 ***
## log(price):brandtropicana         0.7352477  0.0567667  12.952  < 2e-16 ***
## log(price):feat                  -0.4720220  0.0740011  -6.379 1.81e-10 ***
## brandminute.maid:feat             1.1724400  0.0818557  14.323  < 2e-16 ***
## brandtropicana:feat               0.7828053  0.0986278   7.937 2.15e-15 ***
## log(price):brandminute.maid:feat -1.1077365  0.1220970  -9.073  < 2e-16 ***
## log(price):brandtropicana:feat   -0.9833848  0.1239575  -7.933 2.21e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6941 on 28934 degrees of freedom
## Multiple R-squared:  0.5365, Adjusted R-squared:  0.5364
## F-statistic:  2791 on 12 and 28934 DF,  p-value: < 2.2e-16
```

```
b. What demographics significantly (t-value>2) influence demand?
    Looks like the variable **feature** and **store** are signigicantly influencing demand.

c. Use the predict command to determine how well the model predicts logmove and create a new var
iable called logmove_hat. To do so construct the “fair r2” covered in class.  What is the improv
ement relative to the model without the demographic features?
```

```
model2 <- lm(logmove ~ log(price)*brand*feat, data = oj)
logmove_hat2 <- predict(model2, newdata = oj)
logmove_hat1 <- predict(model1, newdata = oj)
RSS1 <- sum((oj$logmove - logmove_hat1)^2)
RSS2 <- sum((oj$logmove - logmove_hat2)^2)
TSS <- sum((oj$logmove - mean(oj$logmove))^2)
n <- nrow(oj)
fair_R1 <- 1 - ((n - 1)/(n - 4 - 1))*(RSS1/TSS)
fair_R2 <- 1 - ((n - 1)/(n - 3 - 1))*(RSS2/TSS)
fair_R1
```

```
## [1] 0.5364797
```

```
fair_R2
```

```
## [1] 0.5353457
```

```
    The prediction accuracy is slightly improved in new model with demographic varaible, as fair
$R^2$ is larger.

d.i & ii. Create a new dataframe which is a random subset of 80% of the data (look at sample_n f
rom the dplyr package).
```

```
oj_temp <- oj %>%
  mutate(binom = rbinom(n, 1, 0.8))

oj80 <- oj_temp %>%
  filter(binom == 1)

oj20 <- oj_temp %>%
  filter(binom == 0)

model3 <- lm(logmove ~ log(price)*brand*feat+store, data = oj80)
model4 <- lm(logmove ~ log(price)*brand*feat, data = oj80)

logmove_hat3 <- predict(model3, newdata = oj20)
logmove_hat4 <- predict(model4, newdata = oj20)

MSE3 = sum((oj20$logmove - logmove_hat3)^2)/n
MSE4 = sum((oj20$logmove - logmove_hat4)^2)/n
```

```
d.ii. Estimate the model with and without demographic characteristics.  Construct MSE for the tr
aining and test set for the models.

     for model with demographic variable is 0.0950531, and MSE for model without demographic var
iable is 0.0952082.

d.iii. Compare the out of sample MSE for the models.  Which is lower implying the model does a b
etter job of fitting the data?

    The new model has slightly lower MSE, which means it does a better job.
```

2. Let’s focus on two variables HHLARGE (“fraction of households that are large”) and EDUC (“fraction of shoppers with advanced education”).

    a. What are the means and percentiles of each of these variables? HINT: summary(oj$EDUC)

```
summary(oj$EDUC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.04955 0.14598 0.22939 0.22522 0.28439 0.52836
```

```
summary(oj$HHLARGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01351 0.09794 0.11122 0.11560 0.13517 0.21635
```

```
b. Using your coefficient estimates from the regression in 1b:

    i. If we move from the median value of HHLARGE to the 75th percentile (3rd quartile), how mu
ch does log(quantity) change each week on average?
```

```
model5 <- lm(logmove ~ log(price)*brand*feat + store + HHLARGE + EDUC, data = oj)
summary(model5)
```

```
##
## Call:
## lm(formula = logmove ~ log(price) * brand * feat + store + HHLARGE +
##     EDUC, data = oj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9968 -0.4195 -0.0195  0.4044  3.3015
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      10.7286138  0.0323278 331.870  < 2e-16 ***
## log(price)                       -2.7948186  0.0384389 -72.708  < 2e-16 ***
## brandminute.maid                  0.0648495  0.0461549   1.405     0.16
## brandtropicana                    0.7284359  0.0502798  14.488  < 2e-16 ***
## feat                              1.0854255  0.0377115  28.782  < 2e-16 ***
## store                             0.0016777  0.0001179  14.234  < 2e-16 ***
## HHLARGE                          -3.4481980  0.1511183 -22.818  < 2e-16 ***
## EDUC                             -0.2103701  0.0402347  -5.229 1.72e-07 ***
## log(price):brandminute.maid       0.7682989  0.0607728  12.642  < 2e-16 ***
## log(price):brandtropicana         0.7264723  0.0562485  12.915  < 2e-16 ***
## log(price):feat                  -0.4570954  0.0733258  -6.234 4.62e-10 ***
## brandminute.maid:feat             1.1553809  0.0811095  14.245  < 2e-16 ***
## brandtropicana:feat               0.7701326  0.0977263   7.881 3.37e-15 ***
## log(price):brandminute.maid:feat -1.0917803  0.1209803  -9.024  < 2e-16 ***
## log(price):brandtropicana:feat   -0.9798686  0.1228215  -7.978 1.54e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6877 on 28932 degrees of freedom
## Multiple R-squared:  0.545,  Adjusted R-squared:  0.5448
## F-statistic:  2476 on 14 and 28932 DF,  p-value: < 2.2e-16
```

```
sum_M = 0
sum_Q = 0
for (cweek in 40:160) {
  oj_cweek <- oj %>%
    filter(week == cweek)

  oj_cweek_mean <- oj_cweek %>%
    mutate(HHLARGE = rep(c(0.11560), each = nrow(oj_cweek)))

  oj_cweek_3rdq <- oj_cweek %>%
    mutate(HHLARGE = rep(c(0.13517), each = nrow(oj_cweek)))

  sum_mean <- sum(predict(model5, newdata = oj_cweek_mean))
  sum_3rdq <- sum(predict(model5, newdata = oj_cweek_3rdq))

  sum_M = sum_M + sum_mean
  sum_Q = sum_Q + sum_3rdq
}

sum_M/120 - sum_Q/120
```

```
## [1] 16.27816
```

```
        The log quantity is increased by 16 units as we move HHLARGE from mean value to 3rd quartile value.

    ii.If we move from the median value of EDUC to the 75th percentile (3rd quartile), how much
does log(quantity) change each week on average?
```

```
sum_M = 0
sum_Q = 0
for (cweek in 40:160) {
  oj_cweek <- oj %>%
    filter(week == cweek)

  oj_cweek_mean <- oj_cweek %>%
    mutate(EDUC = rep(c(0.22522), each = nrow(oj_cweek)))

  oj_cweek_3rdq <- oj_cweek %>%
    mutate(EDUC = rep(c(0.28439), each = nrow(oj_cweek)))

  sum_mean <- sum(predict(model5, newdata = oj_cweek_mean))
  sum_3rdq <- sum(predict(model5, newdata = oj_cweek_3rdq))

  sum_M = sum_M + sum_mean
  sum_Q = sum_Q + sum_3rdq
}

sum_M/120 - sum_Q/120
```

```
## [1] 3.002672
```

```
        The log quantity is increased by 3 units as we move EDUC from mean value to 3rd quartile
value.

    iii. Base on this analysis, which is the more important predictor of demand?

        HHLARGE should be more importand as it increases log quantity by a larger amount.

c. Now let’s see if these variables impact price sensitivity. Add two interaction terms (with lo
gprice) to the model to test this.
```

```
model6 <- lm(logmove ~ log(price)*brand*feat + store + HHLARGE*log(price) + EDUC*log(price), dat
a = oj)
summary(model6)
```

```
##
## Call:
## lm(formula = logmove ~ log(price) * brand * feat + store + HHLARGE *
##     log(price) + EDUC * log(price), data = oj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3867 -0.4101 -0.0143  0.3965  3.1147
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     10.9477862  0.0659363 166.036  < 2e-16 ***
## log(price)                      -3.0992322  0.0824613 -37.584  < 2e-16 ***
## brandminute.maid                 0.0696993  0.0451406   1.544    0.123
## brandtropicana                   0.7828342  0.0492138  15.907  < 2e-16 ***
## feat                             1.0761805  0.0368819  29.179  < 2e-16 ***
## store                            0.0018269  0.0001153  15.840  < 2e-16 ***
## HHLARGE                          0.3782223  0.4092054   0.924    0.355
## EDUC                            -3.1728216  0.1118466 -28.368  < 2e-16 ***
## log(price):brandminute.maid      0.7665863  0.0594379  12.897  < 2e-16 ***
## log(price):brandtropicana        0.6838085  0.0550404  12.424  < 2e-16 ***
## log(price):feat                 -0.4350462  0.0717121  -6.067 1.32e-09 ***
## brandminute.maid:feat            1.1487178  0.0793221  14.482  < 2e-16 ***
## brandtropicana:feat              0.7385191  0.0955802   7.727 1.14e-14 ***
## log(price):HHLARGE              -4.8287076  0.4829774  -9.998  < 2e-16 ***
## log(price):EDUC                  3.7485178  0.1328453  28.217  < 2e-16 ***
## log(price):brandminute.maid:feat -1.0886350  0.1183139  -9.201  < 2e-16 ***
## log(price):brandtropicana:feat   -0.9671499  0.1201174  -8.052 8.48e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6726 on 28930 degrees of freedom
## Multiple R-squared:  0.5649, Adjusted R-squared:  0.5647
## F-statistic:  2348 on 16 and 28930 DF,  p-value: < 2.2e-16
```

```
    i. What are the coefficients on the interaction terms?

        The coefficients are -4.8287076 for HHLARGE and log price, and 3.7485178 for EDUC and lo
g price.

    ii. Does the sign of your estimates make sense based on your intuition?

        The sign of HHLARGE make sense as negative coefficient means negative influence on quant
ity, which means if the fraction for large household is high, then the effect of price on quanti
ty will be magnified, and a negative sign means large family tends to buy less units if price is
high, which correpsonds to what happends in daily lives.

    iii. What are the coefficient estimates on the variables EDUC and HHLARGE that aren’t part o
f the interaction term? How do they compare to your regression from 1b?

        The coefficient for HHLARGE is 0.3782223, and the coefficient for EDUC is -3.1728216. An
d compared to previous model, the sign of coefficient of HHLARGE is changed to positive.

    iv. Similar to 2b, if we move from the median value of each variable to the 3rd quartile, ho
w much does elasticity change? Based on this, which is more important to price sensitivity?
```

```
# HHLARGE change
sum_M = 0
sum_Q = 0
for (cweek in 40:160) {
  oj_cweek <- oj %>%
    filter(week == cweek)

  oj_cweek_mean <- oj_cweek %>%
    mutate(HHLARGE = rep(c(0.11560), each = nrow(oj_cweek)))

  oj_cweek_3rdq <- oj_cweek %>%
    mutate(HHLARGE = rep(c(0.13517), each = nrow(oj_cweek)))

  sum_mean <- sum(predict(model6, newdata = oj_cweek_mean))
  sum_3rdq <- sum(predict(model6, newdata = oj_cweek_3rdq))

  sum_M = sum_M + sum_mean
  sum_Q = sum_Q + sum_3rdq
}

diff_HHLARGE <- sum_M/120 - sum_Q/120


# EDUC change
sum_M = 0
sum_Q = 0
for (cweek in 40:160) {
  oj_cweek <- oj %>%
    filter(week == cweek)

  oj_cweek_mean <- oj_cweek %>%
    mutate(EDUC = rep(c(0.22522), each = nrow(oj_cweek)))

  oj_cweek_3rdq <- oj_cweek %>%
    mutate(EDUC = rep(c(0.28439), each = nrow(oj_cweek)))

  sum_mean <- sum(predict(model6, newdata = oj_cweek_mean))
  sum_3rdq <- sum(predict(model6, newdata = oj_cweek_3rdq))

  sum_M = sum_M + sum_mean
  sum_Q = sum_Q + sum_3rdq
}

diff_EDUC <- sum_M/120 - sum_Q/120

diff_HHLARGE
```

```
## [1] 16.08787
```

```
diff_EDUC
```

```
## [1] 3.335238
```

```
         Form the result, HHLARGE has higher elasticity, as change in HHLARGE leads to higher in
crease in quantity, thus HHLARGE seems to be more important to price sensitivity.

d. You should notice that the coefficients on EDUC and HHLARGE have flipped sign once we include
interaction terms with price. HHLARGE now appears to be a positive demand shifter and increases
price sensitivity. Explain in words or pictures what is going on.

    If we do not consider interaction between price and HHLARGE, HHLARGE alone would have negati
ve influence on quantity, as larger family might drink less orange juice, they might prefer mil
k, coke or other substitutes more. But after considered the interaction between price and HHLARG
E, it is that large family tends to save money, so they are sensitive to change in price, thus H
HLARGE magnifies the impact of price on quantity.
```

3. Create make a new dataframe which takes the previous week's prices as a variable on the same line as the current week. This would enable you to see if there is intertemporal substitution.

    a.

```
oj_week <- oj %>%
  group_by(store, brand) %>%
  mutate(last_price = lag(price)) %>%
  ungroup()
```

```
b. Now run a regression with this week’s log(quantity) on current and last week’s price.
```

```
model7 <- lm(logmove ~ log(price) + log(last_price), data = oj_week)
summary(model7)
```

```
##
## Call:
## lm(formula = logmove ~ log(price) + log(last_price), data = oj_week)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9122 -0.5197 -0.0193  0.5097  3.7811
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.98879    0.01522  656.47   <2e-16 ***
## log(price)      -2.77539    0.02266 -122.48   <2e-16 ***
## log(last_price)  1.73059    0.02255   76.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8267 on 28695 degrees of freedom
##   (249 observations deleted due to missingness)
## Multiple R-squared:  0.344,  Adjusted R-squared:  0.344
## F-statistic:  7525 on 2 and 28695 DF,  p-value: < 2.2e-16
```

```
 c.What do you notice about the previous week’s elasticity?  Does this make sales more or less a
ttractive from a profit maximization perspective?  Why?
    The elasticity for previous week is positive. And this would make sales more attractive as t
he higher the price is during last week, the higher the sales this week. And this coincides with
fact that if the price this week is high, people tends to wait untill the price drops to purchas
e, thus sales for next week would increase.
```