Pricing Analytics - Project 1

Shen, Ruiling; Tao, Chenxi; Liang, Jiawen; Tran, Khanh; Ding, Xiaodan

1/22/2020

```
library("lfe")
```

```
## Warning: package 'lfe' was built under R version 3.6.2

## Loading required package: Matrix
```

```
library("data.table")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::between()   masks data.table::between()
## x tidyr::expand()    masks Matrix::expand()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x tidyr::pack()      masks Matrix::pack()
## x purrr::transpose() masks data.table::transpose()
## x tidyr::unpack()    masks Matrix::unpack()
```

```
rm(list = ls())
setwd("E:/Studying/Simon/Classes/MKT440 - Pricing Analytics/Project 1/car data")
cardata = fread("cars.csv", stringsAsFactors = F)
ironPrice = fread("iron_ore.csv", stringsAsFactors = F)
```

4. Control variables

4.1. Interpreting a log0log regression

```r
# Our colleague's regression
reg = felm(log(qu)~log(eurpr), data=cardata)
summary(reg)
```

```
##
## Call:
##    felm(formula = log(qu) ~ log(eurpr), data = cardata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8279 -1.1271 -0.0056  1.1214  4.3239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.32158    0.20207   56.03   <2e-16 ***
## log(eurpr)  -0.29603    0.02284  -12.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 11547 degrees of freedom
## Multiple R-squared(full model): 0.01434   Adjusted R-squared: 0.01426
## Multiple R-squared(proj model): 0.01434   Adjusted R-squared: 0.01426
## F-statistic(full model):  168 on 1 and 11547 DF, p-value: < 2.2e-16
## F-statistic(proj model):   168 on 1 and 11547 DF, p-value: < 2.2e-16
```

1. Interpretation of the coefficient.

- Intercept: The expected sales of the car is e^11.32158 which is approximately 82585.
- Coefficient of log(eurpr): When price increases 1%, the sales is expected to decrease by 0.296%. In other words, the own price elasticity of the car is |0.29603|.

2. Rationality of the coefficient. The estimated demand from my colleague's demand model is **not reasonable**, and there are two reasons to justify our statement.

- 1) She got a own price elasticity of 0.29603 which stood for an **inelastic** demand of car. An inelastic demand means that consumers are not very price sensitive to the product; however, when it comes to the product like cars which typically cost consumer more than $10,000, consumers tend to be more price sensitive.

- 2) The demand model of the colleague failed to consider any potential variables or factors that may affect the demand, which may issue a serious **omitted variable bias** of her estimation.

Based on aforementioned statement, we **disagree** with her demand estimation.

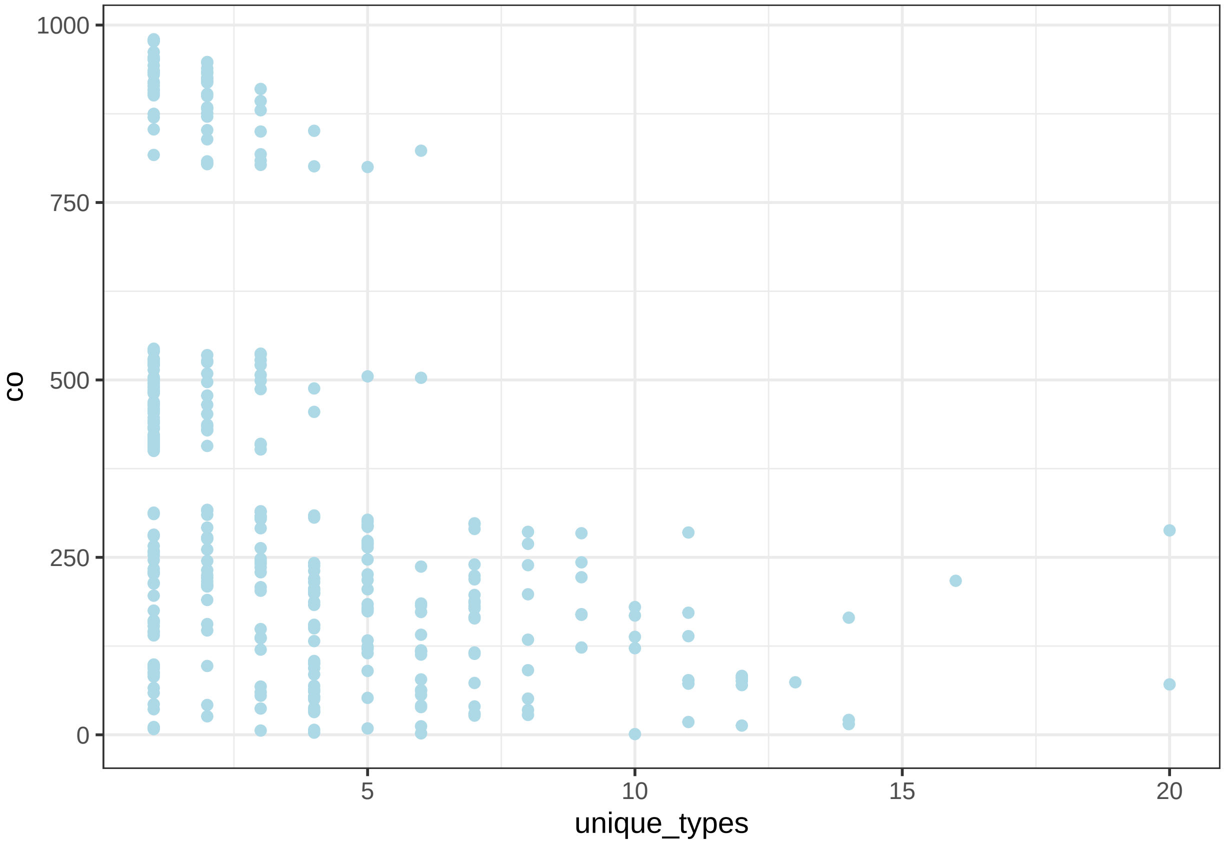
4.2. Adding cotrol variables

At the first glance of this question, we added year(**ye**), market(**ma**), car model(**co**), and the interaction of ma and ye as the fixed effect. The reason for fixed effects is that in different year, different market, and for different model, the demand is different. The reason for ye and ma interaction is that demand in different markets changes over year. For example, in certain years, China was in tension with Japan, then demand for Japanese cars in China market went down during these years. So we set this model as our baseline.

To avoid omitted variable bias, we checked whether car attributes were captured by the variable co. So we did the following scatter plot for each car attribute variable from cy to ac.

```
caratt <- select(cardata,co,cy:ac)

attri <-
    caratt %>%
    group_by(co) %>%
    mutate(unique_types = n_distinct(ac)) %>%
    select(co,unique_types)
unique <- attri[!duplicated(attri$co), ]
print(ggplot(unique, aes(x=unique_types, y=co)) + geom_point(color="lightblue") + theme_bw())
```



Cite the above scatter plot of variable ac as an example. The scatter plot illustrates how many unique value of ac for a unique co. So if ac is fully incorporated in the variable co, we should observe a single vertical line with x value equal to 1. Therefore, ac is not fully incorporated in the variable co. We followed the same method for every car attribute variable, and discoverd that none of the car attribute variable is fully incorported in the variable co.

So we chose and added some car attributes into our model. To get an idea of which variable to add in the demand estimation, we created the following correlation table.

```
carattri <- select(cardata,qu,eurpr,cy:ac)
library('corrplot')
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
## corrplot 0.84 loaded
```

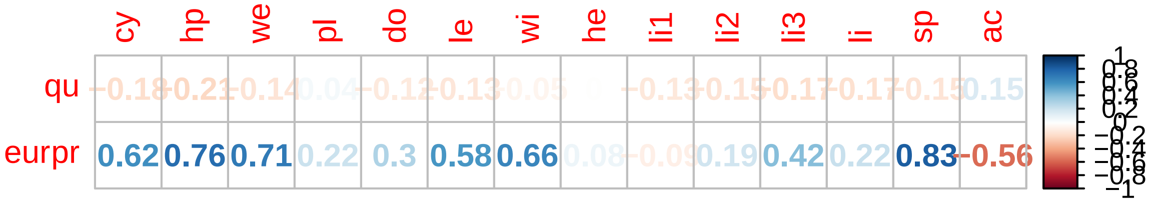
```
library('caret')
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
corrplot(cor(carattri[,1:2],carattri[,3:16], use = "pairwise.complete.obs"),method='number')
```



From the correlation table we can observe the correlation between these variables and demand/price. We first selected variables which are relatively correlated with both demand and price. They are cy, hp, sp, ac. The intution is that horsepower, maximum speed and acceleration time show the performance of cars. These attributes influence car price. At the same time, customers care much about performance when purchasing cars. Cylinder volume would also affect price of cars and be considered by customers.

Then we considered size attributes: wi, le and he. Size of cars affect price because larger cars require more materials to produce. It also influences demand since different customers have different preference for car size. Some prefer mini cooper while someone prefer land cruiser. So, we put wi, le and he into the model.

However, le is insignificant and has almost no impact on coefficient. We decided to remove it and kept wi and he.

The following is our demand estimation model which considered variables that had the potential to cause omitted variable bias.

```
reg4 <- felm(log(qu)~log(eurpr)+cy+hp+wi+he+sp+ac|
                 factor(co)+factor(ye)+factor(ma)+factor(ye):factor(ma),
             data=cardata)
summary(reg4)
```

```
##
## Call:
##    felm(formula = log(qu) ~ log(eurpr) + cy + hp + wi + he + sp +      ac | factor(co) + factor(ye)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0754 -0.6227 -0.0231  0.5950  4.0779
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## log(eurpr) -1.8126981  0.1468450 -12.344  < 2e-16 ***
## cy         -0.0006267  0.0001289  -4.861 1.19e-06 ***
## hp         -0.0207760  0.0029774  -6.978 3.21e-12 ***
## wi          0.0721055  0.0062510  11.535  < 2e-16 ***
## he          0.0198964  0.0074723   2.663  0.00777 **
## sp          0.0284783  0.0029971   9.502  < 2e-16 ***
## ac          0.0887031  0.0071258  12.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 8771 degrees of freedom
##   (2321 observations deleted due to missingness)
## Multiple R-squared(full model): 0.6196   Adjusted R-squared: 0.5998
## Multiple R-squared(proj model): 0.07902   Adjusted R-squared: 0.03114
## F-statistic(full model):31.33 on 456 and 8771 DF, p-value: < 2.2e-16
## F-statistic(proj model): 107.5 on 7 and 8771 DF, p-value: < 2.2e-16
## *** Standard errors may be too high due to more than 2 groups and exactDOF=FALSE
```

However, there should be some variables that are not in the dataset cause omitted variable bias issue. Hence, we listed some below.

- Whether the car had **promotion**, such as discount.
- Whether the car was **advertised** and what is the strength of the advertisement.

5. Instrumental variables

Based our common sense, we think cost of car production is highly correlated with price, but normally, customers don't think much about cost, and cost of a car has nothing to do with the demand of the car.

Similarly, economic indexes like exchange rate, CPI, and tax are also highly correlated with price but seem not to be correlated with demand.

So, we chose production cost, exchange rate, CPI, and tax as potential IVs and checked their validity below:

```
#load cost-relevant data
ironPrice$ye <- as.numeric(substr(ironPrice$year, 3, 4))
newCarData <- merge(cardata, ironPrice[, c("ye", "unit_value_98")], by="ye")
newCarData$totalCost <- newCarData$we
# Validity check for weight
newCarData$pro_cost = newCarData$we*newCarData$unit_value_98
check1 <- lm(log(eurpr)~pro_cost,data=newCarData)
summary(check1)
```

```
##
## Call:
## lm(formula = log(eurpr) ~ pro_cost, data = newCarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13618 -0.43440  0.02376  0.46280  1.98675
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.892e+00  1.696e-02 524.432  < 2e-16 ***
## pro_cost    -1.443e-06  3.329e-07  -4.334 1.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6577 on 11547 degrees of freedom
## Multiple R-squared:  0.001624,   Adjusted R-squared:  0.001538
## F-statistic: 18.79 on 1 and 11547 DF,  p-value: 1.475e-05
```

```
# Validity check for exchange rate
check2 <- lm(log(eurpr)~avdexr,data=newCarData)
summary(check2)
```

```
##
## Call:
## lm(formula = log(eurpr) ~ avdexr, data = newCarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13794 -0.43989  0.02202  0.46353  1.92403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.791e+00  6.675e-03 1317.02   <2e-16 ***
## avdexr      1.108e-04  9.173e-06   12.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6541 on 11547 degrees of freedom
## Multiple R-squared:  0.01247,    Adjusted R-squared:  0.01239
## F-statistic: 145.9 on 1 and 11547 DF,  p-value: < 2.2e-16
```

```r
# Validity check for cpi
check3 <- lm(log(eurpr)~avdcpr,data=newCarData)
summary(check3)
```

```
##
## Call:
## lm(formula = log(eurpr) ~ avdcpr, data = newCarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45036 -0.33558 -0.01732  0.30530  1.81783
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.6829394  0.0109005   704.8   <2e-16 ***
## avdcpr      0.0140848  0.0001241   113.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4525 on 11547 degrees of freedom
## Multiple R-squared:  0.5273, Adjusted R-squared:  0.5272
## F-statistic: 1.288e+04 on 1 and 11547 DF,  p-value: < 2.2e-16
```

```r
# Validity check for tax
check4 <- lm(log(eurpr)~tax,data=newCarData)
summary(check4)
```

```
##
## Call:
## lm(formula = log(eurpr) ~ tax, data = newCarData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16438 -0.42909  0.02959  0.46199  1.96145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.21700    0.02361  390.40   <2e-16 ***
## tax         -1.83550    0.10656  -17.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6499 on 11547 degrees of freedom
## Multiple R-squared:  0.02505,    Adjusted R-squared:  0.02497
## F-statistic: 296.7 on 1 and 11547 DF,  p-value: < 2.2e-16
```

First, we created a new variable called pro_cost(weight*iron_price) representing the production cost of cars.

The result shows that they are all correlated with price, so they may be proper IVs. Therefore, we tried the combinations of these IVs to our model. After testing the IVs, exchange rate, CPI and tax do not make any changes to our model, so we excluded them.

```
reg5 <- felm(log(qu)~cy+hp+wi+he+ac+sp|
                 factor(co)+factor(ye)+factor(ma)+factor(ye):factor(ma)|
                 (log(eurpr)~pro_cost),
             data=newCarData)
summary(reg5)
```

```
##
## Call:
##    felm(formula = log(qu) ~ cy + hp + wi + he + ac + sp | factor(co) +      factor(ye) + factor(ma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6516 -0.8201 -0.0224  0.8047  9.6585
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## cy                   6.122e-04  6.161e-04   0.994  0.32043
## hp                   2.686e-02  2.315e-02   1.160  0.24615
## wi                   1.128e-01  2.106e-02   5.357 8.70e-08 ***
## he                   4.127e-02  1.392e-02   2.964  0.00305 **
## ac                   1.291e-01  2.135e-02   6.046 1.55e-09 ***
## sp                   5.485e-02  1.320e-02   4.154 3.29e-05 ***
## `log(eurpr)(fit)` -1.237e+01  5.069e+00  -2.441  0.01467 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.313 on 8771 degrees of freedom
##   (2321 observations deleted due to missingness)
## Multiple R-squared(full model): 0.3953   Adjusted R-squared: 0.3639
## Multiple R-squared(proj model): -0.4641   Adjusted R-squared: -0.5402
## F-statistic(full model):19.51 on 456 and 8771 DF, p-value: < 2.2e-16
## F-statistic(proj model): 54.79 on 7 and 8771 DF, p-value: < 2.2e-16
## F-statistic(endog. vars):5.958 on 1 and 8771 DF, p-value: 0.01467
## *** Standard errors may be too high due to more than 2 groups and exactDOF=FALSE
```

Adding IVs makes the P-value of price's estimation much higher than before. Since the coefficient is not plausible to be 0, another reason may be due to an unusually high standard error. We wanted to detect whether the low variation in IVs causes the high standard error, so we used IVs as dependent variables to run the following regressions:

```
a = felm(pro_cost~cy+hp+wi+he+ac+sp|
             factor(co)+factor(ye)+factor(ma)+factor(ye):factor(ma),
         data=newCarData)
summary(a)$adj.r.squared
```

```
## [1] 0.9764462
```

The adjusted R-squared is very large, which means the independent variables can almost perfectly explain the variance in pro_cost(IV). So, it is safe to say, the high standard error in the previous regression is because of very low variance in IV itself.

To conclude, the IV we added to our model is reasonable to some extent, but it is not likely to be considered as an ideal indicator because the estimation of elasticity after adding IV to our model becomes too large.

Here are some insights about why our IV is not good enough: Intuitively, the cost of iron itself can probably not determine car prices. Except for raw material, cost of internal parts such as engine as well as labor cost take up most of production costs.

Thus, we decided not to add IV in our model.

6. Recovering costs

If we estimate a log-log regression, we might get the result as following:

$$E(\log(Q)|P) = \beta_0 + \beta_1 \cdot \log(P) + \beta_x \cdot X + \varepsilon$$

We can simulate profit for each value of P.

$$\text{Profit} \ = \ \text{Revenue} \ - \ \text{cost} \ = (P - UC) \cdot e^{\beta_0 + \beta_1 \cdot \log(P) + \beta_x \cdot X + \varepsilon} - FC$$

The optimal price is the one that generates maximum profit, indicating that the derivative of profit with respect to price is 0.

$$\text{Profit}' = e^{\beta_0 + \beta_1 \cdot \log(P) + \beta_x \cdot X + \varepsilon} \cdot \left(1 + \beta_1 - UC \cdot \frac{\beta_1}{P}\right)$$

$$\text{Profit }' = 0 \Rightarrow 1 + \beta_1 - UC \cdot \frac{\beta_1}{P} = 0$$

Then we can safely conclude the relationship between price and marginal cost:

$$MC = \frac{1 + \beta_1}{\beta_1} \cdot P$$

From the final causal model, we know that:

```
B1 = -1.8127
```

#(1) alfa33, 1983, 1

```
P1 = 5507.57683
UC1 = (1+B1)*P1/B1
UC1
```

```
## [1] 2469.249
```

#(2) audi80/90, 1990, 1

```
P2 = 10931.8132
UC2 = (1+B1)*P2/B1
UC2
```

```
## [1] 4901.133
```

#(3) BMW5, 1999, 1

```
P3 = 11314.4215
UC3 = (1+B1)*P3/B1
UC3
```

```
## [1] 5072.671
```

7. Cross-elasticities and competitive effects

```
reg7 <- felm(log(qu)~log(eurpr)+log(avgurprrival)+cy+hp+wi+he+ac+sp|
                 factor(co)+factor(ye)+factor(ma)+factor(ye):factor(ma),
             data=cardata)
summary(reg7)
```

```
##
## Call:
##    felm(formula = log(qu) ~ log(eurpr) + log(avgurprrival) + cy +      hp + wi + he + ac + sp | facto
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0865 -0.6148 -0.0213  0.5965  4.0684
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## log(eurpr)        -0.7085374  0.2122987  -3.337 0.000849 ***
## log(avgurprrival) 78.6881670 10.9551307   7.183 7.39e-13 ***
## cy                -0.0005730  0.0001288  -4.450 8.69e-06 ***
## hp                -0.0195994  0.0029734  -6.592 4.60e-11 ***
## wi                 0.0728828  0.0062340  11.691  < 2e-16 ***
## he                 0.0198032  0.0074508   2.658 0.007878 **
## ac                 0.0849931  0.0071241  11.930  < 2e-16 ***
## sp                 0.0267486  0.0029982   8.922  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.039 on 8770 degrees of freedom
##   (2321 observations deleted due to missingness)
## Multiple R-squared(full model): 0.6218   Adjusted R-squared: 0.6021
## Multiple R-squared(proj model): 0.08441   Adjusted R-squared: 0.0367
## F-statistic(full model):31.56 on 457 and 8770 DF, p-value: < 2.2e-16
## F-statistic(proj model): 101.1 on 8 and 8770 DF, p-value: < 2.2e-16
## *** Standard errors may be too high due to more than 2 groups and exactDOF=FALSE
```

After including the average prices of rival cars (`avgurprrival`), the coefficient of the log(average rival price) is 78.6881. This coefficient means that if the average price of rivals' car increases by 1%, the sales of our car will increase by 78.7%.

The expected coefficient of the log-average rival price is positive, because when rival price increases, our sales are expected to increase. The computed coefficient satisfies this criteria. In addition, the coefficient here is very large, showing that the car markets in Europe are extremely competitive (near perfect competition).

By including rivals' price, the coefficient of log(our price) shifts upwards and decreases in magnitude (from -1.813 to -0.708). This means rivals' price affects both our price and our sales and is an omitted variable.

Finally, because the average rival price are the average price of all rival's cars in the same market and period, it might not show the correct price of the car models directly competing with our car models. In order for the calculation to be more accurate, the average rival price should be computed on the same market, same period and on the **competing car models**.