BLP Assignment

Jose M. Fernandez

MSBA Data Analytics III

Problem Set: Ordered Probit Methodology The San Francisco Airport is very concerned about customer satisfaction. Attached you will find a customer satisfaction survey for the SFO. *Q7ALL* is a categorical variable ranking the customer's satisfaction from *unacceptable* to *outstanding*

```
library(stargazer) # for creating Tables
library(MASS) # For doing ordered Logit/probit
library(AER) # For performing ivreg
library(knitr)

#### Reading in Data
library(readr)
SFO <- read_csv("C:/Users/jmfern02.AD/Downloads/2015_SFO_Customer_Survey.csv")
# Codebook
# Q7ALL SFO Airport as a whole
# 5 Outstanding
# 4
# 3
# 2
# 1 Unacceptable
# 6 Have never used or visited / Not applicable
# 0 Blank
```

a. Create a table summarizing the counts of the potential outcomes and report them as percentages.

```
m1<-table(SFO$Q7ALL)
names(m1)<-c("Blank","Unacceptable","Poor","Average","Good","Outstanding","NA")
kable(round(prop.table(m1),digits = 3),type="html")
```

Var1	Freq
Blank	0.037
Unacceptable	0.002
Poor	0.010
Average	0.178
Good	0.510
Outstanding	0.250
NA	0.014

b. Review the available variables. Which variables capture customers characteristics?

Q18Age 1 Under 18 2 18 - 24 3 25 - 34 4 35 - 44 5 45 - 54 6 55 - 64 7 65 and over 8 Don’t Know / Refused 0 Blank/Multiple responses

Q19Gender 1 Male 2 Female 3 Other 0 Blank/Multiple responses

Q20INCOME Household Income: 1 Under 50,000 2 $50,000 - $100,000 3 $100,001 - $150,000 4 Over $150,000 5 Other Currency (specify) 0 Blank/Multiple responses

Q21FLY Did you fly 100,000 miles or more per year? 1 Yes 2 No 3 Don’t know 0 Blank/Multiple responses

LANGUAGE of questionnaire: 1 English 2 Spanish 3 Chinese 4 Japanese

In an effort to clean up the data. I will be removing all observations that are coded 0 or do not use an informative response (i.e. I don’t know)

I will let student include other factor as long as they are customer characteristics and not airport service and facilities.

```
# First, I only keep the variable of interest
SFO<-SFO[c("Q7ALL","Q18AGE","Q19GENDER","Q20INCOME","Q21FLY","LANG")]
# Next, I remove Blank responses, "Other", and "I Don't Know"
SFO<-SFO[SFO$Q7ALL!=0 & SFO$Q7ALL!=6, ]
SFO<-SFO[SFO$Q18AGE!=0 & SFO$Q18AGE!=8, ]
SFO<-SFO[SFO$Q19GENDER!=0 & SFO$Q19GENDER!=3, ]
SFO<-SFO[SFO$Q20INCOME!=0 & SFO$Q20INCOME!=5, ]
SFO<-SFO[SFO$Q21FLY!=0, ]
```

c. Estimate an ordered probit model of overal satisfaction using customer characteristics as explanatory variables.

```
reg1<-polr(factor(Q7ALL)~factor(Q18AGE)+factor(Q19GENDER)
           +factor(Q20INCOME)+factor(Q21FLY)+factor(LANG),
           data=SFO,method="probit")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
stargazer(reg1,type = "html", dep.var.caption = "",
          covariate.labels = c("18 - 24 years old","25 - 34 years old",
                               "35 - 44 years old","45 - 54 years old",
                               "55 - 64 years old","65 and over years old",
                               "Female"," 50,000 -  100,000","100,001 -  150,000",
                               " 150,000 or over","Other Income","Not a Frequent flyer",
                               "Not Sure","Spanish","Chinese","Japanese"))
```

	Q7ALL
18 - 24 years old	-0.081
	(0.229)
25 - 34 years old	-0.206
	(0.223)
35 - 44 years old	-0.396*
	(0.224)
45 - 54 years old	-0.327

	(0.225)
55 - 64 years old	-0.298
	(0.228)
65 and over years old	-0.254
	(0.233)
Female	0.099**
	(0.048)
50,000 - 100,000	0.069
	(0.074)
100,001 - 150,000	-0.089
	(0.084)
150,000 or over	-0.137*
	(0.080)
Other Income	-0.146
	(0.640)
Not a Frequent flyer	0.122*
	(0.063)
Not Sure	-0.084
	(0.107)
Spanish	0.320
	(0.270)
Chinese	0.297*
	(0.176)
Japanese	-0.837***
	(0.241)
Observations	2,181
Note:	*p<0.1;* ***p<0.05;*** p<0.01

You can skip the marginal effects portion. No points will be deducted here.

Problem Set: BLP Methodology Attached is a table of market shares, prices, and characteristics on the top-selling brands of cereal in 1992. The data are aggregated from household-level scanner data (collected at supermarket checkout counters).

The market shares below are shares of *total cereal purchases* observed in the dataset. For the purposes of this problem set, assume that all households purchased some cereal during 1992 (so that non-purchase is not an option). Assume that brand #51, the composite basket of “all other brands”, is the outside good.

Two sets of prices are given in the table. *Shelf prices* are those listed on supermarket shelves, and do not include coupon discounts. *Transactions prices* are prices actually paid by consumers, net of coupon discounts. Estimate using the transactions prices. Note that you should subtract the price of brand #51, the “outside good”, from the prices of the top fifty brands.

Assume a utility specification for uij , household i’s utility from brand j:

$$u_{ij} = X_j\beta + \alpha p_j + \xi_j + v_{ij}$$

where X_j are characteristics of brand j, ξ_j is an unobserved (to the econometrician) quality parameter for brand j, and v_{ij} is a disturbance term which is identicially and independently distributed (i.i.d.) over households i and brands j. As in Berry (1994), denote the mean utility level from brand j as

$$\delta_j \equiv X_j\beta + \alpha p_j + \xi_j$$

1. Assuming that the v_{ij} 's are distributed i.i.d. type I extreme value, derive the resulting expressions for the market shares of each brand j; $j = 1, \ldots, 51$.

We define the brand name goods share as

$$Pr\left(Y = j\right) = s_{jm} = \frac{exp(\delta_{jm})}{1 + \sum exp(\delta_{km})}$$

We define the outside option as

$$s_{0m} = \frac{1}{1 + \sum exp(\delta_{km})}$$

Now let's take logs of the shares

$$log(s_{jm}) = \delta_{jm} + log(1 + \sum exp(\delta_{km}))$$

$$log(s_{0m}) = log(1) + log(1 + \sum exp(\delta_{km}))$$

The log(1)=0 so we can write the difference in logs as

$$log(s_{jm}) - log(s_{0m}) = \delta_{jm}$$

Next we implement the BLP two-step estimator.

2. Invert the resulting system of demand functions to get estimates of the mean utility levels δ_j as a function of the shares s_j . (Hint: look in Berry (1994, pg. 250)).

```
library(readr)
blp <- read_csv("C:/Users/jmfern02.AD/Downloads/Workbook1 (2).csv")
# We need to normalize our data with respect to the outside option.
#We do this by subtracting the transaction price of the outside basket from all other prices.
#We also subtract the log (market share) of the outside option from all other log market shares
blp$adj.price<-blp$`average transaction price`-blp$`average transaction price`[51]
blp$delta<-log(blp$`in sample market share`)-log(blp$`in sample market share`[51])
# Now we no longer need the basket of goods observation
blp<-blp[-51, ]
```

3. Estimate the second stage regression of δ_j on X_j and p_j in different ways:

(a and b) OLS

```
## In this section we are just going to run OLS on the linear demand curve
blp.reg.1<-lm(delta~fat+suger+cals+adj.price,data=blp)
blp.reg.fe<-lm(delta~fat+suger+cals+adj.price+factor(company),data=blp)
stargazer(blp.reg.1,blp.reg.fe,type = "html", keep = c("fat","suger","cals","adj.price"),covaria
te.labels = c("fat","suger","calories","price difference"))
```

	Dependent variable:	
	delta	
	(1)	(2)
fat	-0.005	0.011
	(0.057)	(0.062)

suger	-0.029**	-0.038**
	(0.014)	(0.015)
calories	0.001	0.002
	(0.003)	(0.003)
price difference	-0.224	-0.400**
	(0.149)	(0.170)
Observations	50	50
R^2	0.150	0.345
Adjusted R^2	0.074	0.197
Residual Std. Error	0.483 (df = 45)	0.450 (df = 40)
F Statistic	1.982 (df = 4; 45)	2.336** (df = 9; 40)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

c. 2SLS: using average characteristics for all other brands produced by the same manufacturer as brand j as instruments for p_j

```
## ave is a group mean function, but you can use it to apply other functions
# Here I calculate the mean calories of products by the same firm excluding the product of inter
est.
blp$Firm <- blp$company
blp$cal.z<-(ave(blp$cals,factor(blp$Firm),FUN=function(x)
  sum(x))-blp$cals)/(ave(blp$cals,factor(blp$Firm),FUN=function(x) length(x))-1)

blp$fat.z<-(ave(blp$fat,factor(blp$Firm),FUN=function(x) sum(x))-blp$fat)/(ave(blp$fat,factor(bl
p$Firm),FUN=function(x) length(x))-1)

blp$suger.z<-(ave(blp$suger,factor(blp$Firm),FUN=function(x) sum(x))-blp$suger)/(ave(blp$suger,f
actor(blp$Firm),FUN=function(x) length(x))-1)

blp$adj.price.z<-(ave(blp$adj.price,factor(blp$Firm),FUN=function(x) sum(x))-blp$adj.price)/(ave
(blp$adj.price,factor(blp$Firm),FUN=function(x) length(x))-1)

blp.reg.2<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z+fat.z+suger.z,data=blp)

blp.reg.price.2<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z+fat.z+suger.z+adj.pri
ce.z,data=blp)
```

d. 2SLS: using average characteristics for all other brands produced by rivals to the manufacturer as brand j as instruments for p_j

```
## We can declare rivals as those in the same segment.
#We will repeat what we did above, but instead of grouping by firm we are grouping by segment.

blp$cal.z2<-(ave(blp$cals,factor(blp$sgmnt),FUN=function(x)
  sum(x))-blp$cals)/ave(blp$cals,factor(blp$sgmnt),FUN=function(x) length(x))

blp$fat.z2<-(ave(blp$fat,factor(blp$sgmnt),FUN=function(x)
  sum(x))-blp$fat)/ave(blp$fat,factor(blp$sgmnt),FUN=function(x) length(x))

blp$suger.z2<-(ave(blp$suger,factor(blp$sgmnt),FUN=function(x)
  sum(x))-blp$suger)/ave(blp$suger,factor(blp$sgmnt),FUN=function(x) length(x))

blp$adj.price.z2<-(ave(blp$adj.price,factor(blp$sgmnt),FUN=function(x)
  sum(x))-blp$adj.price)/ave(blp$adj.price,factor(blp$sgmnt),FUN=function(x) length(x))

blp.reg.3<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z2+fat.z2+suger.z2,data=blp)

blp.reg.price.3<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z2+fat.z2+suger.z2+adj.
price.z2,data=blp)
```

e. 2SLS: using average characteristics for all other brands not produced by manufacturer of product j as instruments for p_j.

```
## We can declare rivals as those in the same segment.
#We will repeat what we did above, but instead of grouping by firm we are grouping by segment.

blp$cal.z3<-(ave(blp$cals,FUN=function(x)
  sum(x))-ave(blp$cals,factor(blp$Firm),FUN=function(x)
  sum(x)))/(ave(blp$cals,FUN=function(x) length(x))-ave(blp$cals,factor(blp$Firm),FUN=function
(x) length(x)))

blp$fat.z3<-(ave(blp$fat,FUN=function(x)
  sum(x))-ave(blp$fat,factor(blp$Firm),FUN=function(x)
  sum(x)))/(ave(blp$fat,FUN=function(x) length(x))-ave(blp$fat,factor(blp$Firm),FUN=function(x)
 length(x)))

blp$suger.z3<-(ave(blp$suger,FUN=function(x)
  sum(x))-ave(blp$suger,factor(blp$Firm),FUN=function(x)
  sum(x)))/(ave(blp$suger,FUN=function(x) length(x))-ave(blp$suger,factor(blp$Firm),FUN=function
(x) length(x)))

blp$adj.price.z3<-(ave(blp$adj.price,FUN=function(x)
  sum(x))-ave(blp$adj.price,factor(blp$Firm),FUN=function(x)
  sum(x)))/(ave(blp$adj.price,FUN=function(x) length(x))-ave(blp$adj.price,factor(blp$Firm),FUN=
function(x) length(x)))

blp.reg.4<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z3+fat.z3+suger.z3,data=blp)

blp.reg.price.4<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z3+fat.z3+suger.z3+adj.
price.z3,data=blp)
```

f. 2SLS: using the average characteristics for all other brands besides product j.(This one is actually a bad idea as you will have perfect multicolinearity)

```
blp$cal.z4<-(ave(blp$cals,FUN=function(x)
  sum(x))-blp$cals)/(ave(blp$cals,FUN=function(x) length(x))-1)

blp$fat.z4<-(ave(blp$fat,FUN=function(x)
  sum(x))-blp$fat)/(ave(blp$fat,FUN=function(x) length(x))-1)

blp$suger.z4<-(ave(blp$suger,FUN=function(x)
  sum(x))-blp$suger)/(ave(blp$suger,FUN=function(x) length(x))-1)

blp$adj.price.z4<-(ave(blp$adj.price,FUN=function(x)
  sum(x))-blp$adj.price)/(ave(blp$adj.price,FUN=function(x) length(x))-1)

blp.reg.5<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z4+fat.z4+suger.z4,data=blp,
 model = TRUE)

blp.reg.price.5<-ivreg(delta~fat+suger+cals+adj.price|fat+suger+cals+cal.z4+fat.z4+suger.z4+adj.
price.z4,data=blp, model = TRUE)

stargazer(blp.reg.2,blp.reg.3,blp.reg.4,blp.reg.5, type = "html", keep = c("fat","suger","cals",
"adj.price"),covariate.labels = c("fat","suger","calories","price difference"))
```

	Dependent variable:			
	delta			
	(1)	(2)	(3)	(4)
fat	-0.042	0.067	-0.062	-0.015
	(0.076)	(0.116)	(0.097)	(0.058)
suger	-0.040**	-0.008	-0.046*	-0.032**
	(0.019)	(0.030)	(0.025)	(0.014)
calories	0.005	-0.008	0.007	0.002
	(0.004)	(0.007)	(0.006)	(0.002)
price difference	0.576	-1.809*	1.023	
	(0.478)	(0.998)	(0.766)	
Observations	50	50	50	50
R^2	-0.396	-1.992	-1.176	0.107
Adjusted R^2	-0.520	-2.258	-1.369	0.049
Residual Std. Error	0.619 (df = 45)	0.907 (df = 45)	0.773 (df = 45)	0.490 (df = 46)
Note:				*p<0.1;* ***p<0.05;*** p<0.01

```
stargazer(blp.reg.price.2,blp.reg.price.3,blp.reg.price.4,blp.reg.price.5, type = "html", keep =
c("fat","suger","cals","adj.price"),covariate.labels = c("fat","suger","calories","price differe
nce"))
```

	Dependent variable:			
	delta			
	(1)	(2)	(3)	(4)
fat	-0.033	-0.009	-0.022	-0.005
	(0.069)	(0.058)	(0.062)	(0.057)
suger	-0.038**	-0.031**	-0.034**	-0.029**

	(0.017)	(0.014)	(0.015)	(0.014)
calories	0.004	0.001	0.002	0.001
	(0.003)	(0.003)	(0.003)	(0.003)
price difference	0.395	-0.136	0.141	-0.224
	(0.368)	(0.174)	(0.283)	(0.149)
Observations	50	50	50	50
R^2	-0.177	0.143	0.036	0.150
Adjusted R^2	-0.281	0.067	-0.050	0.074
Residual Std. Error (df = 45)	0.569	0.485	0.515	0.483
Note:		*p<0.1;* ***p<0.05;*** p<0.01		

(g)First Stage F statistics for each model

	IV 1	IV 2	IV 3	IV 4
F-stat	2.7181917	1.2193006	1.5353191	NaN

	IV 1	IV 2	IV 3	IV 4
F-stat	3.0792586	29.6315611	4.7845599	6.6394585^{29}