Streaming Services

Michael Chan

Introduction

Streaming services have become a common service that offer consumers the ability to watch their favorite television or movie content on demand. Firms that provide these services have varying choices to make when trying to maximize their profit. These choices include price to consumers of the service provided, advertising space provided for commercials, and selling customer usage data. When trying to understand how a firm should maximize profit, the representation of how a firm may calculate their profit is:

$$\text{Profit} = \text{Revenue} - \text{Cost}$$

$$= P \times Q + \text{ADV} \times Q + \text{INFO} \times Q - \text{Cost}$$

The revenue variables with P representing price to customers, ADV representing number of commercials sold to advertisers, and INFO representing how much data is sold are all affected by the decisions streaming firms make when providing their services. When focusing on revenue, there are questions that a firm may focus on to help drive decisions to maximize profit. These include finding optimal prices since consumers tend to lower quantity (Q) demanded for higher priced products, an aversion to commercial breaks, and a preference for data privacy.

Costs can become quite expensive for firms, since licensing content such as television shows and movies can become quite costly as more watchable content becomes available for viewers in an attempt to draw in more subscriptions.

Understanding the individual demand for consumers is important when addressing how to adjust these decisions to increase profits for firms. Utility theory is an important concept when trying to model consumer preferences for firms that provide streaming services. A logit model for random utility models, and an extension called the mixed logit, are used to explore how consumer demand is affected by the discussed variables.

Data

The dataset includes panel data that examines 100 observations of people over multiple periods making a choice about which streaming service bundle of features they would prefer. There is also a choice to have no streaming service. Each person is observed in 11 separate situations (or periods) making one decision out of the five varying bundles that they are presented. No streaming is one of choice bundles available. The variables in each bundle include the variable price, which is the only continuous variable in the bundle sets. The rest of the variables are dummy variables including sharing exclusively user usage data, sharing all user data, whether the service has commercials, high speed delivered content, more television content, more movie content, and finally a no service variable representing a no streaming service choice.

First Model

The basic random utility model takes a basic form that will be estimated as a logistic regression problem since there is a discrete choice being made by each person in a given choice-set. Since we are trying to model individual utility of consumers, we can represent utility for person i's choice j as:

$$U_{\text{ij}} = x'_j\beta + \varepsilon_{\text{ij}}$$

Using this representation of utility, we can then show the probability that a person will choose choice j with J number of choices as:

$$\Pr(\text{choose } j) = \frac{\exp\left(x'_j\beta\right)}{\exp\left(x'_j\beta\right) + \ldots + \exp\left(x'_J\beta\right)}$$

Standard errors for this model will be estimated with a non-parametric bootstrapping method.

Final Model

A potential issue with the random utility model in the previous section is that the model has constant β effects on utility for all the variables in the bundle. However, it is possible that preferences for certain attributes in bundles can differ. The extension for this random utility model involves a latent class model that enables some coefficients to be fixed, and some to come from a random distribution. In this model, the variables price and commercials will have coefficients estimated as random in what is callled a mixed logit model with latent classes. The model is similar to the original random utility, but distinguishes itself with the random coefficients categorizing customers into two different types that will be classified as type A and type B customers. The new parameters that will be estimated in this latent class model include:

π_A, π_B for the share of the population for the types of customers.

β^A, β^B for the differing coefficients for the difference types.

α vector of coefficients for the variables designated as fixed.

This model can also be estimated with maximum likelihood estimation by maxmizing the function:

$$\text{LL}(\theta) = \sum_{i=1} \ln\left[\pi_A L_i(\beta^A) + \pi_B L_i(\beta^B)\right]$$

Standard errors will also be estimated with a non-parametric bootstrapping method similar to the previous model.

Results

Estimation of the random utility model without the latent class extension resulted in estimates similar to NERA's streaming service.

```
variables = {'price','shareusageonly','shareall','speed','noservice','commercials','moretv','mo
RUM_betas =  [-0.1033   -0.0503   -0.2231    0.3210   -1.7809   -0.2753    0.0368    0.5155]';
RUM_SE = [0.0146605597221620 0.0802686930376359 0.0793916329994764 0.0639508294770176 0.0633028
T = table(RUM_betas,RUM_SE,'RowNames',variables)
```

T = 8×2 table

	RUM_betas	RUM_SE
1 price	-0.1033	0.0147
2 shareusageonly	-0.0503	0.0803

	RUM_betas	RUM_SE
3 shareall	-0.2231	0.0794
4 speed	0.3210	0.0640
5 noservice	-1.7809	0.0633
6 commercials	-0.2753	0.0843
7 moretv	0.0368	0.0781
8 moremovies	0.5155	0.1972

At a glance, these estimates pass a quick check in sign direction. The results imply that customers dislike paying higher prices, having their data shared, and commercials with coefficients statistically significant from zero at a 95% level. The remaining coefficients that are positive also all make sense, speedier service, more content in the form of television or movies results in more attractive bundles for customers. However, the standard error estimated for more television resulted in a coefficient that we cannot conclude is statistically difference from zero.

Estimation of the latent class extension with the mixed logit model returns values as follows:

```
fixed_variables = {'shareusageonly','shareall','speed','noservice','moretv','moremovies'};
fixed_coefficients_alpha = [-0.0310    -0.2316    0.3176    -3.4573    0.0477    0.5406]';
alpha_SE = [0.0734    0.0841    0.0600    0.2021    0.0871    0.0815]';
table(fixed_coefficients_alpha,alpha_SE,'RowNames',fixed_variables)
```

ans = 6×2 table

	fixed_coefficients_alpha	alpha_SE
1 shareusageonly	-0.0310	0.0734
2 shareall	-0.2316	0.0841
3 speed	0.3176	0.0600
4 noservice	-3.4573	0.2021
5 moretv	0.0477	0.0871
6 moremovies	0.5406	0.0815

```
random_variables = {'price','commercials'};
type_A = [-0.397352092613121 -0.305084000377067]';
A_SE = [0.0352 .2012]';
type_B = [0.0342731279033918 -0.296841155200895]';
B_SE = [.0412 .8320]';
table(type_A,A_SE,type_B,B_SE,'RowNames',random_variables)
```

ans = 2×4 table

	type_A	A_SE	type_B	B_SE
1 price	-0.3974	0.0352	0.0343	0.0412
2 commercials	-0.3051	0.2012	-0.2968	0.8320

```
probability = {'probability'};
type_A_pr = 0.409609856981468;
```

```
type_b_pr = 0.590390143018532;
table(type_A_pr,type_b_pr,'RowNames',probability)
```

ans = 1×2 table

	type_A_pr	type_b_pr
1 probability	0.4096	0.5904

While the signs and significance of the coefficients in the fixed variables remain the same, the two types of consumers that we estimated with the different classes ended up with type B consumers with statistically insignificant results for price and commercial coefficients. Type A customers also have a an estimate that was not found to be statistically significant from zero with commercials, but kept a similar effect on utility.

While this approach is more interesting, the original random utility model seems to be a more reliable estimate for understanding consumer preferences.

Code

```
D = readtable('videodata.csv');

Y = D{:,'choice'};
X = D{:,'price','shareusageonly','shareall','speed','noservice','commercials','moretv','moremovies'};

est_mle = estimRUM(Y,X,5);

%bootstrap

n = 1000;
betas = zeros([n 8]);

for i = 1:n
Dnew = datasample(D,size(D,1));

Ynew = Dnew{:,3};
Xnew = Dnew{:,4:11};
ESTMLE_bs = estimRUM(Ynew,Xnew,5);
betas(i,:) = ESTMLE_bs.b';
end

RUMse = std(betas);

Xf = D{:,{'shareusageonly','shareall','speed','noservice','moretv','moremovies'}};
Xr = D{:,{'price','commercials'}};

pers = D.person;
estLC = estimLC(Y,Xf,Xr,pers,5,2)

LCbetas = struct();
LCbetas.oneone = zeros([n 1]); %just naming these as the position that the beta values are in the estLC.beta o
LCbetas.onetwo = zeros([n 1]);
LCbetas.twoone = zeros([n 1]);
LCbetas.twotwo = zeros([n 1]);
LCbetas.threeone = zeros([n 1]);
LCbetas.threetwo = zeros([n 1]);
LCalphas = zeros([n 6]);
LCpi = zeros([n 2]);
```

```
for i = 1:n
Dnew = datasample(D,size(D,1));

Ynew = Dnew{:,3};
Xnew = Dnew{:,4:11};
Xfnew = Dnew{:,{'price','shareusageonly','shareall','speed','noservice'}};
Xrnew = Dnew{:,{'commercials','moretv','moremovies'}};
ESTMLE_bs = estimLC(Ynew,Xfnew,Xrnew,pers,5,2);
LCbetas.oneone(i,:) = ESTMLE_bs.beta(1,1);
LCbetas.onetwo(i,:) = ESTMLE_bs.beta(1,2);
LCbetas.twoone(i,:) = ESTMLE_bs.beta(2,1);
LCbetas.twotwo(i,:) = ESTMLE_bs.beta(2,2);
LCalphas(i,:) = ESTMLE_bs.alpha';
LCpi(i,:) = ESTMLE_bs.pi';
disp('iteration') %just so I can keep an eye on this while I do other things..
disp(i)
end

alphasSE = std(LCalphas)
LCbetasSE = struct()
LCbetasSE.oneone = std(LCbetas.oneone)
LCbetasSE.onetwo = std(LCbetas.onetwo)
LCbetasSE.twoone = std(LCbetas.twoone)
LCbetasSE.twotwo = std(LCbetas.twotwo)
LCpiSE = std(LCpi)
```