Refining Taste: An Analysis of Preference Formation and Persistent Consumption of Food

Lindsey Currier, Jon Greenberger, Eric Li, Johnny Ma, Kevin Tse
University of Chicago

June 3, 2016

Introduction

Investigating taste formation for food, with special consideration towards how initial exposure to food affects future behavior.

Builds upon prior literature surrounding preference formation for cultural goods.

- Becker mentions food addiction in his landmark paper!
- Our model is unique in examining the effects of initial exposure.

Using Yelp data as a modern method of tracking individual level food consumption.

Can shed light on preference formation for quality goods in general, where rigorous work is desperately needed.

Theory Roadmap and Literature Review

Three main models of preference formation have previously been proposed:

1. Habit Formation
2. Learning By Consumption
3. Rational Addiction

Habit Formation

- Initially proposed by Houthhaker and Taylor (1970)
- Consumers build habit of consumption, and face utility loss if they deviate.
- Consumers are not forward looking.

$$D_t = \beta_0 + \beta_1 D_{t-1} + \beta_2 P_t + \beta_3 I_t + X'\gamma + U_t$$

- D_t is demand at time t, P_t is price, and X' is vector of controls.

Learning By Consumption

- Initially proposed by Levy-Garboua and Mont-marquette (1996)
- Consumers discover their preferences through repeated consumption.
- Still not forward looking.

$$S_t = (1 - \delta)S_{t-1} + \epsilon_t$$

$$D_t = \beta_0 + \beta_1 S_t + \beta_2 P_t + X'\gamma + U_t$$

- $1 - \delta$ represents memory decay, S_t is accumulated experience, and ϵ is the shock.

Rational Addiction

- Initially proposed by Stiglitz (1977) and pioneered by Becker and Murphy (1988).
- Features forward looking, infinitely lived agents.
- Agents build up "consumption Capital" i.e addiction which affects future utility.
- Agents are aware of this, and rationally maximize lifetime utility, taking this effect into account.

Model Overview

We seek to analyze *habit formation* and *learning-by-consumption* with respect to food preferences using the following four models. Using Yelp reviews, we are able to pinpoint individuals' initial exposure to a cuisine.

We focus primarily on Asian Cuisines, the models are:

1. Cuisine-Specific Probit Learning-by-Consumption
2. Asian-General Probit Learning-by-Consumption
3. Habit Formation
4. General Learning-by-Consumption

Notes for Models

- We consider "initial" exposures to be first time visits during 2013 and 2014.
  - Using 2015 data, we examine how this shock affects preference.
- We subset our data to two time periods: 2014 and 2015.
  - These years contain the mass majority of reviews.
- These models were both informed by literature and summary statistics of our data.

Probit Learning by Consumption

We regress on whether consumers revisits a particular Asian cuisine, upon first exposure to that same cuisine.
(i.e. Thai $\rightarrow$ Thai).

$$Y_1 = \begin{cases} 1, \text{if } \beta_0 + \beta_1 R_0 + \gamma_3 O_{2014} + \gamma_4 C_1 + ... + \gamma_9 C_6 + U > 0 \\ 0, \text{ if otherwise} \end{cases}$$

$$\Pr(Y_1 = 1) = \Phi(\beta_0 + \beta_1 R_0 + \gamma_3 O_{2014} + \gamma_4 C_1 + ... + \gamma_9 C_6)$$

Broad Probit Learning by Consumption

We also investigate whether consumer preferences for Asian food in general is affected by experience with a new cuisine.
(i.e. Thai $\rightarrow$ any Asian).

$$\Pr(Y_2 = 1) = \Phi(\beta_0 + \beta_1 R_0 + \gamma_3 O_{2014} + \gamma_4 C_1 + ... + \gamma_9 C_6)$$

Habit Formation

For consumers who first experienced a new Asian Cuisine prior to 2014, we examine if 2015 consumption behavior is dictated by habits formed in 2014.

$$D_{2015} = \beta_0 + \beta_1 D_{2014} + \gamma_1 P_{2015} + \gamma_2 Q_{2015} + \gamma_3 O_{2014} + \gamma_4 C_1 + \ldots + \gamma_9 C_6 + U_{201}$$

General Learning by Consumption

We now apply the traditional learning-by-consumption model on consumption behaviors in 2015.

ρ_0 is the difference between user's initial review of new cuisine, R_0, and their average review for Asian restaurants. This is the "shock" in learning by consumption model.

$$D_{2015} = \beta_0 + \beta_1 \rho_0 + \gamma_1 P_{2015} + \gamma_2 Q_{2015} + \gamma_3 O_{2015} + \gamma_4 C_1 + \ldots + \gamma_9 C_6 + U_{2015}$$

Data

- Data is taken from round 7 of "Yelp Data Challenge"
- We have 2.2M reviews, 552k users, 77k businesses, from 2006-2015.
- Isolated cities to Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison.
- Considered Asian cuisines: Chinese, Indian, Japanese, Korean, Mongolian, Thai, and Vietnamese.
- Reviews are on 5 star scale, with half point increments.

Data Justification

- Consider the trade-off between number of restaurants and "exotic" nature of cuisine.
- Asian cuisine is established enough for significant sample size.
- Asian Cuisine is foreign enough to isolate new consumers of Asian food.
- Inspired by Washington Post Article about Asian food's role in society.
- Grouping Asian cuisine may be problematic, but contains enough similar elements and typically co-exist.

Assumptions

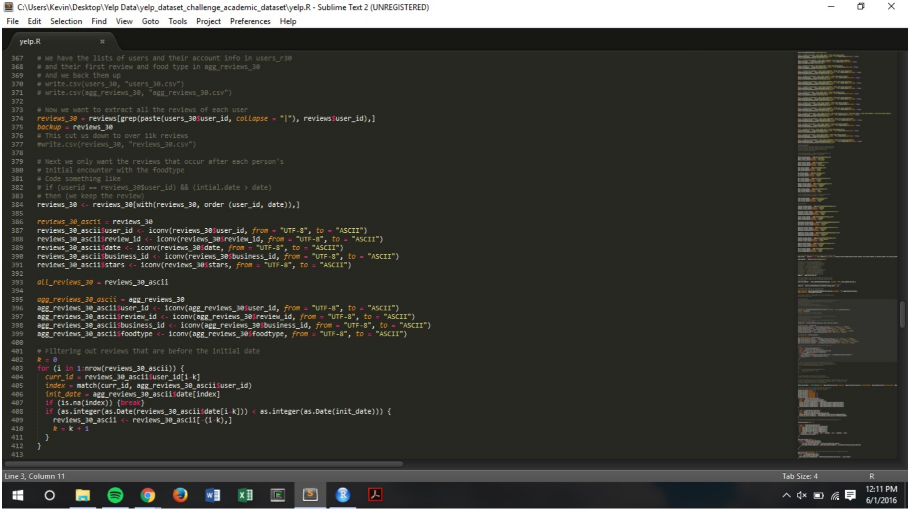
- Consumers review all restaurants that they go to.
  - To justify this, we isolated, and only considered chronic reviewers. (Davis, Dingle, and Monras, 2016) addresses problems with this assumption and why they are not significant.
- Rating of the restaurant reflects the user's true feelings about the cuisine experience (this will be discussed later).
- Users do not re-review restaurants. From manually looking through data, this is typically true.
- Reviewers are inhabitants in the city in which they review.
- Day of review is same as day of consumption.
  - Given that we group by year, should not be a problem.

Typical Yelp Reviewer

Methodology

- Chronic reviewers are defined as reviewers who had written over 30 reviews between 2014 and 2015.
- First time consumers isolated via two step procedure.
    1. Use pattern matching to identify reviews based on key phrases such as "first time eating" and "never had before".
    2. Identified reviews are manually sorted through to eliminate false positives.
- Record the types of cuisine to which the users are initially exposed
- Look at all their subsequent reviews to identify what types of restaurants the users visit after the initial exposure

R Codes


C:\Users\Kevin\Desktop\Yelp Data\yelp_dataset_challenge_academic_dataset\yelp.R - Sublime Text 2 (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
yelp.R
We have the lists of users and their account info in users_r30
and their first review and food type in agg_reviews_30
And we back them up
write.csv(users_30, "users_30.csv")
write.csv(agg_reviews_30, "agg_reviews_30.csv")

# Now we want to extract all the reviews of each user
reviews_30 = reviews[grep(paste(users_30$user_id, collapse = "|"), reviews$user_id),]
backup = reviews_30
# This cut us down to over 11k reviews
#write.csv(reviews_30, "reviews_30.csv")

# Next we only want the reviews that occur after each person's
# Initial encounter with the foodtype
# Code something like
# if (userid == reviews_30$user_id) && (intial.date > date)
# then (we keep the review)
reviews_30 <- reviews_30[with(reviews_30, order (user_id, date)),]

reviews_30_ascii = reviews_30
reviews_30_ascii$user_id <- iconv(reviews_30$user_id, from = "UTF-8", to = "ASCII")
reviews_30_ascii$review_id <- iconv(reviews_30$review_id, from = "UTF-8", to = "ASCII")
reviews_30_ascii$date <- iconv(reviews_30$date, from = "UTF-8", to = "ASCII")
reviews_30_ascii$business_id <- iconv(reviews_30$business_id, from = "UTF-8", to = "ASCII")
reviews_30_ascii$stars <- iconv(reviews_30$stars, from = "UTF-8", to = "ASCII")

all_reviews_30 = reviews_30_ascii

agg_reviews_30_ascii = agg_reviews_30
agg_reviews_30_ascii$user_id <- iconv(agg_reviews_30$user_id, from = "UTF-8", to = "ASCII")
agg_reviews_30_ascii$review_id <- iconv(agg_reviews_30$review_id, from = "UTF-8", to = "ASCII")
agg_reviews_30_ascii$business_id <- iconv(agg_reviews_30$business_id, from = "UTF-8", to = "ASCII")
agg_reviews_30_ascii$foodtype <- iconv(agg_reviews_30$foodtype, from = "UTF-8", to = "ASCII")

# Filtering out reviews that are before the initial date
k = 0
for (i in 1:nrow(reviews_30_ascii)) {
  curr_id = reviews_30_ascii$user_id[i-k]
  index = match(curr_id, agg_reviews_30_ascii$user_id)
  init_date = agg_reviews_30_ascii$date[index]
  if (is.na(index)) {break}
  if (as.integer(as.Date(reviews_30_ascii$date[i-k])) < as.integer(as.Date(init_date))) {
    reviews_30_ascii <- reviews_30_ascii[-(i-k),]
    k = k + 1
  }
}
Line 3, Column 11
Tab Size: 4
R
12:11 PM
6/1/2016

Methodology (Controls)

We control for both restaurant and user characteristics.

1. P_t is the average price of all (Asian) restaurants the consumer visited in period t. Price level is determined by yelp dollar sign system, with $ representing under $10, $$, $11-$30, $$$, $31-$60, and $$$$, above $61.
2. Quality, Q_t, is the average of aggregate ratings of all (Asian) restaurants a consumer visited in period t.
3. Control for location by assigning dummy varibles to the cities of the restaurants.
4. O_t is defined as the users average Yelp rating, as some reviewers may be generally harsher than others.

Summary Stats

Table: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Reviews/user after initial period	34	45.559	61.338	3	277
Average rating, all restaurants	34	3.723	0.554	2.167	4.700
Average rating, Asian restaurants	34	2.688	1.841	0.000	5.000
Initial rating of new cuisine	34	3.824	1.058	1	5
Visits to Asian restaurants, 2014	34	4.618	6.030	0	24
Visits to Asian restaurants, 2015	34	1.647	3.103	0	15

Table: Two Probit Models (Models 1, 2)

	Dependent variable:	
	Y_1	Y_2
	(1)	(2)
ρ_0 ("surprise")	-0.259	-0.099
	(0.1995)	(0.49992)
C_{PA} (location dummy for PA)	-4.347	-0.243
	(0.9934)	(0.79008)
C_{NC} (location dummy for NC)	-3.950	-6.005
	(0.9954)	(0.98880)
C_{AZ} (location dummy for AZ)	-0.053	-0.865
	(0.9421)	(0.15543)
Constant	-0.591^{**}	1.005^{***}
	(0.0432)	(0.00218)
Observations	34	34
Note: p-value shown in parenthesis. Legend:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

Probit Models Findings

- No significant results for the effect of "surprise" initial consumption
  1. p values of $.1995$ and $.49992$
- This suggests either the learning by consumption model does not hold in this case or first ratings are a poor measure of initial experience

Table: Habit Formation Model (Model 3)

	Dependent variable:
	D_{2015}
D_{2014} (visits to Asian restaurants, 2014)	0.355***
	(0.078)
P_{2015} (average price of Asian restaurants visited, 2015)	2.069*
	(1.014)
Q_{2015} (average aggregate ratings of Asian restaurants visited, 2015)	−0.487
	(0.401)
C_{PA} (location dummy for PA)	1.519
	(1.306)
C_{NC} (location dummy for NC)	0.102
	(1.592)
C_{AZ} (location dummy for AZ)	0.137
	(0.951)
Constant	−0.552
	(0.580)
Observations	33
R^2	0.649
Adjusted R^2	0.568
Residual Std. Error	2.064 (df = 26)
F Statistic	8.021*** (df = 6; 26)
Note: standard errors shown in parenthesis. Legend:	*p<0.1; **p<0.05; ***p<0.01

Table: Learning-by-Consumption Model (Model 4)

	Dependent variable:
	D_{2015}
ρ_0 ("surprise")	-0.832^{***}
	(0.276)
P_{2015} (average price of Asian restaurants visited, 2015)	2.012
	(1.189)
Q_{2015} (average quality of Asian restaurants visited, 2015)	0.325
	(0.452)
C_{PA} (location dummy for PA)	4.343^{**}
	(1.775)
C_{NC} (location dummy for NC)	1.814
	(2.085)
C_{AZ} (location dummy for AZ)	1.139
	(1.159)
Constant	-1.979
	(2.963)
Observations	33
R^2	0.536
Adjusted R^2	0.406
Residual Std. Error	2.421 (df = 25)
F Statistic	4.128^{***} (df = 7; 25)
Note: standard errors shown in parenthesis. Legend:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Habit Models Findings

- Regressions for models 3 and 4 give significant results
- Model 3: the coefficient for D_{2014}, consumption in 2014, is $.355$, suggesting that current consumption of a new cuisine is strongly correlated with future consumption of the same cuisine
    1. supports the *habit-formation* theory
    2. the coefficient to price is positive and significant: $B = 2.069$
- Model 4: the coefficient for ρ_0, "surprise," is $-.832$, the opposite of what we expected
    1. odd results may be due to failure of R_0 to represent preference shock. We find R_0 is highly correlated to the user's previous rating

TABLE 5. Validity of Initial Rating as Proxy for Experience

	Dependent variable:
	R_0
Q_{2015} (average quality of Asian restaurants visited, 2015)	0.120
	(0.212)
P_{2015} (average price of Asian restaurants visited, 2015)	0.151
	(0.159)
O_{2014} (overall rating behavior)	0.856***
	(0.161)
C_{PA} (location dummy for PA)	0.392
	(0.824)
C_{NC} (location dummy for NC)	−0.252
	(0.815)
C_{AZ} (location dummy for AZ)	0.085
	(0.793)
C_{NV} (location dummy for NV)	−0.196
	(0.790)
C_{WI} (location dummy for WI)	0.457
	(1.107)
C_{QC} (location dummy for QC)	
Constant	0.016
	(1.365)
Observations	116
R^2	0.254
Adjusted R^2	0.199
Residual Std. Error	0.774 (df = 107)
F Statistic	4.564*** (df = 8; 107)

Note: standard errors shown in parenthesis. Legend: *p<0.1; **p<0.05; ***p<0.01

Acknowledgements and Thanks

Thanks to Sebastien Gay and Kyle Kost for being great teachers and giving us the methods to run these regressions.

Special Thanks to Jim Marrone (Ph.D. candidate, Economics) and Jonathan Dingel (Assist. Professor at Booth) for help in conceptualization and methodology.

Thanks to Colin Camerer, Melissa Tartari, Sylvia Klosin, and Cameron Taylor for general advice.

Appendix

- Levy-Garboua, L. and Montmarquette, C. (1996), "A Microeconometric Study of Theatre Demand," *Journal of Cultural Economics* 20: 25-50, 1996.
- Becker, G., Grossman, M., and Murphy, K. (1994), "An Empirical Analysis of Cigarette Addiction," *The American Economic Review* Vol. 84, No. 3, pp. 396-418, Jun., 1994.
- Becker, G. and Stigler, G. (1977), "De Gustibus Non Est Disputandum," *The American Economic Review* Vol. 67, No. 2, pp. 76-90, Mar., 1977.
- Davis, D., Dingel, J., Monras, J., and Morales, E. (2016), "How Segregated is Urban Consumption?" Working Paper.
- Houthakker, Hendrik S., and Lester D. Taylor. (1970) Consumer Demand in the United States: Analyses and Projections. Cambridge, MA: Harvard UP, 1970. Web.
- Pollak, R. (1970), "Habit Formation and Dynamic Demand Functions," *Journal of Political Economy*, Vol: 78, No. 4, pp. 745-63, 1970.
- Marrone, J. (forthcoming), "Culture as a Habit: Assimilation and Language Learning Over the Lifecycle," Working Paper.
- Yamamura, E. (2008), "Socio-economic effects on increased cinema attendance: The case of Japan," *Journal of Behavioral and Experimental Economics* Vol: 37, No. 6, pp. 2546-2555, December 2008.