Topic 2:

Introduction to choice data and choice models

Takeaki Sunada[1]

[1]Simon Business School
University of Rochester

Utilizing choice data to estimate demand

- Thanks to technological advancement, more and more firms have access to individual-level data of purchase history.
- For example, loyalty cards of a retailer let us identify transaction by each individual customer. The data typically contain what each consumer bought at what price at each shopping trip.
- Unlike sales data, which are aggregated up to product-level, individual choice data of course provide much finer pieces of information. Moreover, oftentimes we have some demographic information from each customer (age, gender, residential zip code, etc).
- For the rest of the course, we study how we can extract full information out of choice data.

What is choice data?

- In general, choice data come with the following information.
  - Each consumer's ID.
  - For each ID and for each transaction that that person made, the UPC code of items purchased, number of units purchased and the price for each item, and the timestamp of the transaction.
  - For each ID, certain demographic information (depending on the context).

Assumption on the industry environment

- Throughout the rest of the course, we make a couple of assumptions.
- We assume that you are a manufacturer and a choice data set is provided by a retailer that carries your products.
- Hence we not only observe occasions where customers selected your product, but also the ones where they selected your rival's product over yours, and the ones where they bought nothing. This is the other dimension where choice data are better than sales data.

Assumption of single-unit demand

- We assume that people buy at most one unit per shopping trip.
- Quite often, this is what you see in the data.
    - Milk, dish soap, coffee creamer, etc... Not many people buy multiple bottles of them at once.
    - Of course there are exceptions - if you study Costco data, this probably isn't a right assumption.
- If you are a manufacturer of milk or dish soap, you don't lose much with this assumption, and it helps simplify models we develop later.

Choice data we use

id	week	trip	price.0	price.KB	price.KR	price.MB	choice
1	96	1	0	1.43	1.43	1.43	0
2	14	1	0	1.43	1.43	1.65	0
2	25	2	0	1.43	1.43	1.65	0
2	26	3	0	1.43	1.43	1.65	0
2	31	4	0	1.43	0.88	1.65	KB
2	89	5	0	1.43	1.43	1.32	KB
2	94	6	0	0.9	0.89	1.43	KR
2	96	7	0	1.43	1.43	1.43	0
2	114	8	0	1.43	1.43	1.34	MB
2	126	9	0	1.43	1.43	1.26	MB
2	146	10	0	1.47	1.43	1.33	KR

- For now, let's assume that "KB" is your product and "KR" and "MB" are other products. "0" means not buying anything.

Models of individual purchase decisions

- By nature of choice data, the demand we study is that of *each consumer* - we need a demand model that provides prediction of individual purchase decisions.
- In other words, our model now approximates individual consumer's decision process, rather than merely approximating overall demand for the product.
- Because we model demand at a finer scale, we can extract a lot more insights (segmentation, competitive analysis).

Regression models with choice data

- Before jumping into choice models, let's look at how regression models do in this environment. We can create "individual consumer version" of a regression model.
- In the data, we can see choice occasions where a consumer ended up buying our product (KB). Denote by q_{nt}^{KB} if an individual n purchases KB at shopping trip t (it is one if "buy" and otherwise zero).
- We also see the price of KB, P_{nt}^{KB}, at which consumers make each decision.

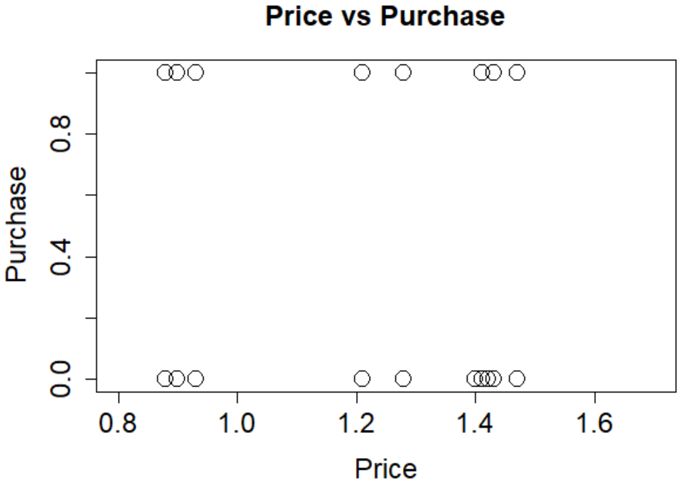
Regression models with choice data

- Then I can apply a regression model to the choice-data environment.

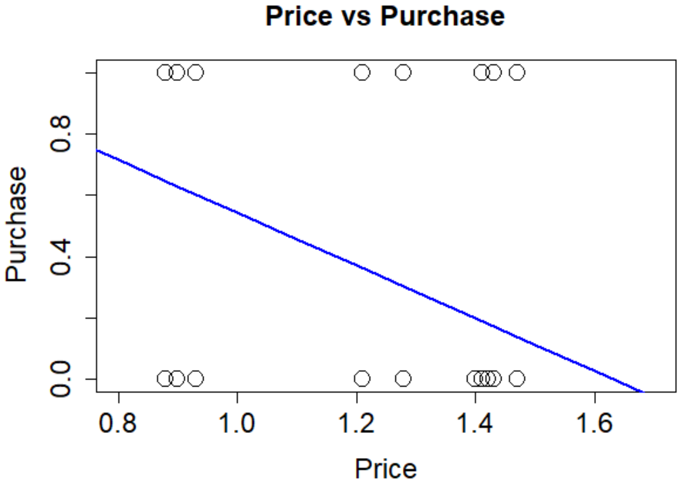
$$q_{nt}^{KB} = \beta_0 + \beta_1 P_{nt}^{KB} + \epsilon_{nt}$$

- Each consumer(n)-shopping trip(t) combination constitutes one observation.
- Note that when $q^{KB} = 0$, the consumer may buy other products (KR or MB), or buy nothing. We do not distinguish between them. Not buying KB is not buying KB (assumption we impose for the rest of this deck of slides).

Regression line



Regression line



Interpretation of regression models in choice data environment

$$q_{nt}^{KB} = \beta_0 + \beta_1 P_{nt}^{KB} + \epsilon_{nt}$$

- Note that dependent variable q_{nt}^{KB} is binary. In this case,
  1. What is the interpretation of the estimated regression line?
  2. How are we estimating β_1?
  3. What is the interpretation of the error term?

What is the interpretation of the estimated regression line?

- Formally, the regression line corresponds to the expected value of q^{KB} conditional on P^{KB}.

$$E(q^{KB} \mid P^{KB}) = \beta_0 + \beta_1 P^{KB} = Pr(q^{KB} = 1 \mid P^{KB}).$$

- But q^{KB} is either 0 or 1 in this context. Hence the expectation of q^{KB} represents the *probability* that $q^{KB} = 1$ at that P^{KB}.
- Hence regression models applied to choice data is often called *linear probability* regression model. "What is the probability that at price P^{KB} a consumer chooses KB?"
- By construction, the probability that a consumer doesn't buy KB at P^{KB} is $1 - Pr(q^{KB} = 1 \mid P^{KB})$.

How are we estimating β_1?

$$Pr(q^{KB} = 1 \mid P^{KB}) = \beta_0 + \beta_1 P^{KB}.$$

- In sales-data case, we estimated β_1 using "lower P is associated with higher Q". Hence β_1 was estimated by something like "as we decrease P by one, q^{KB} increases by three on average", which is no longer the case here.
- Here, q^{KB} is zero or one. Hence we instead estimate β_1 by "how many $q^{KB} = 1$ is associated with each P^{KB}". As P^{KB} goes down, more consumers start choosing KB, and we have more observations of $q^{KB} = 1$ when P^{KB} is lower.

What is the interpretation of the error term?

- Just like sales-data case, our prediction (choice probability) does not necessarily match with the actual realization.

$$\begin{aligned} q^{KB} &= \beta_0 + \beta_1 P^{KB} + \epsilon \\ &= Pr(q^{KB} = 1 \mid P^{KB}) + \epsilon. \end{aligned}$$

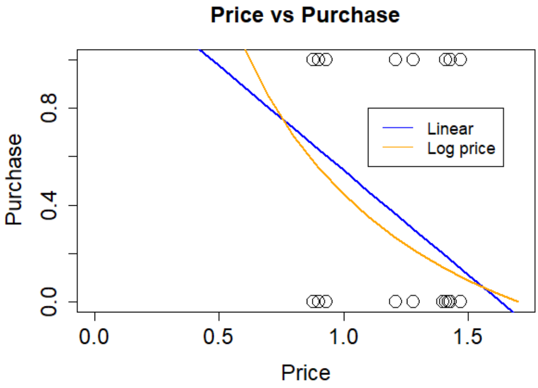
- The realization q^{KB} is always either 0 or 1. On the other hand, our prediction is choice probability, which is *between* 0 and 1. Hence there's an error between our prediction and the observed realization, which is ϵ.
- This is the same logic as in the case of sales data - ϵ captures "anything beyond price", but now in the context of individual's purchase decisions.

Note: Causality in choice data

- In the context of choice data, we tend to be less concerned about the non-causal relationship between P^{KB} and q^{KB}.
- It doesn't mean that we no longer need a causal relationship! Because we later manipulate the price and predict the change in sales, we still need a causal relationship. The data we use have changed, but our objective remains the same - so does this restriction.
- We are less concerned because usually, less contamination is involved between P^{KB} and q^{KB} of *an individual consumer*. How come a single consumer's purchase decision could possibly influence the firm's pricing decisions?
- Henceforth, we interpret any price coefficients from choice models as causal.

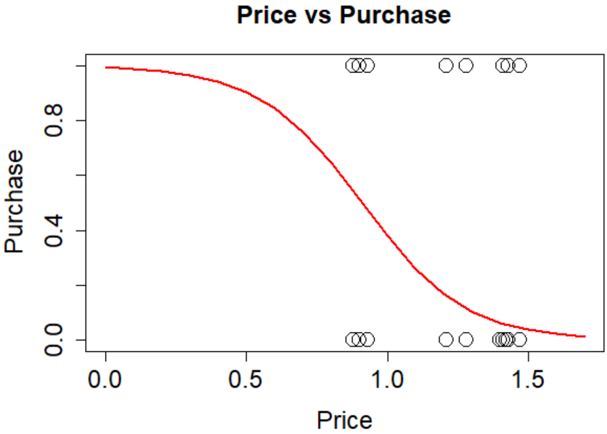
Model fit

- Clearly, regression models aren't quite suitable to predict choice probability of consumers.



- The linearity (or log-linearity) assumption is obviously wrong here.

Model fit



- We would probably prefer this sort of prediction...

Choice models

- Choice models enable us to utilize choice data to consider much richer demand systems.
- Today, we start from seeing that they offer a much better out-of-sample prediction (fit within zero-one).
- We consider the most basic version today, and gradually add bells and whistles over coming weeks.

Choice models

- The main problem with a regression model: the model prediction is not bounded within $[0,1]$ interval.
- Recall that regression line in the environment of choice data represents the probability that a consumer chooses KB at a price P^{KB}.

$$Pr(q^{KB} = 1 \mid P^{KB}) = \beta_0 + \beta_1 P^{KB},$$

- Because we are predicting the probability that an event occurs, we want to have a function bounded within $[0,1]$.

Choice models

- Let's consider a choice model, where the probability of purchase is modeled by the following function.

$$Pr(q^{KB} = 1 \mid P^{KB}) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})}.$$

- $\exp(x)$ is the same as e^x. We assume that $\beta_0 + \beta_1 P^{KB}$ determines $Pr(q^{KB} = 1)$ through the ratio of exponentials.
- I will call this model "binary logit model". "Binary" means that there are two outcomes (buy or not buy). "Logit" comes from the probability distribution this model is based on (we will cover this in an online content).

A binary logit model

$$Pr(q^{KB} = 1 \mid P^{KB}) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})}.$$

- $Pr(q^{KB} = 1)$ is always less than 1. For any β_0, β_1 and P^{KB}, the denominator is larger than the numerator.
- $Pr(q^{KB} = 1)$ is always positive. This is because $\exp(x)$ cannot go negative.
- Hence this "ratio of exponential" form is ideal for representing the probability that an event occurs.

A binary logit model

$$Pr(q^{KB} = 1 \mid P^{KB}) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})}.$$

- Because the model is nonlinear in β_0 and β_1, they do not correspond to any well-known concept anymore. However, how they work here remains the same as in regression models.
    1. Higher β_0 increases the probability that KB is selected by a consumer at any P^{KB}. It captures some sort of popularity (or consumers' inherent preference) of the product.
    2. Higher β_1 (closer to zero - note that $\beta_1 < 0$) implies that demand decreases less when price goes up. Consumers are less responsive to price change.

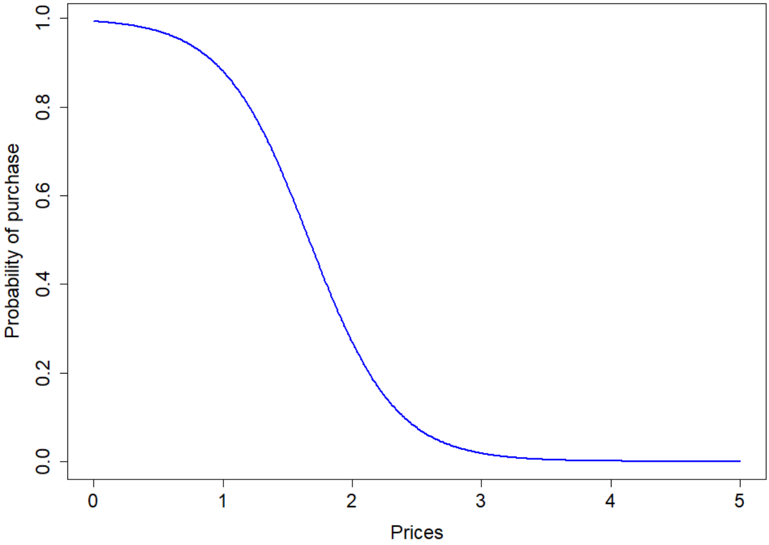
A numerical example

- Suppose that $\beta_0 = 5$ and $\beta_1 = -3$. What is the probability that a consumer buys "KB" at P between 0 and 5?

```
##
#Choice models
#Write choice probability as a function
demand=function(price,beta0,beta1){
  prob=exp(beta0+beta1*price)/(1+exp(beta0+beta1*price))
  return(prob)
}

pricespace=seq(0,5,0.01)
plot(pricespace,demand(pricespace,5,-3),type='l',xlab='Prices',
     ylab='Probability of purchase',col="blue",lwd=2
     ,cex=2,cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
```

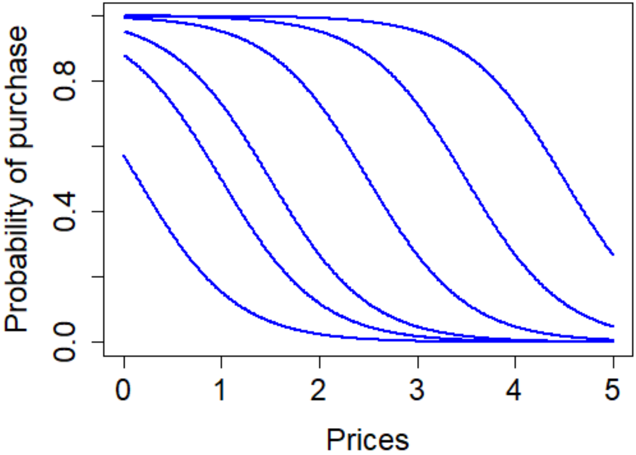
Choice probability



A binary logit model

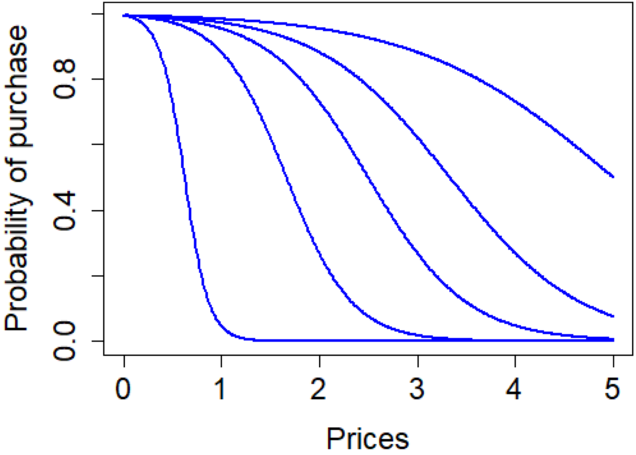
- This is a simple model with two parameters β_0 and β_1. But it is well consistent with our intuition of demand structure.
- By changing parameters, we can flexibly capture how the probability of purchase changes as prices change.

Different β_0



- Higher β_0 increases the probability of purchase at any P.

Different β_1



- β_1 closer to zero makes demand less responsive to price change (flatter slope).

A profit maximization problem

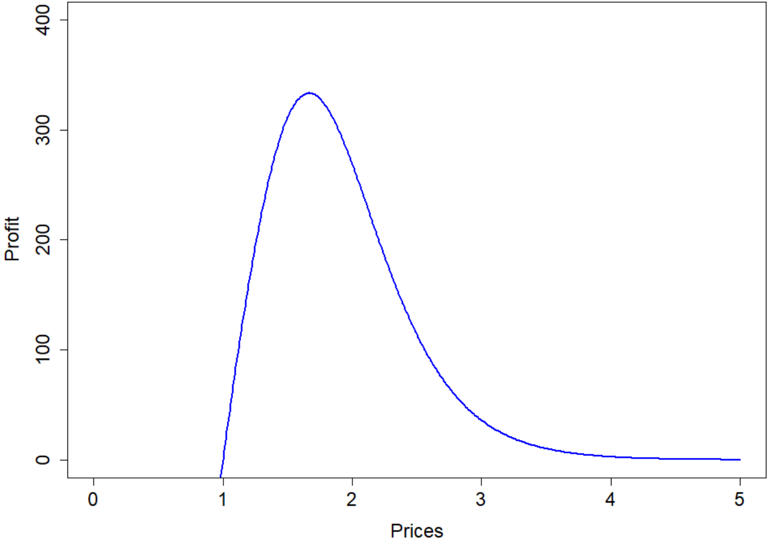
- Assume again $\beta_0 = 5$ and $\beta_1 = -3$. Also suppose that we have 1000 consumers in the market and the unit cost of the product is $uc = 1$. What is the profit-maximizing price?
- Recall profit = demand $\times$ (price - unit cost). Here the demand is "choice probability $\times$ number of consumers". Thus,

$$\text{Profit} = 1000 \times Pr(q^{KB} = 1 | P^{KB})(P^{KB} - 1).$$

```
#unit cost
uc=1

#What is profit maximizing price?
profit=1000*(demand(pricespace,5,-3)*pricespace-demand(pricespace,5,-3)*uc)
plot(pricespace,profit,type='l',xlab='Prices',
     ylab='Profit',ylim=c(0,400),col="blue",lwd=2,
     cex=2,cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
```

Prices and profit



Elasticity with a binary logit model

- One thing we did with regression models is to recover price elasticity of demand (recall log-log regression). Can we calculate demand elasticity with a binary logit model?
- It turns out, the elasticity in a binary logit model is given by a fairly simple formula.

$$\eta \equiv \frac{\frac{\partial Pr(q^{KB}=1)}{Pr(q^{KB}=1)}}{\frac{\partial P^{KB}}{P^{KB}}} = -\beta_1 P^{KB}(1 - Pr(q^{KB} = 1 \mid P^{KB})).$$

- Because the demand we now consider is the probability that a consumer buys KB, the elasticity is "one-percent change in price results in η percent changes in the probability of purchase".

Assumptions in a logit model

$$Pr(q^{KB} = 1 \mid P^{KB}) = \beta_0 + \beta_1 P^{KB}.$$

$$Pr(q^{KB} = 1 \mid P^{KB}) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})}.$$

- Model is a collection of assumptions we impose on the data to supplement missing observations.
- This simple binary logit is not any more flexible than the linear probability model - we have two parameters, β_0 and β_1 to represent the probability of purchase. We replaced the "linearity" assumption with "exponential ratio" assumption, which is more suitable for an environment of choice data.

Estimating logit models

- In practice, we need to first estimate parameters of the model (β_0, β_1) with the data, and then make demand predictions with the estimated parameters (just like regression models).
- One big problem in estimating a choice model is - well, it is not a regression model.
- Hence we cannot use a regression to estimate β_0 and β_1. A different approach is needed.

Estimating logit models

- "Maximum likelihood estimation" (MLE) is a useful, and sometimes essential, tool to estimate models that are nonlinear in parameters (i.e. models that cannot be written in a regression form).
- We discuss how it works later - but first, let's implement MLE to a binary logit model in R.
- We use two packages - package "gmnl" and "mlogit". The actual MLE algorithm is implemented in "gmnl". "mlogit" is used to pass the data to the algorithm.

Estimation with "gmnl"

- With "gmnl", think of the environment as follows - consumers face two options, one "KB" with price P^{KB}, and the other "option zero (outside option)", with price zero.

Estimation with "gmnl"

id	week	trip	price.0	price.KB	choice
1	96	1	0	1.43	0
2	14	1	0	1.43	0
2	25	2	0	1.43	0
2	26	3	0	1.43	0
2	31	4	0	1.43	KB
2	89	5	0	1.43	KB
2	94	6	0	0.9	0
2	96	7	0	1.43	0
2	114	8	0	1.43	0
2	126	9	0	1.43	0
2	146	10	0	1.47	0
3	1	1	0	1.43	0
3	2	2	0	1.43	0
3	4	3	0	1.43	0

kiwi_bubbles_binary

- We use a binary data set "kiwi_bubbles_binary.csv", which only records the choices of "KB or not".

Estimation with "gmnl"

- "gmnl" does not take our data set directly - we need to reshape our data set using "mlogit.data" function.

```
#Convert it to mlogit.data form. Column 4 is "price.0" and Column 5 is "price.KB".
mlogitdata=mlogit.data(data,id="id",varying=4:5,choice="choice",shape="wide")
```

- "mlogit.data" function takes the following arguments.
  - "data", which is the data.frame containing our observations.
  - "id", which specifies the name of the column that records each individual's ID.
  - "varying", which specifies the columns that contain option-specific attributes that vary across observations. In our case, we specify columns of prices (column 4 and 5).
  - "choice", which specifies the name of the column that tracks consumer's choices. In our case, it is "choice".
  - "shape", which we always set to "wide".

Estimation with "gmnl"

```
#Run MLE.
mle= gmnl(choice ~  price, data = mlogitdata)
summary(mle)
```

- Once the data is converted, running MLE is as simple as running a regression.

```
Coefficients:
                 Estimate Std. Error z-value  Pr(>|z|)
KB:(intercept)  4.94931    0.74781  6.6184 3.631e-11 ***
price          -4.48271    0.54976 -8.1539 4.441e-16 ***
```

Estimated demand (= probability of purchase)

- Once the estimation is done, I'd recommend to write the choice probability formula (the ratio of exponentials) as a function.
- The function takes prices and the estimated parameters as inputs and return the choice probability.
- Often, we need to write the choice probability expression multiple times (plotting, profit maximization, etc). With this function, we don't have to write out the whole expression every time.

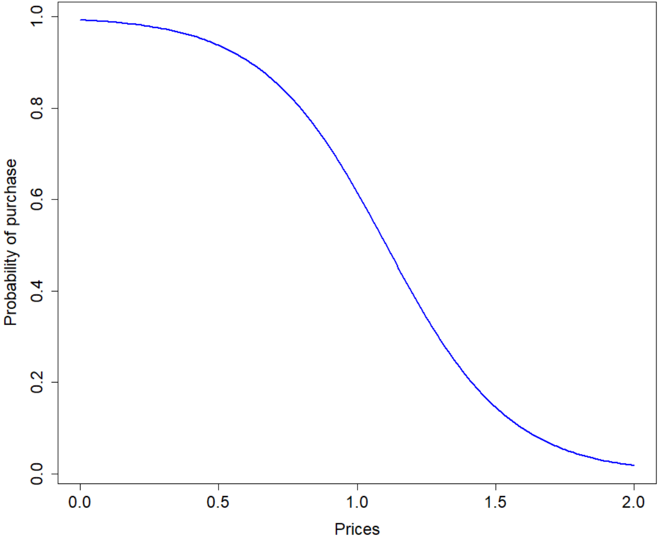
```
#Write choice probability as a function
demand=function(price,beta0,beta1){
  prob=exp(beta0+beta1*price)/(1+exp(beta0+beta1*price))
  return(prob)
}
```

Estimated demand (= probability of purchase)

- To draw figures, we use the generic "plot" function. On the X axis we provide a vector of candidate prices (at which we evaluate choice probability). On the Y axis we provide the corresponding choice probability (which comes from the "demand" function we defined above).

```
#Plot choice probability using the "demand" function
pricespace=seq(0,2,0.01)
plot(pricespace,demand(pricespace,coef[1],coef[2])
     ,type='l',xlab='Prices',ylab='Probability of purchase',col="blue",lwd=2
     ,cex=2,cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)
```

Estimated demand (= probability of purchase)



Flowchart of binary logit models with "mlogit"

1. From the data set, find an outcome variable (choices) and input variables (price). Put them into "mlogit.data" and then "gmnl". It gives you the estimator of β_0 and β_1.
2. In order to calculate optimal price, write the logit choice probability as a function.

$$Pr(q^{KB} = 1) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})},$$

   where you plug in β_0 and β_1 estimated by mlogit. Then the model lets you predict the probability that someone buys "KB" at any possible P^{KB}. Hence you can calculate profit, elasticity, etc.

An essence of maximum likelihood estimation

- So, how does "gmnl" estimate the model? Again, as the model is nonlinear in β_0 and β_1, we cannot estimate them by a regression.
- I will cover the essence of maximum likelihood estimation, which is used as an algorithm in "gmnl". It is also the most predominantly used estimation strategy for any nonlinear models.

An essence of maximum likelihood estimation

- To the extent that a model tries to predict outcomes observed in the data, we can calculate the "probability (likelihood) that the model explains the observed outcomes". The closer to one, the better the model fit is.
- Models with different parameter values (different β_0 and β_1) have different likelihood of explaining the outcome.
- The idea of MLE is "let's find the parameter values that maximize the likelihood that the model explains the data".

Maximum likelihood estimation in a coin-tossing problem

- Let's consider a following simple example.
- Consider tossing an uneven (=not 50/50) coin N times. As a result we observe outcome as a series of "heads" and "tails", such as [T,H,H,H,T].
- Suppose that we build a model, which says "this coin has an underlying true probability of coming up heads, θ, and it comes up tails with probability $1 - \theta$".
- Let's estimate θ from these observations using maximum likelihood.

Likelihood ($\approx$ probability)

- First, let's understand what the likelihood is. Likelihood is "the probability that the model with parameter θ explains the observed outcome".
- Given the model with a parameter θ, what is the likelihood of observing...
  - a coin tossed once coming up with head?
  - a coin tossed once and coming up with tail?
  - a coin tossed twice and coming out "head and then tail"?
  - a coin tossed three times and coming out "head-head-tail"?

Likelihood ($\approx$ probability)

- Given the model with a parameter θ, what is the likelihood of observing...
  - a coin tossed once coming up with head?
    L(H| θ)=θ.
  - a coin tossed once and coming up with tail?
    L(T| θ)=$1-\theta$.
  - a coin tossed twice and coming out "head and then tail"?
    L([H,T]| θ)=$\theta \times (1-\theta)$.
  - a coin tossed three times and coming out "head-head-tail"?
    L([H,H,T]| θ)=$\theta \times \theta \times (1-\theta)$.

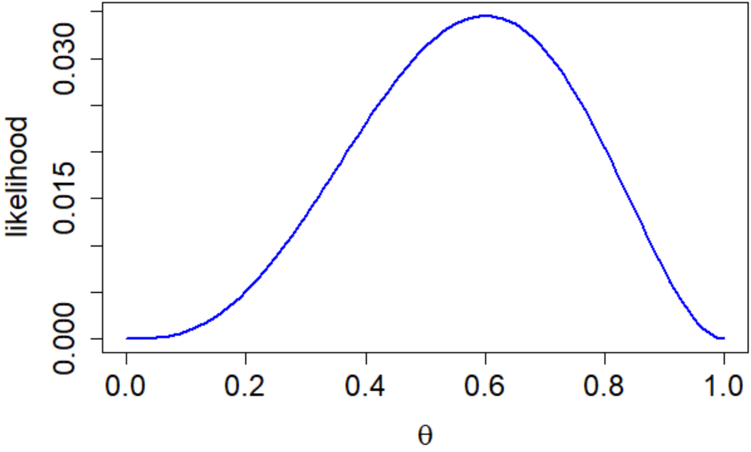
Maximizing likelihood

- In this course, we denote by $L(Y \mid \text{model input})$ the likelihood that a model assigns to an outcome Y.
- Maximum likelihood - find the value of the parameter θ that maximizes the probability that the model explains the observed outcomes.
- Suppose that we see [T,H,H,H,T] as an outcome. Then the likelihood that the model assigns to this outcome is

$$
\begin{aligned}
L([T,H,H,H,T] \mid \theta) &= (1-\theta) \times \theta \times \theta \times \theta \times (1-\theta) \\
&= \theta^3 (1-\theta)^2.
\end{aligned}
$$

- Our estimator of θ is the one that maximizes $L([T,H,H,H,T] \mid \theta)$. i.e. θ that "best explains the outcome in the data".

$L([T, H, H, H, T] \mid \theta)$ at different values of θ



Maximizing likelihood

- $\theta = 0.6$ maximizes the likelihood and hence is our estimator of θ (btw, the true θ that generated the data is 0.65, so our estimator isn't that far even with just 5 observations).
- What is the intuition of different likelihood values?
  - If $\theta = 0$, observations with tails get likelihood=1, but heads get likelihood=0.
  - If $\theta = 1$, observations with heads get likelihood=1, but tails get likelihood=0.
  - So we want to set θ somewhere in the middle. If more heads are observed in the data, set a higher θ, and vice versa.

Maximizing likelihood: a general case

- More generally, if we have N observations with M coming out head, then the likelihood is

$$L(M, N-M \mid \theta) = \theta^M (1-\theta)^{N-M}.$$

- With more heads observed in the data (higher M), we want to increase θ^M part to maximize this function, so we assign higher θ.
- Conversely, with less heads in the data, we want to increase $(1-\theta)^{N-M}$ part and hence assign lower θ.

MLE of a logit model

- Now let's go back to the logit case. In the data, we see if a consumer selected KB or not. This is analogous to coin-flip: it is "head" if one buys KB, and "tail" otherwise.
- In a binary logit model, the probability that someone chooses KB at price P^{KB} is

$$Pr(q^{KB} = 1 \mid P^{KB}) = \frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})},$$

  which is a function of parameters β_0 and β_1.
- This $Pr(q^{KB} = 1 \mid P^{KB})$ is our "θ" now - instead of directly estimating θ, we assume that θ is a function of β_0 and β_1, and estimate β_0 and β_1.

MLE of a logit model

- We plug in $Pr(q^{KB} = 1 \mid P^{KB})$ in place of θ in the previous example.
- Suppose that we observe someone making 5 shopping trips, and purchased KB in her second, third and fourth trip (just like [T,H,H,H,T]).

MLE of a logit model

- The likelihood of observing "$\{0, 1, 1, 1, 0\}$" is then:

$$
\begin{aligned}
& L(q^{KB} = \{0, 1, 1, 1, 0\} \mid P^{KB}) \\
& = Pr(q^{KB} = 1 \mid P^{KB})^3 \times (1 - Pr(q^{KB} = 1 \mid P^{KB}))^2 \\
& = \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})} \right)^3 \left(1 - \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})} \right) \right)^2.
\end{aligned}
$$

- "gmnl" is maximizing this function with a numerical solver.

MLE of a logit model

- More generally, if we have N observations and M ends up with purchase of KB, then the likelihood is

$$
\begin{aligned}
& L(M, N-M \mid P^{KB}) \\
& = Pr(q^{KB} = 1 \mid P^{KB})^M \times (1 - Pr(q^{KB} = 1 \mid P^{KB}))^{N-M} \\
& = \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})} \right)^M \left(1 - \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1 + \exp(\beta_0 + \beta_1 P^{KB})} \right) \right)^{N-M}.
\end{aligned}
$$

- In practice, P^{KB} that the consumer faced at each shopping trip may be different, so that the formula becomes more involved. But the key idea remains the same.

Intuition behind how β is estimated

$$L(M, N-M \mid P^{KB})$$
$$= \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1+\exp(\beta_0 + \beta_1 P^{KB})}\right)^M \left(1 - \left(\frac{\exp(\beta_0 + \beta_1 P^{KB})}{1+\exp(\beta_0 + \beta_1 P^{KB})}\right)\right)^{N-M}.$$

- β_0 is high when we have more observations of purchases *at any price levels*. If the product is more popular, we want to increase the first part of the likelihood, and hence assign a higher β_0.
- β_1 is estimated by how the proportion between $q^{KB} = 1$ and $q^{KB} = 0$ changes as P^{KB} changes. If the proportion decreases quickly as P^{KB} goes up, then we have large negative β_1.

Summary

- Choice models are a class of models that approximate each individual consumer's purchase decisions.
- A binary logit model is a preferred alternative to linear probability model in the environment of choice data. Its estimation relies on "MLE" method.
- Of course we just covered the simplest possible example - just "KB" or "not buy", and there's no consumer heterogeneity (segmentation) whatsoever. This limits our ability to study competition, right product-line allocation, etc.
- We will relax these assumptions in the following sessions.