ApEc 8213: Econometric Analysis III -- Lecture #4

Multinomial Response (Multiple Choice) Models
Hansen, Chapter 26

Lecture 2 introduced probit and logit models, where the dependent variable is a binary (dummy) variable that takes only two values, 0 or 1. This lecture extends this approach to more than two values.

I. Multinomial Response (26.2)

Consider the case where the dependent variable, Y, takes several integer values, from 1 to J: $Y \in \{1, 2, \ldots J\}$. These values may have an order, but **for now we assume no order**. One **example** is **transportation choices**: car, bicycle, bus, train, or airplane.

If there are no explanatory variables, the probability distribution of Y is full described by $P_j = \text{Prob}[Y = j]$. However, we usually want to know how some X variables predict or determine Y. For a vector of k X variables the pair (Y, X) is a **multinomial response model**. The distribution of Y, *conditional* on X, is the **response probability**:

$$P_j(x) = \text{Prob}[Y = j | X = x]$$

Another example of a multinomial response model is when Y is **marital status**, for which the possible outcomes are: married, divorced, separated, or never married. Hansen uses some data from the U.S. Current Population Survey (CPS) to estimate the explanatory power of age (in years) on people's marital status. He first estimates **simple (binary) logit** models **separately for each possible Y_j**. However, the **probabilities** at any age generally **do not add up to 1** (although they are close). This is **statistically inefficient**; multinomial models use the information that the probabilities across the different choices should add up to 1, and in fact are specified so that they add up to 1.

The **general set-up for a multinomial model** is to assume that **choices reflect people's underlying utility**, and that **individuals choose** the option that gives **the highest utility**. More specifically, let the "latent" utility from option j be:

$$U_j^* = X'\beta_j + \varepsilon_j \qquad (26.1)$$

where * indicates that U_j^* is not observed. Note that $\boldsymbol{\beta_j}$ **has a j subscript**, which indicates that the impact of the X variables on utility is different for the different options.

Question: What would happen if β_j were the same for all j?

While we do **not observe U_j^*,** we **do observe Y** (the option chosen). They are related as follows:

$$Y = j \text{ if } U_j^* \geq U_\ell^* \text{ for all } \ell$$

Question: Suppose we add $X'\gamma$ to U_j^* for all J options, so that $U_j^* = X'\beta_j + X'\gamma + \varepsilon_j = X'(\beta_j + \gamma) + \varepsilon_j$. Will Y change?

This shows that β_j is not identified. **Only the differences, such as $\beta_j - \beta_\iota$, are identified**. Thus **a normalization is needed**, and usually one of the β's, such as β_1 or β_J, is set equal to 0. This choice is called the **base alternative**, or **base option**. Thus, the actual ***β*'s estimated are always *relative β*'s**, relative to some base option.

Question: Suppose we multiply U_j^* by some positive constant, for all J options. Does Y change?

This suggests that **another normalization is needed**. The most **common choice** is to **fix the variance of one of the ε_j terms** to be equal to some constant, such as 1.

II. Multinomial Logit (26.3)

The **simple multinomial logit** model is:

$$P_j(x) = \frac{e^{x'\beta_j}}{\sum_{\ell=1}^{J} e^{x'\beta_\ell}} \qquad (26.2)$$

Question: What is the sum of these J probabilities?

This functional form arises from assuming that each ε_j in the U_j* equation **has the following individual univariate distribution and joint distribution**:

Definition 26.1. **Type I Extreme Value** distribution:

$$F(\varepsilon) = e^{-e^{-\varepsilon}} = \exp(-\exp(-\varepsilon))$$

Definition 26.2. **Generalized Extreme Value (GEV)** joint distribution function:

$$F(\varepsilon_1, \varepsilon_2, \ldots \varepsilon_J) = \exp\left(-\left[\sum_{\ell=1}^{J} e^{-\varepsilon_j/\tau}\right]^{\tau}\right) \qquad (26.3)$$

For $J > 1$ and $\tau = 1$, the joint distribution in (26.3) is **equal to the product of the univariate distributions** in (26.2), which means that those distributions are independent of each other. **If $\tau < 1$**, then **the different ε terms are correlated**, with a correlation equal to $1 - \tau^2$. The τ parameter is called the **dissimilarity** parameter. Note that some authors use "$1 - \sigma$" for τ, and call σ the "similarity" parameter.

The GEV distribution leads to the following theorem:

Theorem 26.1. Assume that the utility of option j equals $U_j^* = X'\beta_j + \varepsilon_j$, and that the error terms $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_J$ follow the GEV distribution function. Then the response probabilities (probabilities as a function of x) equal:

$$P_j(x) = \frac{e^{x'\beta_j/\tau}}{\sum_{\ell=1}^{J} e^{x'\beta_\ell/\tau}}$$

Recall that multiplying U_j^* by some constant does not change the probabilities for $P_j(x)$. This means that multiplying the β_j terms by some constant also does not change those probabilities, which in turn means that $\boldsymbol{\tau}$ **is not identified**. Thus, we might as well set $\tau = 1$.

Maximum Likelihood Estimation

If we assume that the ε terms follow a GEV distribution, we can use maximum likelihood estimation to estimate the (differences in the) β_j parameters. The **probability of the observed value of *Y* for a given observation** can be written as:

$$\pi(Y|\, X, \beta) = \prod_{j=1}^{J} P_j(X|\, \beta)^{1[Y=j]}$$

where $\beta = (\beta_1, \beta_2, \ldots \beta_J)$ and 1[ ] is an indicator function that equals 1 if the expression in the brackets is "true" and equals 0 if it is not "true".

We can **take the log** of $\pi(Y|X,\beta)$, and **sum over all *n* observations**, to get the **log-likelihood function**:

$$\ell_n(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J} 1[Y_i = j]\log(P_j(X_i|\beta))$$

The **maximum likelihood estimator** is the **value of β that maximizes the log-likelihood function**, denoted by $\hat{\beta}_{\text{mle}}$:

$$\hat{\beta}_{\text{mle}} = \underset{\beta}{\text{argmax}}\, \ell_n(\beta)$$

To find the value of β that maximizes the log-likelihood function, you **need to search for it** using numerical optimization methods. Since the likelihood function is **globally concave**, such optimization should quickly find a maximum.

Note also that we **need a normalization** here, so **one of the β_j's** needs to be specified as being a **vector of zeroes**.

Marginal Effects

Unlike linear models, the marginal effects, denoted by $\delta_j(x)$, are **not constants** but instead are functions of the data:

$$\delta_j(x) = \partial P_j(x)/\partial x = P_j(x)\left(\beta_j - \sum_{\ell=1}^{J}\beta_\ell P_\ell(x)\right) \qquad (26.4)$$

Note: $\delta_j(x)$ is a vector of k derivatives, and β's are $k\times 1$ vectors.

The estimate of $\delta_j(x)$, denoted by $\hat{\delta}_j(x)$, is estimated by replacing β_j and $P_j(x)$ in (26.4) with their estimated values. Of course, this is for only one specific value of x, so it is **useful to calculate average marginal effects (AME) over all observations**:

$$\widehat{\text{AME}}_j = (1/n) \sum_{i=1}^{n} \hat{\delta}_j(x_i) \qquad (26.5)$$

Standard errors are calculated by delta method or bootstrap.

III. Conditional Logit (26.4)

The X variables in the (simple) multinomial logit model vary over individuals, but not over the different choice options. Yet **in many situations important determinants of choice can vary across these options**. Perhaps the most obvious example is price, when the options are consumer purchases. The **conditional logit** model adjusts the (simple) multinomial logit to **allow for such *X* variables**.

In most basic conditional logit model has the following specification for the decision maker's latent utility:

$$U_j^* = X_j'\gamma + \varepsilon_j \qquad (26.6)$$

For this specification, ***X* varies across choices**, **but the γ coefficient does not vary**. For example, if one of the X_j

variables is price, the corresponding γ coefficient effectively measures the marginal utility of money, which should be the same for all choices.

The **multinomial** logit and the **conditional** logit **can be combined** to include both types of X variables. This is **often still called the conditional logit** (Wooldridge calls it a "mixed logit"). For it, U_j^* is:

$$U_j^* = W'\beta_j + X_j'\gamma + \varepsilon_j \qquad (26.7)$$

For this model, the β_j coefficient vectors are identified only relative to each other (by setting one of them to 0), as in the multinomial model, yet the **γ coefficient vector is identified**. More specifically the differences in the β_j parameters, and the γ parameters, are identified **up to a scale**, since multiplying both sides of (26.7) by some positive constant would not change the option chosen by the decision maker. **In most cases the scale is set by choosing the variance of one of the ε_j terms.**

The assumption that the ε_j terms follow a Type I extreme value distribution yields the following **probability response functions**:

$$P_j(w, x) = \frac{e^{w'\beta_j + x_j'\gamma}}{\sum_{\ell=1}^{J} e^{w'\beta_\ell + x_\ell'\gamma}} \qquad (26.8)$$

The corresponding **log-likelihood function** is:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} 1[Y_i = j] \log(P_j(W_i, X_i| \theta)$$

where $\theta = (\beta_1, \beta_2, \ldots \beta_J, \gamma)$ and $X_i = \{X_{1i}, X_{2i}, \ldots X_{Ji}\}$. The **maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{\text{mle}}$,** is defined as:

$$\hat{\theta}_{\text{mle}} = \underset{\theta}{\text{argmax}}\, \ell_n(\theta)$$

Again, it must be estimated by numerical optimization.

As with the (simple) multinomial logit, the marginal effects of a change in the *W* or *X* variables depend on the values of those variables. The **average marginal effects for the *W* variables** are calculated as:

$$\widehat{\text{AME}}_j = (1/n) \sum_{i=1}^{n} \hat{\delta}_j(w_i, x_i) \qquad (26.5')$$

where

$$\hat{\delta}_j(w_i, x_i) = \partial \hat{P}_j(w_i, x_i)/\partial w_i = \hat{P}_j(w_i, x_i)\left(\hat{\beta}_j - \sum_{\ell=1}^{J} \hat{\beta}_\ell \hat{P}_\ell(w_i, x_i)\right)$$

The **average marginal effects for the *X* variables** depend on whether the *X* variable pertains to the option of interest (e.g. the effect of a change in the price of a train ticket on the probability of taking the train) or to one of the other

options (e.g. the effect of a change in the price of a train ticket on the probability of taking the bus). The marginal effects **for a given observation** are calculated as:

$$\delta_{jj}(w, x) = \partial P_j(w, x)/\partial x_j = \gamma P_j(w, x)(1 - P_j(w, x)) \qquad (26.9)$$

$$\delta_{j\ell}(w, x) = \partial P_j(w, x)/\partial x_\ell = -\gamma P_j(w, x) P_\ell(w, x) \qquad (26.10)$$

The **average marginal effects** for the ***X* variables** are then:

$$\widehat{\mathrm{AME}}_{j\ell} = (1/n) \sum_{i=1}^{n} \hat{\delta}_{j\ell}(w_i, x_i) \qquad (26.5'')$$

In **some situations** it may be that the **impact of a choice characteristic *X*** (e.g. price) **could vary by** one of the **decision maker's characteristics *W*** (e.g. income). This is easily accommodated by adding an interaction term to the latent utility function. For example, if *X* and *W* are scalars, the latent utility function could be written as:

$$U_j^* = W\beta_j + X_j\gamma_1 + X_j W\gamma_2 + \varepsilon_j$$

IV. Nested Logit (26.5 and 26.6)

Both the (simple) multinomial logit model and the (generalized) conditional logit model can be criticized for having the **independence of irrelevant alternatives (IIA)**

property. To understand this problem, note that the relative probability of two options (alternatives), j and ℓ, is:

$$\frac{P_j(W,X|\,\theta)}{P_\ell(W,X|\,\theta)} = \frac{e^{W'\beta_j + X_j'\gamma}}{e^{W'\beta_\ell + X_\ell'\gamma}} \qquad (26.11)$$

This expression is called the **odds ratio**. The important point is that this ratio **does not depend on the characteristics of any of the other options**. In many cases, this seems unrealistic. For example, some transportation options, such as train and bus, are similar. One would expect that the relative probability of driving a car or taking a train would depend on the existence of a bus option (the "appearance" of this option would probably affect the probability of taking the train more than the probability of driving a car).

McFadden proposed the **nested logit** to **address this problem**. This model **assigns each alternative to a smaller number of groups**. Changing the notation slightly, there are now ***J* groups**, and within each **group *j*** there are ***k*$_j$ options**. This changes the latent utility function only slightly:

$$U_{jk}^* = W'\beta_{jk} + X_{jk}'\gamma + \varepsilon_{jk} \qquad (26.12)$$

The **bigger change** is in the **joint distribution of the ε_{jk} terms**. They have the following joint GEV distribution:

$$F(\varepsilon_{11}, \varepsilon_{12}, \ldots \varepsilon_{JKJ}) = \exp\left(-\sum_{j=1}^{J}\left[\sum_{k=1}^{K_j} e^{-\varepsilon_{jk}/\tau_j}\right]^{\tau_j}\right) \quad (26.13)$$

For **each group *j*** there is a **"dissimilarity" parameter τ_j**. **If all the τ_j** parameters were **equal to 1**, then we would be **back to the conditional logit**. When any group has **$\tau_j < 1$**, then the ε_{jk} terms for the options within that group are **positively correlated**. However, if two options are in different groups then they are not correlated. Note that **the τ_j parameters can be estimated**, unlike the τ in the standard multinomial logit model.

As in the standard (conditional) multinomial logit model, the **β terms are not identified**, but the differences between them are identified, so **one option should be chosen as the base option** and its β terms should be set to 0. The γ terms are identified. To sum up…

Theorem 26.2. Assume that the utility for option *jk* is given as in equation (26.12), and that the ε_{jk} terms follow the distribution in equation (26.13). Then the response probability P_{jk}, which equals $P_{k|j} \times P_j$, is given by:

$$P_{k|j} = \frac{e^{W'(\beta_{jk}/\tau_j) + X'_{jk}(\gamma/\tau_j)}}{\sum_{m=1}^{K_j} e^{W'(\beta_{jm}/\tau_j) + X'_{jm}(\gamma/\tau_j)}}$$

and

$$P_j = \frac{\left(\sum_{m=1}^{K_j} e^{W'(\beta_{jm}/\tau_j) + X_{jm}'(\gamma/\tau_j)}\right)^{\tau_j}}{\sum_{\ell=1}^{J}\left(\sum_{m=1}^{K_\ell} e^{W'(\beta_{\ell m}/\tau_\ell) + X_{\ell m}'(\gamma/\tau_\ell)}\right)^{\tau_\ell}}$$

The **log-likelihood function** can be written as follows, where $\theta = (\beta_{11}, \beta_{12}, \ldots \beta_{1K1}, \beta_{21}, \ldots \beta_{JKJ}, \gamma, \tau_1, \ldots \tau_J)$:

$$\ell_n(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j} 1[Y_i = jk](\log(P_{k|j}(W_i, X_i|\,\theta) + \log(P_j(W_i, X_i|\,\theta))$$

As above, the **maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{\mathbf{mle}}$,** is defined as:

$$\hat{\theta}_{\text{mle}} = \underset{\theta}{\text{argmax}}\, \ell_n(\theta)$$

Again, $\hat{\theta}_{\text{mle}}$ must be found using numerical optimization.

Marginal effects can be derived, but they are very messy.

In Section 26.7, Hansen presents the "mixed logit", which allows for random coefficients for the γ terms. This is rarely used, so this is optional material.

See Wooldridge (2010, pp.651-53) for control function and other approaches when some of the X or W variables are endogenous (Hansen does not discuss this).

V. Multinomial Probit (26.8 and 26.9)

The **multinomial logit** models are useful because the error term assumptions make it **relatively easy to estimate models with many options**. One can also specify that the error term in latent utility function follows a normal distribution, but this requires computationally demanding methods when the number of options (J) is > 3 *and* one allows those error terms to be correlated across choices.

To **start with a simple model**, **assume that ε_j is normally distributed** and that the **ε_j terms are not correlated** across choices. Hansen calls this the **simple multinomial probit**. The **latent utility** is the same as for multinomal logit models. More specifically, we have:

$$U_j^* = W'\beta_j + \varepsilon_j \qquad (26.16)$$

or

$$U_j^* = W'\beta_j + X_j'\gamma + \varepsilon_j \qquad (26.17)$$

with ε_j identically and independently distributed as N(0, 1). Strictly speaking, this simple multinomial probit model does not have the IIA property, but Hansen says that "its properties are similar to IIA." In particular, since ε_j is i.i.d. any two options are not correlated with each other after conditioning on observed variables and so are not

"close equivalents". In practice, the results are similar to those from the (non-nested) multinomial logit models.

The **identification properties are exactly the same as those for the multinomial logit models**: the $\boldsymbol{\beta_j}$ **coefficients are identified only relative to each other**, and so a **base category needs to be chosen** for which all β coefficients = 0, and both β_j and γ are identified only up to a scale (which is set by normalizing all the ε_j terms to have a variance equal to 1.

Unlike the multinomial logit, it is **not possible to express** the response probabilities, $\boldsymbol{P_j(W, X)}$**, directly**. However, they can be expressed as (one-dimensional) integrals:

Theorem 26.3. In the simple multinomial probit and simple conditional multinomial probit models the response probabilities equal:

$$P_j(W, X) = \int_{-\infty}^{\infty} \prod_{\ell \neq j} \Phi(W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma + v)\phi(\mathrm{v})dv \quad (26.18)$$

where $\Phi(v)$ and $\phi(v)$ are the standardized normal cumulative and density functions, respectively.

Here is the **derivation for** $\boldsymbol{P_j(W, X)}$. The individual chooses option j instead of option ℓ if $U_j^* > U_\ell^*$. The probability that this happens is:

$$\text{Prob}[U_j^* = W'\beta_j + X_j'\gamma + \varepsilon_j > U_\ell^* = W'\beta_\ell + X_\ell'\gamma + \varepsilon_\ell]$$

$$= \text{Prob}[W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma > \varepsilon_\ell - \varepsilon_j]$$

Treat ε_j as fixed and focus on ε_ℓ (which distributed as N(0, 1):

$$= \text{Prob}[\varepsilon_\ell < W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma + \varepsilon_j]$$

$$= \Phi(W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma + \varepsilon_j)$$

This compares option j to option ℓ, but for j to have higher utility than all other options **we need this relationship to hold for all $\ell \neq j$**. Since all ε's are assumed to be independent, this is just the product of the Φ terms for all $\ell \neq j$:

$$\prod_{\ell \neq j} \Phi(W'(\beta_j - \beta_\ell) + (X_j - X_\ell)'\gamma + \varepsilon_j)$$

This for a particular value of ε_j. The **last step** is to **integrate over all possible values of ε_j**. This gives (26.18).

The **log-likelihood** for this multinomial probit model is:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} 1[Y_i = j] \log(P_j(W_i, X_i | \theta)$$

where $\theta = (\beta_1, \beta_2, \ldots \beta_J, \gamma)$. The **maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{\text{mle}}$,** is defined as:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} \ell_n(\theta)$$

Again, it must be estimated by numerical optimization.

The assumption that the ε_j terms are uncorrelated is very restrictive. The **general multinomial probit allows them to be correlated**. The model for U_j^* is the same as in equations (26.16) and (26.17), but now we specify that the error terms in the vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_J) \sim \text{N}(0, \boldsymbol{\Sigma})$, with no constraint on $\boldsymbol{\Sigma}$. Unlike logit models, $\text{Var}(\varepsilon_j)$ can vary over j.

In fact, $\boldsymbol{\Sigma}$ cannot be estimated, but **a smaller "differenced" version can be estimated**. To see how this works. write the model for U_j^* in differenced form:

$$U_j^* - U_J^* = W'(\beta_j - \beta_J) + (X_j - X_J)'\gamma + \varepsilon_{jJ} \qquad (26.19)$$

where $\varepsilon_{jJ} = \varepsilon_j - \varepsilon_J$. Define the $(J-1)\times(J-1)$ covariance matrix for ε_{jJ} as $\boldsymbol{\Sigma}_J$. As with all the models in this lecture the scale of γ and the (differenced) β's cannot be identified, so **one of the diagonal elements of $\boldsymbol{\Sigma}_J$ is normalized**, usually to 2 (see Hansen for some details).

To estimate this model we **need to integrate over a $J-1$ dimensional multivariate normal distribution**. This is difficult for $J = 4$, and pretty much impossible for $J > 4$. So many applications of the multinomial probit focus on J

= 3. However, with faster computers these days it is possible to simulate these integrals for $J \geq 4$. See the references in Hansen for how this can be done. This is called **simulated maximum likelihood estimation**. Sometimes this works well, but sometimes the likelihood function is not concave and convergence does not occur.

VI. Ordered Response Models (26.10)

Up to this point we have assumed that **the 3+ options** to choose from **had no particular order**. But there are cases where there is a natural order, and this makes things much easier. Examples of this are:

1. Opinion polls (strongly disagree, disagree, agree, strongly agree)
2. Self-reported health (poor, fair, good, excellent)
3. Level of schooling (less than high school, high school graduate, some college, college graduate)

The standard approach here is to **start with a simple latent variable framework**:

$$U^* = X'\beta + \varepsilon$$

where ε follows some cumulative distribution function $G(\varepsilon)$.

As we will see, we do not need a constant term in X.
Y has J possible **ordered** options: 1, 2, … J.

Draw a picture to give the intuition.

We do not observe U^*, but we do observe Y. They are related as follows:

$$\begin{aligned} Y &= 1 \quad \text{if } U^* \le \alpha_1 \\ Y &= 2 \quad \text{if } \alpha_1 < U^* \le \alpha_2 \\ &\vdots \qquad\qquad \vdots \\ Y &= \text{J} \quad \text{if } U^* > \alpha_{J\text{-}1} \end{aligned}$$

where the α_j terms are thresholds, with $\alpha_1 < \alpha_2 < \ldots < \alpha_{J\text{-}1}$.

If ε is normally distributed, this is an **ordered probit**, while if ε follows a logistic distribution, it is an **ordered logit**. As with standard probits and logits, the β parameters are identified only up to a scale, so the variance of ε is normalized, e.g. to be 1 for an ordered probit.

The **response probabilities are**:

$$\begin{aligned} P_j(x) &= \text{Prob}[Y = j \mid X = x] \\ &= \text{Prob}[\alpha_{j\text{-}1} < U^* \le \alpha_j \mid X = x] \\ &= \text{Prob}[\alpha_{j\text{-}1} - X'\beta < \varepsilon \le \alpha_j - X'\beta \mid X = x] \\ &= G(\alpha_j - x'\beta) - G(\alpha_{j\text{-}1} - x'\beta) \end{aligned}$$

where we can write $\alpha_0 = -\infty$ and $\alpha_J = \infty$.

The **marginal effects** are:

$$\partial P_j(x)/\partial x = \beta[g(\alpha_{j-1} - x'\beta) - g(\alpha_j - x'\beta)]$$

where ***g*( ) is the density function** corresponding to $G(\ )$.

The **log-likelihood** for this multinomial probit model is:

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} 1[Y_i = j] \log(P_j(X_i|\ \theta)$$

where $\theta = (\beta, \alpha_1, \alpha_2, \ldots \alpha_{J-1})$. The **maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{\mathbf{mle}}$,** is defined as:

$$\hat{\theta}_{\text{mle}} = \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta)$$

Again, it must be estimated by numerical optimization.

Section 26.11 presents count models, which are like ordered models but with a potentially large number for *J*. Wooldridge (2010) presents this in much more detail.

Section 26.12 presents an extension of conditional logit models to market demand. This is used a lot in the industrial organization literature. This is optional reading.