Econometrics II TA Session #4

Hiroki Kato

1 Empirical Application of Ordered Probit and Logit Model: Housing as Status Goods

Breif Background. Social image may affect consumption behavior. Specifically, a desire to signal high income or wealth may cause consumers to purchase status goods. In this application, we explore whether living in an upper floor serves as a status goods.

Data. We use the housing data originally coming from the American Housing Survey conducted in 2013 [1]. We use the following variable

- `Level`: ordered value of a story of respondent's living (1:Low - 4:High)
- `Levelnum`: variable we recode the response `Level` as 25, 50, 75, 100. This represents the extent of floor height.
- `lnPrice`: logged price of housing (proxy for quality of house)
- `Top25`: a dummy variable taking one if household income is in the top 25 percentile in sample.

```
house <- read.csv(file = "./data/housing.csv", header = TRUE,  sep = ",")
house <- house[,c("Level", "lnPrice", "Top25")]
house$Levelnum <- ifelse(
  house$Level == 1, 25,
  ifelse(house$Level == 2, 50,
  ifelse(house$Level == 3, 75, 100)))
head(house)
```

```
##   Level  lnPrice Top25 Levelnum
## 1     3 11.51294     0       75
## 2     4 11.51294     1      100
## 3     3 11.60824     0       75
## 4     3 11.69526     0       75
## 5     3 12.57764     0       75
## 6     3 12.64433     0       75
```

Model. The outcome variable is `Level` taking $\{1, 2, 3, 4\}$. Consider the following regression

[1] https://www.census.gov/programs-surveys/ahs.html. This is a repeated cross-section survey. We use the data at one time.

equation of a latent variable:

$$y_i^* = \mathbf{x}_i \beta + u_i,$$

where $\mathbf{x}_i = (lnPrice, Top25)$ and u_i is an error term. The relationship between the latent variable y_i^* and the observed outcome variable is

$$Level = \begin{cases} 1 & \text{if} \quad -\infty < y_i^* \leq a_1 \\ 2 & \text{if} \quad a_1 < y_i^* \leq a_2 \\ 3 & \text{if} \quad a_2 < y_i^* \leq a_3 \\ 4 & \text{if} \quad a_3 < y_i^* < +\infty \end{cases}.$$

Consider the probability of realization of y_i, that is,

$$\begin{aligned} \mathbb{P}(y_i = k | \mathbf{x}_i) &= \mathbb{P}(a_{k-1} - \mathbf{x}_i \beta < u_i \leq a_k - \mathbf{x}_i \beta | \mathbf{x}_i) \\ &= G(a_k - \mathbf{x}_i \beta) - G(a_{k-1} - \mathbf{x}_i \beta), \end{aligned}$$

where $a_4 = +\infty$ and $a_0 = -\infty$. Then, the likelihood function is defined by

$$p((y_i | \mathbf{x}_i), i = 1, \dots, n; \beta, a_1, \dots, a_3) = \prod_{i=1}^{n} \prod_{k=1}^{4} (G(a_k - \mathbf{x}_i \beta) - G(a_{k-1} - \mathbf{x}_i \beta))^{I_{ik}}.$$

where I_{ik} is a indicator variable taking 1 if $y_i = k$. Finally, the log-likelihood function is

$$M(\beta, a_1, a_2, a_3) = \sum_{i=1}^{n} \sum_{k=1}^{4} I_{ik} \log(G(a_k - \mathbf{x}_i \beta) - G(a_{k-1} - \mathbf{x}_i \beta)).$$

Usually, $G(a)$ assumes the standard normal distribution, $\Phi(a)$, or the logistic distribution, $1/(1 + \exp(-a))$.

In `R`, the library (package) `MASS` provides the `polr` function which estimates the ordered probit and logit model. Although we can use the `nlm` function when we define the log-likelihood function, we do not report this method. To compare results, we use the variable `Levelnum` as outcome variable, and apply the linear regression model.

```
library(MASS)
library(tidyverse) #use case_when()

ols <- lm(Levelnum ~ lnPrice + Top25, data = house)

model <- factor(Level) ~ lnPrice + Top25
oprobit <- polr(model, data = house, method = "probit")
ologit <- polr(model, data = house, method = "logistic")

a_oprobit <- round(oprobit$zeta, 3)
a_ologit <- round(ologit$zeta, 3)
```

```r
xb_oprobit <- oprobit$lp
xb_ologit <- ologit$lp

hatY_oprobit <- case_when(
  xb_oprobit <= oprobit$zeta[1] ~ 1,
  xb_oprobit <= oprobit$zeta[2] ~ 2,
  xb_oprobit <= oprobit$zeta[3] ~ 3,
  TRUE ~ 4
)
hatY_ologit <- case_when(
  xb_ologit <= ologit$zeta[1] ~ 1,
  xb_ologit <= ologit$zeta[2] ~ 2,
  xb_ologit <= ologit$zeta[3] ~ 3,
  TRUE ~ 4
)

pred_oprobit <- round(sum(house$Level == hatY_oprobit)/nrow(house), 3)
pred_ologit <- round(sum(house$Level == hatY_ologit)/nrow(house), 3)
```

1.1 Interepretations

Table 1 shows results. OLS model shows that respondents whose household income is in the top 25 percentile live in 3.7% higher floor than other respondents. This implies that high earners want to live in higher floor, which may serve as a status goods. The ordered probit and logit model are in line with this result. To evaluate two models quantitatively, consider the following equation.

$$E[Levelnum|\mathbf{x}_i] = 25\mathbb{P}[level = 1|\mathbf{x}_i] + 50\mathbb{P}[level = 2|\mathbf{x}_i] + 75\mathbb{P}[level = 3|\mathbf{x}_i] + 100\mathbb{P}[level = 4|\mathbf{x}_i].$$

We compute this equation with $Top25 = 1$ and $Top25 = 0$ at mean value of $lnPrice$ and take difference.

```r
quantef <- function(model) {
  b <- coef(model)
  val1 <- mean(house$lnPrice)*b[1] + b[2]
  val0 <- mean(house$lnPrice)*b[1]

  prob <- matrix(c(rep(val1, 3), rep(val0, 3)), ncol = 2, nrow = 3)
  for (i in 1:3) {
    for (j in 1:2) {
      prob[i,j] <- pnorm(model$zeta[i] - prob[i,j])
    }
  }
  Ey1 <- 25*prob[1,1] + 50*(prob[2,1]-prob[1,1]) +
    75*(prob[3,1]-prob[2,1]) + 100*(1-prob[3,1])
```

```r
  Ey0 <- 25*prob[1,2] + 50*(prob[2,2]-prob[1,2]) +
    75*(prob[3,2]-prob[2,2]) + 100*(1-prob[3,2])

  return(Ey1 - Ey0)
}

ef_oprobit <- round(quantef(oprobit), 3)
ef_ologit <- round(quantef(ologit), 3)
```

As a result, we obtain similar values to OLSE. In the ordered probit model, earners in the top 25 percentile live in 4.2% higher floor than others. In the ordered logit model, earners in the top 25 percentile live in 5.9% higher floor than others. Note that, in this application, model fitness seems to be bad because the percent correctly predicted is low (16.7%).

```r
library(stargazer)
stargazer(
  ols, oprobit, ologit,
  report = "vcstp", keep.stat = c("n"),
  omit = c("Constant"),
  add.lines = list(
    c("Cutoff value at 1|2", "", a_oprobit[1], a_ologit[1]),
    c("Cutoff value at 2|3", "", a_oprobit[2], a_ologit[2]),
    c("Cutoff value at 3|4", "", a_oprobit[3], a_ologit[3]),
    c("Quantitative Effect of Top25", "", ef_oprobit, ef_ologit),
    c("Percent correctly predicted", "", pred_oprobit, pred_ologit)
  ),
  omit.table.layout = "n", table.placement = "t",
  title = "Floor Level of House: Ordered Probit and Logit Model",
  label = "housing",
  type = "latex", header = FALSE
)
```

2 Empirical Application of Multinomial Model: Gender Discremination in Job Position

Brief Background. Recently, many developed countries move toward women's social advancement, for example, an increase of number of board member. In this application, we explore whether the U.S. bank hindered the entrance of female into the workhorse.

Data. We use a built-in dataset called `BankWages` in the library `AER`. This dataset contains choice of three job position: `custodial`, `admin` and `manage`. The rank of position is `custodial < admin < manage`. Other variables are `education`, `gender`, and `minority`. We use former two variables as explanatory variables.

Table 1: Floor Level of House: Ordered Probit and Logit Model

	Dependent variable:		
	Levelnum	Level	
	OLS	*ordered probit*	*ordered logistic*
	(1)	(2)	(3)
lnPrice	0.348	0.012	0.019
	(0.430)	(0.016)	(0.026)
	t = 0.810	t = 0.777	t = 0.745
	p = 0.418	p = 0.438	p = 0.457
Top25	3.714	0.156	0.239
	(1.723)	(0.064)	(0.106)
	t = 2.156	t = 2.426	t = 2.259
	p = 0.032	p = 0.016	p = 0.024
Cutoff value at 1|2		-0.149	-0.25
Cutoff value at 2|3		0.246	0.384
Cutoff value at 3|4		0.97	1.574
Quantitative Effect of Top25		4.17	5.488
Percent correctly predicted		0.167	0.167
Observations	1,612	1,612	1,612

```r
library(AER)
data(BankWages)
dt <- BankWages
dt$job <- as.character(dt$job)
dt$job <- factor(dt$job, levels = c("admin", "custodial", "manage"))
head(BankWages, 5)
```

```
##       job education gender minority
## 1 manage        15   male       no
## 2  admin        16   male       no
## 3  admin        12 female       no
## 4  admin         8 female       no
## 5  admin        15   male       no
```

Model. The outcome variable y_i takes three values $\{0, 1, 2\}$. Then, the multinomial logit model has the following response probabilities

$$P_{ij} = \mathbb{P}(y_i = j | \mathbf{x}_i) = \begin{cases} \frac{\exp(\mathbf{x}_i \beta_j)}{1+\sum_{k=1}^2 \exp(\mathbf{x}_i \beta_k)} & \text{if} \quad j = 1, 2 \\ \frac{1}{1+\sum_{k=1}^2 \exp(\mathbf{x}_i \beta_k)} & \text{if} \quad j = 0 \end{cases}.$$

The log-likelihood function is

$$M_n(\beta_1, \beta_2) = \sum_{i=1}^n \sum_{j=0}^3 d_{ij} \log(P_{ij}),$$

where d_{ij} is a dummy variable taking 1 if $y_i = j$.

In `R`, some packages provide the multinomial logit model. In this application, we use the `multinom` function in the library `nnet`.

```
library(nnet)
est_mlogit <- multinom(job ~ education + gender, data = dt)

# observations and percent correctly predicted
pred <- est_mlogit$fitted.value
pred <- colnames(pred)[apply(pred, 1, which.max)]
n <- length(pred)
pcp <- round(sum(pred == dt$job)/n, 3)

# Log-likelihood and pseudo R-sq
loglik1 <- as.numeric(nnet:::logLik.multinom(est_mlogit))
est_mlogit0 <- multinom(job ~ 1, data = dt)
loglik0 <- as.numeric(nnet:::logLik.multinom(est_mlogit0))
pr2 <- round(1 - loglik1/loglik0, 3)
```

2.1 Interpretations

Table 2 summarizes the result of multinomial logit model. The coefficient represents the change of $\log(P_{ij}/P_{i0})$ in corresponding covariate. For example, eduction decreases the log-odds between `custodial` and `admin`, $\log(P_{i,custodial}/P_{i,admin})$ by -0.562. This implies that those who received higher education are more likely to obtain the position `admin`. Highly-educated workers are also more likely to obtain the position `manage`. Moreover, a female dummy decrease the log-odds between `manage` and `admin` by -0.748, which implies that females are less likely to obtain higher position `manage`. From this result, we conclude that the U.S. bank disencouraged females to assign higher job position.

Finally, we should check the model fitness. The predicted position is the outcome with the highest estimated probability. The multinomial logit model correctly predicts many cases (correction rate: 85.2%).

Table 2: Multinomial Logit Model of Job Position

	Dependent variable:	
	custodial	manage
	(1)	(2)
Education	−0.562	1.661
	(0.098)	(0.247)
	t = −5.721	t = 6.715
	p = 0.000	p = 0.000
Female = 1	−10.976	−0.748
	(27.808)	(0.429)
	t = −0.395	t = −1.743
	p = 0.694	p = 0.082
Constant	5.030	−26.730
	(1.130)	(3.874)
	t = 4.450	t = −6.899
	p = 0.00001	p = 0.000
Observations	474	
Percent correctly predicted	0.852	
Log-likelihood	-144.928	
Pseudo R-sq	0.546	

```r
stargazer(
  est_mlogit,
  covariate.labels = c("Education", "Female = 1"),
  report = "vcstp", omit.stat = c("aic"),
  add.lines = list(
    c("Observations", n, ""),
    c("Percent correctly predicted", pcp, ""),
    c("Log-likelihood", round(loglik1, 3), ""),
    c("Pseudo R-sq", pr2, "")
  ),
  omit.table.layout = "n", table.placement = "t",
  title = "Multinomial Logit Model of Job Position",
  label = "job",
  type = "latex", header = FALSE
)
```