Econometrics Homework 2

Utpalraj Kemprai
MDS202352

Question 1

We have the following ordinal regression model:

$$z_i = x_i'\beta + \epsilon_i \quad \forall i = 1, \cdots, n$$
$$\gamma_{j-1} < z_i \leq \gamma_j \implies y_i = j, \quad \forall i, j = 1, \cdots, J$$

where (in the first equation) z_i is the latent variable for individual i, x_i is a vector of covariates, β is a $k \times 1$ vector of unknown parameters, and n denotes the number of observations. The second equation shows how z_i is related to the observed discrete response y_i, where $-\infty = \gamma_0 < \gamma_1 < \gamma_{J-1} < \gamma_J = \infty$ are the cut-points (or thresholds) and y_i is assumed to have J categories or outcomes.

(a)

Probability of success

We assume that $\epsilon_i \sim N(0,1)$, for $i = 1, 2, \cdots, n$. Therefore we have,
The probability of success,

$$\begin{aligned} Pr(y_i = j) &= Pr(\gamma_{j-1} < z_i \leq \gamma_j) \\ &= Pr(\gamma_{j-1} < x_i'\beta + \epsilon_i \leq \gamma_j) \\ &= Pr(\gamma_{j-1} - x_i'\beta < \epsilon_i \leq \gamma_j - x_i'\beta) \\ &= \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta) \qquad \text{[where } \Phi(\cdot) \text{ is the cdf of } N(0,1)\text{]} \end{aligned}$$

Likelihood function

The likelihood function for the ordinal probit model is,

$$\begin{aligned} L(\beta, \gamma; y) &= \prod_{i=1}^{n}\prod_{j=1}^{J} Pr(y_i = j|\beta, \gamma)^{I(y_i=j)} \qquad \text{[where } I(\cdot) \text{ is the indicator function]} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J} \Big(\Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta)\Big)^{I(y_i=j)} \end{aligned}$$

(b)

Probability of success

We assume that $\epsilon_i \sim \mathcal{L}(0,1)$, for $i = 1, 2, \cdots, n$. Therefore the probability of success,

$$\begin{aligned} Pr(y_i = j) &= Pr(\gamma_{j-1} < z_i \leq \gamma_j) \\ &= Pr(\gamma_{j-1} < x_i'\beta + \epsilon_i \leq \gamma_j) \\ &= Pr(\gamma_{j-1} - x_i'\beta < \epsilon_i \leq \gamma_j - x_i'\beta) \\ &= \frac{1}{1 + e^{-(\gamma_j - x_i'\beta)}} - \frac{1}{1 + e^{-(\gamma_{j-1} - x_i'\beta)}} \qquad \text{[as } \epsilon_i \sim \mathcal{L}(0,1)\text{]} \\ &= \frac{e^{-(\gamma_{j-1} - x_i'\beta)} - e^{-(\gamma_j - x_i'\beta)}}{(1 + e^{-(\gamma_j - x_i'\beta)})(1 + e^{-(\gamma_{j-1} - x_i'\beta)})} \\ &= \frac{e^{x_i'\beta}(e^{-\gamma_{j-1}} - e^{-\gamma_j})}{(1 + e^{-(\gamma_j - x_i'\beta)})(1 + e^{-(\gamma_{j-1} - x_i'\beta)})} \end{aligned}$$

Likelihood function

The likelihood function for the ordinal logit model is,

$$L(\beta,\gamma;y) = \prod_{i=1}^{n}\prod_{j=1}^{J} Pr(y_i = j|\beta,\gamma)^{I(y_i=j)} \qquad \text{[where } I(\cdot) \text{ is the indicator function]}$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{J}\left(\frac{e^{x_i'\beta}(e^{-\gamma_{j-1}} - e^{-\gamma_j})}{(1+e^{-(\gamma_j - x_i'\beta)})(1+e^{-(\gamma_{j-1}-x_i'\beta)})}\right)^{I(y_i=j)}$$

(c)

Probability remain unchanged on adding a constant c to cut-points and the mean

If we add a constant c to the cut-points γ_j and the means $x_i'\beta$, then $\forall j = 1, 2, \cdots, J-1$ and $\forall i = 1, 2, \cdots, n$, the latent variable z_i becomes $x_i'\beta + c + \epsilon_i$ and the cut-point γ_j becomes $\gamma_j + c$.

The probability of y_i taking the value j is,

$$\begin{aligned} Pr(y_i = j) &= Pr(\gamma_{j-1} + c < z_i \leq \gamma_j + c) \\ &= Pr(\gamma_{j-1} + c < x_i'\beta + c + \epsilon_i \leq \gamma_j + c) \\ &= Pr(\gamma_{j-1} - x_i'\beta < \epsilon_i \leq \gamma_j - x_i'\beta) \\ &= \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta) \end{aligned}$$

which is the same as the value obtained in part(a) of Question 1. So adding a constant c to the cut-point γ_j and the mean $x_i'\beta$ does not change the outcome probability.

Identification Problem

This identification problem can be solved by fixing the value of one of $\gamma_1, \gamma_2, \cdots, \gamma_{J-1}$. In particular setting $\gamma_1 = 0$ will solve this identification problem.

(d)

Rescaling the parameters (γ_j, β) and scale of distribution does not change outcome probability

Rescaling the parameters (γ_j, β) and the scale of the distribution of ϵ_i by some constant $d > 0$, the latent variable z_i becomes $x_i'd\beta + \epsilon_i$ where $\epsilon_i \sim N(0, d^2)$, $\forall i = 1, 2, \cdots, n$ and the cut-point γ_j becomes $d\gamma_j, \forall j = 1, 2, \cdots, J-1$.

The probability of y_i taking the value j is,

$$\begin{aligned} Pr(y_i = j) &= Pr(d\gamma_{j-1} < z_i \leq d\gamma_j) \\ &= Pr(d\gamma_{j-1} < x_i'd\beta + \epsilon_i \leq d\gamma_j) \\ &= Pr(\gamma_{j-1} - x_i'\beta < \frac{\epsilon_i}{d} \leq \gamma_j - x_i'\beta) \\ &= \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta) \qquad [\text{as } \epsilon_i \sim N(0,d^2), \frac{\epsilon_i}{d} \sim N(0,1)] \end{aligned}$$

which is the same as the value obtained in part(a) of Question 1. So rescaling the parameters (γ_j, β) and the scale of the distribution by some arbitrary constant d lead to same outcome probabilities.

Identification problem

This identification problem can be solved by fixing the scale of the distribution of ϵ_i. In particular we can set the scale of the distribution of ϵ_i to 1, i.e. $\text{var}(\epsilon_i) = 1$.

(e)

(i) Descriptive Summary of the data

VARIABLE		MEAN	STD
LOG AGE		3.72	0.44
LOG INCOME		10.63	0.98
HOUSEHOLD SIZE		2.74	1.42
	CATEGORY	COUNTS	PERCENTAGE
PAST USE		719	48.19
MALE		791	53.02
EDUCATION	BACHELORS & ABOVE	551	36.93
	BELOW BACHELORS	434	29.09
	HIGH SCHOOL & BELOW	507	33.98
TOLERANT STATES		556	37.27
EVENTUALLY LEGAL		1154	77.35
RACE	WHITE	1149	77.01
	AFRICAN AMERICAN	202	13.54
	OTHER RACES	141	9.45
PARTY AFFILIATION	REPUBLICAN	333	22.32
	DEMOCRAT	511	34.25
	INDEPENDENT & OTHERS	648	43.43
RELIGION	PROTESTANT	550	36.86
	ROMAN CATHOLIC	290	19.44
	CHRISTIAN	182	12.20
	CONSERVATIVE	122	8.18
	LIBERAL	348	23.32
PUBLIC OPINION	OPPOSE LEGALIZATION	218	14.61
	LEGAL ONLY FOR MEDICINAL USE	640	42.90
	LEGAL FOR PERSONAL USE	634	42.49

Table 1: Descriptive Summary of the variables

(ii) Public opinion on extent of marijuana legalization

We estimate Model 8 and replicate the results from the lecture.

coef	Estimate	Std. Error	t value	p value
intercept	0.35	0.48	0.72	0.47
log_age	-0.35	0.08	-4.51	< 0.05
log_income	0.09	0.04	2.47	< 0.05
hh1(household size)	-0.02	0.02	-0.87	0.38
pastuse	0.69	0.06	10.67	< 0.05
bachelor_above	0.24	0.08	3.01	< 0.05
below_bachelor	0.05	0.08	0.62	0.54
tolerant_state	0.07	0.07	1.04	0.30
expected_legal	0.57	0.07	7.76	< 0.05
black	0.03	0.10	0.26	0.79
other_race	-0.28	0.11	-2.56	< 0.05
democrat	0.44	0.09	5.03	< 0.05
other_party	0.36	0.08	4.56	< 0.05
male	0.06	0.06	1.00	0.32
christian	0.16	0.10	1.60	0.11
roman_catholic	0.10	0.09	1.19	0.23
liberal	0.39	0.09	4.39	< 0.05
conservative	0.09	0.12	0.76	0.45
cut_point(γ_1)	1.46	0.05	29.58	< 0.05
LR (χ^2) Statistic	377			
McFadden's R^2	0.13			
Hit-Rate	58.91			

Table 2: Model 8: Estimation Results

(iii) Covariate effects of the variables

Covariate	Δ P(not legal)	Δ P(medicinal use)	Δ P(personal use)
Age, 10 years	0.015	0.012	-0.028
Income, $ 10,000	-0.005	-0.003	0.008
Past Use	-0.129	-0.113	0.243
Bachelors & Above	-0.045	-0.035	0.080
Eventually Legal	-0.126	-0.060	0.186
Other Races	0.059	0.031	-0.089
Democrat	-0.080	-0.066	0.147
Other parties	-0.070	-0.051	0.121
Liberal	-0.068	-0.066	0.134

Table 3: Average covariate effects from Model 8.

Question 2

Grunfeld Investment Study

Investment is modeled as a function of market value, capital, and firm. There are 200 observation on 10 firms from 1935-1954 (20 years). We have excluded the data on the company American Steel.

Variable Description

- invest: Gross investment in 1947 dollars.
- value: Market value as of Dec. 31 in 1947 dollars.
- capital: Stock of plant and equipment in 1947 dollars.
- firm: General Motors, US Steel, General Electric, Chrysler, Atlantic Refining, IBM, Union Oil, Westinghouse, Goodyear, Diamond Match.

(a)

Pooled Effects model

	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
intercept	-42.71	9.511	-4.49	< 0.001
capital	0.23	0.025	9.05	< 0.001
value	0.12	0.006	19.80	< 0.001

Table 4: Covariates of the Pooled Effects model

- R-Squared: 0.812
- Adj. R-Squared: 0.811
- F-Statistic: 426.58 with 2 and 197 df, p-value $< 2.22 \times 10^{-16}$

Interpretation of Coefficients

- capital: It's coefficient is statistically significant(p-value < 0.001) and has a positive value (0.23). So investment increases with increase in capital as per our Pooled Effects model.
- value: It's coefficient is also statistically significant(p-value < 0.001) and is positive (0.12). So a increase in value also increases the investment, but relatively lesser compared to capital.

The model is able to explain about 81.2% variance in the data and the high value of the F-Statistic also shows that the variables are statistically significant.

(b)

Fixed Effects model

	Estimate	Std. Error	t-value	Pr(> |t|)
capital	0.31	0.017	17.87	< 0.001
value	0.11	0.012	9.29	< 0.001

Table 5: Covariates of the Fixed Effects model

	Estimate	Std. Error	t-value	Pr(> |t|)
Atlantic Refining	-114.62	14.17	-8.09	< 0.001
Chrysler	-27.81	14.08	-1.98	0.05
Diamond Match	-6.57	11.83	-0.56	0.58
General Electric	-235.57	24.43	-9.64	< 0.001
General Motors	-70.30	49.71	-1.41	0.16
Goodyear	-87.22	12.89	-6.77	< 0.001
IBM	-23.16	12.67	-1.83	0.07
Union Oil	-66.55	12.84	-5.18	< 0.001
US Steel	101.91	24.94	4.09	< 0.001
Westinghouse	-57.55	13.99	-4.11	< 0.001

Table 6: Firm specific intercepts

- R-Squared: 0.767
- Adj. R-Squared: 0.753
- F-Statistic: 309.01 with 2 and 188 df, p-value $< 2.22 \times 10^{-16}$

Test for Individual Effects

```
#Test for Individual Effects
#------------------------------------------------------------------------------
pFtest(fe_model_plm, pooled_effect_plm)

# OUTPUT:
# F test for individual effects
#
# data: invest ~ capital + value
# F = 49.177, df1 = 9, df2 = 188, p-value < 2.2e-16
# alternative hypothesis: significant effects
```

Listing 1: R Code for Test for Individual Effects

As the p-value is very small($< 2.2 \times 10^{-16}$), the null hypothesis is rejected in favor of the alternative that there are significant fixed effects. Hence, a Fixed Effect model will be a better choice than a Pooled Effect model, in this case.

Firm–Specific Intercepts

The firm-specific intercepts reveal interesting patterns in baseline investment behavior across companies. US Steel stands out with a significantly positive intercept (101.91), suggesting this firm maintains higher baseline investment levels compared to others, even after accounting for capital and value effects. In contrast, General Electric shows the most negative intercept (-235.57), indicating substantially lower baseline investment. The statistical significance of most intercepts (6 out of 10 firms showing p-values < 0.001) reinforces the importance of controlling for firm-specific effects when modeling investment behavior.

Interpretation of Coefficients

- capital: It's coefficient is statistically significant(p-value < 0.001) and has a positive value (0.31). So investment increases with increase in capital as per our Fixed Effects model.
- value: It's coefficient is also statistically significant(p-value < 0.001) and is positive (0.11). So a increase in value also increases the investment, but relatively lesser compared to capital.

(c)

Random Effects model

	var	std.dev	share
idiosyncratic	2784.46	52.77	0.282
individual	7089.80	84.20	0.718
$\theta = 0.861$			

Table 7: Variance Components of the Random Effects Model

	Estimate	Std. Error	z-value	Pr($>\|z\|$)
intercept	-57.83	28.90	-2.00	0.045
capital	0.31	0.017	17.93	< 0.001
value	0.11	0.010	10.46	< 0.001

Table 8: Covariates of the Random Effects model

- R-Squared: 0.770
- Adj. R-Squared: 0.767
- χ^2 Statistic: 657.67 on 2 df, p-value $< 2.22 \times 10^{-16}$

Hausman Test

```
# Hausman Test
#-------------------------------------------------------------------------
phtest(fe_model_plm,re_model_plm)

# OUTPUT:
# Hausman Test
#
# data: invest ~ capital + value
# chisq = 2.3304, df = 2, p-value = 0.3119
# alternative hypothesis: one model is inconsistent
```

Listing 2: R Code for Hausman Test

The null hypothesis cannot be rejected here (p-value = 0.3119). Hence it makes sense to use a Random Effects model instead of a Fixed Effects model.

Idiosyncratic and Individual Variances

- Idiosyncratic variance: Idiosyncratic variance refers to the within-firm variation. This captures how a firm's values change over time, representing the deviation from that firm's own average. This is essentially the variance of the error term after accounting for the firm-specific effects.
- Individual variance: Individual variance refers to the between-firm variation. This captures how different firms vary from each other in their average level of the dependent variable (investment in your case). It represents the variation in the random intercepts across different firms.

In our Random Effects model results, the individual variance (7089.80) being much larger than the idiosyncratic variance (2784.46) tells us that differences between firms explain more of the variation in investment than changes within firms over time. The share values confirm this: 71.8% of the total variance comes from between-firm differences, while only 28.2% comes from within-firm variation over time.

The high theta value ($\theta = 0.861$) further indicates the relative importance of the variation between firms in the total variation.

Interpretation of Coefficients

- capital: It's coefficient is statistically significant (p-value < 0.001) and is positive (0.31). So increase in capital increases the investment as per our Random Effects model.
- value: It's coefficient is also statistically significant (p-value < 0.001 and is positive (0.11). So increase in value also increases the investment but has a smaller effect than capital.