Econometrics Homework 1

Utpalraj Kemprai
MDS202352

January 25, 2025

Question 1

(a)

We have the binary logit model expressed in terms of a continuous latent random variable z_i as

$$z_i = x_i'\beta + \epsilon_i, \forall i = 1, \cdots, n$$

$$y_i = \begin{cases} 1 & \text{if } z_i > 0, \\ 0 & \text{otherwise} \end{cases}$$

where $\epsilon_i \sim L(0, \frac{\pi^2}{3})$, here L denotes a logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$ and variance

Therefore the probability of success

$$\begin{aligned} Pr(y_i = 1) &= Pr(z_i > 0) = Pr(x_i'\beta + \epsilon_i > 0) \\ &= Pr(\epsilon_i > -x_i'\beta) \\ &= 1 - Pr(\epsilon_i \leq -x_i'\beta) \\ &= 1 - \frac{1}{1 + e^{-(-x_i'\beta)}} && \text{[using the fact } \epsilon_i \sim L(0, \frac{\pi^2}{3}) \text{]} \\ &= \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \end{aligned}$$

(b)

The likelihood for a binary model expressed as a function of β is

$$\ell(\beta; y) = \prod_{i=1}^{n} (Pr(y_i = 0)I(y_i = 0) + Pr(y_i = 1)I(y_i = 1))$$

where $I(.)$ is an indicator function.
From part (a) of question 1, we have

$$Pr(y_i = 1) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

and

$$Pr(y_i = 0) = 1 - Pr(y_i = 1) = \frac{1}{1 + e^{x_i'\beta}}$$

So the likelihood function of the Logit model is,

$$\ell(\beta; y) = \prod_{i=1}^{n} (\frac{1}{1 + e^{x_i'\beta}} I(y_i = 0) + \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} I(y_i = 1))$$

(c)

Here the dependent variable is the probability that the subject dies before age 65, and the primary explanatory variable of interest is whether the person smoked (at all) in the years prior to age 65.

And the latent equation is,

$$z_i = \beta_1 + \text{Smoke}_{2i}\beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k + \epsilon_i$$

where Smoke_{2i} is an indicator for smoking status

The odds of mortality by age 65 for an individual i are:

Case 1: individual i was a smoker ($\text{Smoke}_{2i} = 1$)

$$\begin{aligned}
\text{Odds}_{\text{Smoker}} &= \frac{Pr(y_i = 1|\text{Smoke}_{2i} = 1)}{1 - Pr(y_i = 1|\text{Smoke}_{2i} = 1)} \\
&= \frac{Pr(z_i > 0|\text{Smoke}_{2i} = 1)}{1 - Pr(z_i > 0|\text{Smoke}_{2i} = 1)} \\
&= \frac{Pr(\epsilon_i > -(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))}{1 - Pr(\epsilon_i > -(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))} \\
&= \frac{1 - Pr(\epsilon_i \leq -(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))}{Pr(\epsilon_i \leq -(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))} \\
&= \frac{1 - \frac{1}{1+e^{(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}}{\frac{1}{1+e^{(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}} \qquad \text{[using the fact } \epsilon_i \sim L(0, \frac{\pi^2}{3}) \text{]} \\
&= e^{(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}
\end{aligned}$$

Case 2: individual i was a nonsmoker ($\text{Smoke}_{2i} = 0$)

$$\begin{aligned}
\text{Odds}_{\text{NonSmoker}} &= \frac{Pr(y_i = 1|\text{Smoke}_{2i} = 0)}{1 - Pr(y_i = 1|\text{Smoke}_{2i} = 0)} \\
&= \frac{Pr(z_i > 0|\text{Smoke}_{2i} = 0)}{1 - Pr(z_i > 0|\text{Smoke}_{2i} = 0)} \\
&= \frac{Pr(\epsilon_i > -(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))}{1 - Pr(\epsilon_i > -(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))} \\
&= \frac{1 - Pr(\epsilon_i \leq -(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))}{Pr(\epsilon_i \leq -(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k))} \\
&= \frac{1 - \frac{1}{1+e^{(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}}{\frac{1}{1+e^{(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}} \qquad \text{[using the fact } \epsilon_i \sim L(0, \frac{\pi^2}{3}) \text{]} \\
&= e^{(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}
\end{aligned}$$

The log-odds ratio of mortality for a smoker vs nonsmoker is:

$$\begin{aligned}
\text{log-odds ratio} &= \ln(\frac{\text{Odds}_{\text{Smoker}}}{\text{Odds}_{\text{NonSmoker}}}) \\
&= \ln(\frac{e^{(\beta_1 + \beta_{\text{Smoke}} + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}{e^{(\beta_1 + x_{3i}\beta_2 + \cdots + x_{ki}\beta_k)}}) \\
&= \ln(e^{\beta_{\text{Smoke}}}) \\
&= \beta_{\text{Smoke}}
\end{aligned}$$

Question 2

(a)

The variables in the model are as follows:

- Discrete : cars, intercept, depend
- Continuous : dcost, dovtt, divtt

Descriptive summary of continuous variables

Variable	Mean	Standard Deviation
dcost	-12.94	37.97
dovtt	12.85	10.06
divtt	17.05	17.96

Table 1: Descriptive summary of continuous variables

Descriptive summary of discrete variables

cars	Count	Percentage (%)
0	81	9.620
1	359	42.637
2	322	38.242
3	64	7.601
4	12	1.425
5	3	0.356
7	1	0.119

Table 2: cars Count and Percentage

depend	Count	Percentage (%)
0	135	16.03
1	707	83.97

Table 3: depend Count and Percentage

(b)

We now estimate the Probit and Logit models by regressing the dependent variable depend on intercept, dcost, cars, dovtt and divtt.

Regression Coefficients: Estimate and Standard errors

	Estimate	**Standard Error**
(Intercept)	-1.222	0.304
DCOST	0.017	0.004
CARS	2.308	0.226
DOVTT	0.062	0.019
DIVTT	0.009	0.009

Table 4: Regression Coefficients for Logit Model

	Estimate	**Standard Error**
(Intercept)	-0.601	0.166
DCOST	0.010	0.002
CARS	1.225	0.114
DOVTT	0.032	0.010
DIVTT	0.005	0.005

Table 5: Regression Coefficients for Probit Model

Interpretation of the coefficient of cars

Model	**Estimate**	**Std. Error**	**z value**	**Pr($>|z|$)**
Probit	1.225	0.114	10.75	$< 2e^{-16}$
Logit	2.308	0.226	10.21	$< 2e^{-16}$

Table 6: Coefficient of cars for Probit and Logit Model

- **Probit:**

  The p-value for the coefficient of cars is less than 2^{-16}, so the coefficient is significant. The coefficient is positive with a value of 1.225, so people with more cars are more likely to choose automobile than transit. More precisely, a unit increase in the number of cars owned will result in an increase of 1.225 in the log-odds.

- **Logit:**

  The p-value for the coefficient of cars for the Logit model is also less than 2^{-16}, so it is also significant. The coefficient is positive with a value of 2.308. So in the Logit model, cars have a greater impact for choosing automobile than the Probit model. As was the case with Probit model, people with more cars more likely to choose automobile than transit. A unit increase in the number of cars owned will result in an increase of 2.308 in the log-odds.

(c)

Model	**AIC**	**BIC**	**log-likelihood**	**Hit-rate**
Probit	470.33	494.00	-230.16	0.9038
Logit	465.74	489.42	-227.87	0.9038

Table 7: AIC,BIC,log-likelihood and Hit-rate for the models

Question 3

(a)

Descriptive summary of variables

	Mean	Standard Deviation
WHRS	740.58	871.31
WomenEduc	12.29	2.28
WomenExp	10.63	8.07
WomenAge	42.54	8.07
childl6	0.24	0.52

Table 8: Descriptive summary of variables of interest

(b)

Linear Regression on only positive values of WHRS

	Estimate	Std. Error	t-value
(Intercept)	1829.75	292.54	6.25
WomenEduc	-16.46	15.58	-1.06
WomenExp	33.94	5.01	6.77
WomenAge	-17.11	5.46	-3.13
childl6	-305.31	96.45	-3.17

Table 9: Linear Regression model on only positive values of WHRS

Linear Regression framework is not appropriate

The linear regression model is not suitable for work hours because work hours data are censored (many people work zero hours), which violates the assumptions of linear regression. Linear regression assumes continuous, unbounded data, while work hours are bounded (cannot be negative) and may exhibit zero-inflation (many people working zero hours). The Tobit model is specifically designed for censored data and can handle these issues, making it a better choice for modeling work hours.

(c)

Tobit model and corresponding likelihood

The Tobit model is:

$$\begin{aligned} z_i &= \beta_1 + \text{WomenEduc}_i\beta_2 + \text{WomenExp}_i\beta_3 + \text{WomenAge}_i\beta_4 + \text{childl6}_i\beta_5 + \epsilon_i, \forall i = 1, \cdots, 753 \\ &= x_i'\beta + \epsilon_i, \forall i = 1, \cdots, 753 \end{aligned}$$

$$y_i = \text{WHRS}_i = \begin{cases} z_i & \text{if } z_i > 0, \\ 0 & \text{otherwise} \end{cases}$$

where,
$\epsilon_i \sim N(0, \sigma^2)$, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$, $x_i = (1, \text{WomenEduc}_i, \text{WomenExp}_i, \text{WomenAge}_i, \text{childl6}_i)'$

The corresponding likelihood function for the Tobit Model is:

$$\begin{aligned}
\ell(\beta; y) &= \prod_{i:y_i=0} Pr(y_i = 0) \prod_{i:y_i>0} \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}) \\
&= \prod_{i:y_i=0} Pr(z_i \leq 0) \prod_{i:y_i>0} \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}) \\
&= \prod_{i:y_i=0} Pr(\frac{\epsilon_i}{\sigma} \leq -\frac{x_i'\beta}{\sigma}) \prod_{i:y_i>0} \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}) \\
\Rightarrow \ell(\beta; y) &= \prod_{i:y_i=0} \Phi(\frac{-x_i'\beta}{\sigma}) \prod_{i:y_i>0} \frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}) \qquad [\text{as } \epsilon_i \sim N(0, \sigma^2),\ \frac{\epsilon_i}{\sigma} \sim N(0,1)\ ]
\end{aligned}$$

(d)

	Estimate	**Std. Error**	**z-value**
(Intercept)	1349.88	386.30	3.49
WomenEduc	73.29	20.47	3.58
WomenExp	80.54	6.29	12.81
WomenAge	-60.77	6.89	-8.82
childl6	-918.92	111.66	-8.23

Table 10: Tobit model summary

Effect each variable on the response variable

- WomenEduc:

  The positive coefficient suggests that an additional year of education for women is associated with an increase of approximately 73.29 units in the latent work hours variable. The z-value (3.58) indicates this effect is statistically significant.

- WomenAge:

  The negative coefficient indicates that as women age, their latent work hours decrease by about 60.77 units per year in the latent work hours variable. This effect is statistically significant (z-value = -8.82), suggesting a strong relationship between age and work hours.

- WomenExp:

  The positive and significant coefficient implies that each additional year of work experience for women increases their latent work hours by about 80.54 units. The very high z-value (12.81) highlights strong statistical significance.

- childl6:

  The large negative coefficient implies that having a child under the age of 6 is associated with a significant reduction in latent work hours, by approximately 918.92 units. This effect is highly significant (z-value = -8.23), reflecting the substantial impact of childcare responsibilities on women's work hours.

(e)

Marginal Effect of another year of education on observed hours of work

The marginal effect of year of education on observed hours of work is evaluated as:

$$\frac{\partial E(\text{WHRS}_p|x_p)}{\partial \text{WomenEduc}}\bigg| z_p > 0 = \beta_2 \Phi(\frac{\bar{x}_p'\beta}{\sigma})$$

Setting WomenEduc, WomenExp, and WomenAge at the corresponding mean values and childl6 = 1, and inputting the value of β and σ obtained from the Tobit model in the above equation we get:

$$\frac{\partial E(\text{WHRS}_p|x_p)}{\partial \text{WomenEduc}}\bigg| z_p > 0 = 26.6$$