HW1 - Econometrics

Himanshu, MDS202327

2025-01-25

```
# Import necessary libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(knitr)
library(pastecs)
library(AER)
```

Question 1

(a) Find the probability of success $\Pr(y_i = 1)$?

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(z_i > 0) \\
&= \Pr(x_i'\beta + \epsilon_i > 0) \\
&= \Pr(\epsilon_i > -x_i'\beta) \\
&= 1 - \Pr(\epsilon_i \leq -x_i'\beta) \\
&= 1 - (1 - \Pr(\epsilon_i \leq x_i'\beta)) \\
&= \Pr(\epsilon_i \leq x_i'\beta) \\
&= \Lambda(x_i'\beta) \\
&= \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}
\end{aligned}$$

(b) Derive the likelihood function of the logit model.

$$\begin{aligned}
\Pr(y_i = 0) &= \Pr(z_i \leq 0) \\
&= \Pr(x_i'\beta + \epsilon_i \leq 0) \\
&= \Pr(\epsilon_i \leq -x_i'\beta) \\
&= \Lambda(-x_i'\beta) \\
&= \frac{\exp(-x_i'\beta)}{1 + \exp(-x_i'\beta)}
\end{aligned}$$

$$\begin{aligned}
l(\beta, y) &= \prod_{i=1}^{n} [\Pr(y_i = 0)\mathrm{I}(y_i = 0) + \Pr(y_i = 1)\mathrm{I}(y_i = 1)] \\
&= \prod_{i=1}^{n} [\Lambda(-x_i'\beta)\mathrm{I}(y_i = 0) + \Lambda(x_i'\beta)\mathrm{I}(y_i = 1)]
\end{aligned}$$

(c) Consider a study in which 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. Let Smoke_{i2} be an indicator for smoking status, and $\boldsymbol{\beta_{smoke}}$ be the corresponding coecient. Then the latent regression in equation becomes:

$$z_i = \beta_1 + \text{Smoke}_{i2}\beta_{smoke} + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k$$

Find the odds of mortality by age 65 if individual $\boldsymbol{i}$ was a smoker ($\text{Smoke}_{i2} = 1$) and the odds if individual $\boldsymbol{i}$ was a nonsmoker ($\text{Smoke}_{i2} = 0$). What is the log-odds ratio of mortality for a smoker vs nonsmoker?

$p = \Pr(y_i = 1) = \frac{\exp(z_i)}{1+\exp(z_i)}$

$\frac{p}{1-p} = \exp(z_i)$

$\frac{p}{1-p} = \exp(\beta_1 + \text{Smoke}_{i2}\beta_{smoke} + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k)$

Odds for smoker is given by,

$$\exp(\beta_1 + \beta_{smoke} + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k)$$

Odds for non-smoker is given by,

$$\exp(\beta_1 + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k)$$

Log-odds ratio is given by,

$$\log\left(\frac{\exp(\beta_1 + \beta_{smoke} + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k)}{\exp(\beta_1 + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k)}\right) = \beta_{smoke}$$

Question 2

(a) Present the descriptive summary of the variables (i.e., mean and standard deviation for continuous variables and count and percentage for discrete variables) in a table.

```
data_transport = read.csv("transport.csv", header = TRUE)
summary_transport = stat.desc(data_transport)
summary_transport[c(9,8,13),]
```

Statistics	dcost	cars	dovtt	divtt
Mean	-12.941	1.502	12.854	17.052
Median	-6.50	1.00	11.00	13.00
Std. Dev	37.974	0.871	10.064	17.964

Statistics	intcpt	depend
Count	842	707
Percentage	100%	83.96%

Table 1: Descriptive summary of the variables.

(b) Estimate Probit and Logit models by regressing the dependent variable depend on intercept, dcost, cars, dovtt and divtt. Present the regression coecients and the standard errors in a table. Numbers should be reported to 3 digits after the decimal. Interpret the coefficient for cars.

```
probit_model = glm(depend ~ dcost + cars + dovtt + divtt,
                   family = binomial(link = "probit"),
                   data = data_transport)

summary(probit_model)
```

Coefficients	Estimate	Std. Error
(Intercept)	-0.601	0.165
dcost	0.009	0.002
cars	**1.225**	0.114
dovtt	0.032	0.009
divtt	0.005	0.004

Table 2: Results of the probit model.

The coefficient estimate of 1.225 can be interpreted as a unit increase in the number of cars owned by the traveler's household will result in 1.225 increase in log-odds.
That is $(\exp(1.225) - 1) * 100\% = \mathbf{240\%}$ increase in the odds assuming other variables remain fixed.

```r
logit_model = glm(depend ~ dcost + cars + dovtt + divtt,
                  family = binomial(link = "logit"),
                  data = data_transport)

summary(logit_model)
```

Coefficients	Estimate	Std. Error
(Intercept)	-1.221	0.303
dcost	0.016	0.003
cars	**2.308**	0.226
dovtt	0.062	0.018
divtt	0.009	0.009

Table 3: Results of the logit model.

The coefficient estimate of 2.308 can be interpreted as a unit increase in the number of cars owned by the traveler's household will result in 2.308 increase in log-odds.
That is $(\exp(2.308) - 1) * 100\% = \mathbf{905\%}$ increase in the odds assuming other variables remain fixed.

(c) Calculate the sum of the log-likelihood, Akaike Information Criterion, Bayesian Information Criterion and Hit-rate for the Probit and Logit models

```r
# Log-likelihood
logLik(probit_model)
logLik(logit_model)

# AIC
AIC(probit_model)
AIC(logit_model)

# BIC
BIC(probit_model)
BIC(logit_model)

# Hit Rate
mean((logit_model$fitted.values > 0.5) == data_transport$depend)
mean((probit_model$fitted.values > 0.5) == data_transport$depend)
```

Measures	Probit Model	Logit Model
Log-Likelihood	-230.16	-227.86
AIC	470.33	465.74
BIC	494.00	489.41
Hit Rate	90.38%	90.38%

Table 4: Log-likelihood, AIC, BIC, HR for the probit and logit model.

Question 3

(a) Present a descriptive summary (mean and standard deviation) of the variables of interest. Report all results to two digits after the decimal.

```
data_mroz = read.csv("mroz.csv", header = TRUE)
summary_mroz = stat.desc(data_mroz)
summary_mroz[c(9,8,13),]
```

Statistics	WomenEduc	WomenExp	WomenAge	childl6	WHRS
Mean	-12.28	10.63	42.53	0.23	740.57
Median	12.00	9.00	43.00	0.00	288.00
Std. Dev	2.28	8.06	8.07	0.52	871.31

Table 5: Table 1: Descriptive summary of the variables.

(b) Estimate a linear regression model only on positive values of WHRS and report the coefficient estimates, standard errors, and t-values in a table. Are there reasons to believe that a linear regression framework will not be appropriate for this data? Please explain.

```
pos_idx = data_mroz$WHRS>0
data_mroz_pos = data_mroz[pos_idx,]
lin_reg_pos = lm(WHRS ~ WomenEduc+WomenExp+WomenAge+childl6,
                 data = data_mroz_pos)
summary(lin_reg_pos)
```

Coefficients	Estimate	Std. Error	t-value
(Intercept)	1829.746	292.536	6.255
WomenEduc	-16.462	15.581	-1.057
WomenExp	33.936	5.009	6.775
WomenAge	-17.108	5.458	-3.135
childl6	-305.309	96.449	-3.165

Table 6: Results of the linear regression model on postive values of WHRS.

Yes, due to the following reasons, believe linear regression will not be appropriate model for this data.

- The R-squared value is 0.1251 which indicates that the model is not able to explain the variance in the data.

(c) Write down a Tobit model and the corresponding likelihood.

The tobit model is given by,

$$z_i = x_i'\beta + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2)$$

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

The corresponding likelihood function is given by,

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \left[\Phi\left(\frac{-x_i'\beta}{\sigma}\right) I(y_i = 0) + \frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) I(y_i > 0) \right]$$

(d) Fit a Tobit model and report coefficient estimates, standard errors, and t-values in a table. Comment on the effect of each variable on the response variable.

```
tobit_model = AER::tobit(WHRS ~ WomenEduc+WomenExp+WomenAge+childl6,
                         left=0, data = data_mroz)
summary(tobit_model)
```

Coefficients	Estimate	Std. Error	t-value
(Intercept)	1349.876	386.299	3.494
WomenEduc	73.291	20.474	3.580
WomenExp	80.535	6.287	12.808
WomenAge	-60.767	6.888	-8.822
childl6	-918.918	111.660	-8.230

Table 7: Results of the tobit regression model on WHRS.

A unit increase in `WomenEduc` and `WomenExp` corresponds to an increase in the `WHRS` while a unit increase in `WomenAge` and `childl6` corresponds to a decrease in the `WHRS`.

(e) What is the marginal effect of on observed hours of work for another year of education? Assume `WomenEduc`, `WomenExp`, and `WomenAge` are set at the corresponding mean values and `childl6` = 1

```
mean_vals_df = data.frame(WomenEduc = mean(data_mroz$WomenEduc),
                          WomenExp = mean(data_mroz$WomenExp),
                          WomenAge = mean(data_mroz$WomenAge),
                          childl6=1)

# Marginal effect
pnorm(predict(tobit_model, mean_vals_df)/tobit_model$scale) * tobit_model$coefficients[2]
```

Marginal effect on `WHRS` for another year of education is **26.605**.