Problem 3 - OLS, Logit, and Probit Model

Taufiqur Rohman

2022-06-20

Exercise 3 gives data for 2,000 women regarding work (1 = a woman works, 0 = otherwise), age, marital status (1 = married, 0 = otherwise), number of children, and education (number of years of schooling). Out of a total of 2,000 women, 657 were recorded as not being wage earners.

The problems are written in below. For now, let's upload and clean the data set to the R environment.

Let's upload the data set. Looking from the data set, they skip 2 rows to the column names. So, I separate the column names and the data frame first, and combine it again.

```
path <- '~/Documents/RU/Econometrics II/06-HW-Jun/Exercise_3.xls'

cols_exec3 <- as.character(read_excel(path, skip = 2, n_max = 1, col_names = FALSE))
```

```
## New names:
## * ‘‘ -> ...1
## * ‘‘ -> ...2
## * ‘‘ -> ...3
## * ‘‘ -> ...4
## * ‘‘ -> ...5
## * ...
```

```
df_exec3 <- read_excel(path, skip = 3, col_names = cols_exec3)

head(df_exec3)
```

```
## # A tibble: 6 x 15
##       c1      c2      u       v county   age education married children select
##    <dbl>   <dbl>  <dbl>   <dbl>  <dbl> <dbl>     <dbl>   <dbl>    <dbl>  <dbl>
## 1 -0.436 -0.0969 -0.218 -0.376       1    22        10       1        0   16.8
## 2  0.352  0.300   0.176  0.461       2    36        10       1        0   32.4
## 3  1.08  -1.60    0.539 -0.376       3    28        10       1        0   19.2
## 4  1.02  -1.71    0.511 -0.497       4    37        10       1        0   21.3
## 5 -0.443  0.308  -0.221 -0.0925      5    39        10       1        1   32.0
## 6 -0.440  0.613  -0.220  0.126       6    33        10       1        2   37.2
## # ... with 5 more variables: wagefull <dbl>, wage <dbl>, lw <dbl>, work <dbl>,
## #   lwf <dbl>
```

Looking from the data set, there's no problem in it. All of the data types are correct (dbl). We can move on now to the analysis.

Problem 1

Using these data, estimate the linear probability model (LPM).

```r
formula <- work ~ age + married + children + education
data <- df_exec3

p1_lpm <- lm(formula, data)
summary(p1_lpm)
```

```
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0703 -0.4142  0.1372  0.3437  0.8060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.207323   0.054111  -3.831 0.000131 ***
## age          0.010255   0.001227   8.358  < 2e-16 ***
## married      0.111112   0.021948   5.063 4.52e-07 ***
## children     0.115308   0.006772  17.028  < 2e-16 ***
## education    0.018601   0.003250   5.724 1.20e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4199 on 1995 degrees of freedom
## Multiple R-squared:  0.2026, Adjusted R-squared:  0.201
## F-statistic: 126.7 on 4 and 1995 DF,  p-value: < 2.2e-16
```

From the result above, the LPM regression will be like this:

$E(work) = -2.073 + 0.010 age + 0.111 married + 0.115 children + 0.019 education$

Problem 2

Using the same data, estimate a logit model and obtain the marginal effects of the various variables.

```r
p2_logit <- glm(formula, data, family = binomial(link = "logit"))

summary(p2_logit) # summary of logit regression
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6212  -0.9292   0.4614   0.8340   2.0455
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -4.159247   0.332040 -12.526  < 2e-16 ***
## age          0.057930   0.007221   8.022 1.04e-15 ***
## married      0.741777   0.126471   5.865 4.49e-09 ***
## children     0.764488   0.051529  14.836  < 2e-16 ***
## education    0.098251   0.018652   5.268 1.38e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2532.4  on 1999  degrees of freedom
## Residual deviance: 2055.8  on 1995  degrees of freedom
## AIC: 2065.8
## 
## Number of Fisher Scoring iterations: 5
```

```
logitmfx(formula, data) # look for the marginal effects
```

```
## Call:
## logitmfx(formula = formula, data = data)
## 
## Marginal Effects:
##               dF/dx Std. Err.       z     P>|z|    
## age       0.0115031 0.0014236  8.0801 6.469e-16 ***
## married   0.1545671 0.0270286  5.7186 1.074e-08 ***
## children  0.1518030 0.0093768 16.1893 < 2.2e-16 ***
## education 0.0195096 0.0036991  5.2742 1.333e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## dF/dx is for discrete change for the following variables:
## 
## [1] "married"
```

From the result above, the Logit regression will be like this:

$E(work) = -4.159 + 0.058age + 0.742married + 0.764children + 0.098education$

Where the marginal effects is 1. If a woman is one year older, then the probability would be 0.012 higher 2. If a woman is married, then the probability would be 0.155 higher 2. If a woman has one more child, then the probability would be 0.152 higher 3. If a woman has an additional year of schooling, then the probability would be 0.020 higher

In this context, it is the probability that a woman is working or not.

Problem 3

Repeat (2) for the probit model.

```
p3_probit <- glm(formula, data, family = binomial(link = "probit"))

summary(p3_probit) # summary of probit regression
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "probit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7594  -0.9414   0.4552   0.8459   2.0427
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.467365   0.192291 -12.831  < 2e-16 ***
## age          0.034721   0.004232   8.204 2.33e-16 ***
## married      0.430857   0.074310   5.798 6.71e-09 ***
## children     0.447325   0.028642  15.618  < 2e-16 ***
## education    0.058365   0.011018   5.297 1.18e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2532.4  on 1999  degrees of freedom
## Residual deviance: 2054.1  on 1995  degrees of freedom
## AIC: 2064.1
##
## Number of Fisher Scoring iterations: 5
```

```
probitmfx(formula, data) # look for the marginal effects
```

```
## Call:
## probitmfx(formula = formula, data = data)
##
## Marginal Effects:
##               dF/dx Std. Err.       z     P>|z|
## age       0.0117210 0.0014235  8.2341 < 2.2e-16 ***
## married   0.1504779 0.0264380  5.6917 1.258e-08 ***
## children  0.1510059 0.0091814 16.4469 < 2.2e-16 ***
## education 0.0197024 0.0037176  5.2997 1.160e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "married"
```

From the result above, the Probit regression will be like this:

E(work) = -2.467 + 0.035age + 0.431married + 0.447children + 0.058education

Where the marginal effects is 1. If a woman is one year older, then the probability would be 0.012 higher 2. If a woman is married, then the probability would be 0.150 higher 2. If a woman has one more child, then the probability would be 0.151 higher 3. If a woman has an additional year of schooling, then the probability would be 0.020 higher

In this context, it is the probability that a woman is working or not.

Problem 4

With proper transformation compare three estimated results, OLS, logit, and probit. Which model would you choose? Why?

```r
print("R2 for OLS model is")
```

```
## [1] "R2 for OLS model is"
```

```r
summary(p1_lpm)$r.squared
```

```
## [1] 0.2026228
```

```r
print("Pseudo R2 for Logit model is")
```

```
## [1] "Pseudo R2 for Logit model is"
```

```r
print(with(summary(p2_logit), 1 - deviance/null.deviance))
```

```
## [1] 0.188204
```

```r
print("Pseudo R2 for Probit model is")
```

```
## [1] "Pseudo R2 for Probit model is"
```

```r
print(with(summary(p3_probit), 1 - deviance/null.deviance))
```

```
## [1] 0.1888775
```

For the binary variable prediction, the choice will be between Logit and Probit model. The problem lies in the OLS is that there is possibility that we'll obtain y < 0 or y > 1. It doesn't make sense since a probability can't be less than 0 or more than 1. This is a fundamental issue with the LPM that we are unable to resolve.

In GLM model, we use pseudo R-squared as a measurement for goodness of fit. Comparing the value between Logit and Probit model, there's only a slight difference between both of them. But we can choose the Probit model because the pseudo is slightly higher than that of the Logit model. There are also fewer possibilities of it becoming heteroskedastic.