Problem 4 - OLS, Logit, Probit, and Tobit Model

Taufiqur Rohman

2022-06-20

The file Exercise 4 contains the 1998 Current Population Survey data in the United States. To estimate the linear probability model for the likelihood of being a union, suppose we specify the following model.

$Union = \beta_0 + \beta_1(Potexp) + \beta_2(Potexp)^2 + \beta_3(Grade) + \beta_4(Married) + \beta_5(High) + e$

where,

- Potexp (potential experience) = age – year of schooling – 5, which for men is often reasonable approximation of the number of years they have been in the labor force.
- Grade = number of years of schooling completed.
- Married = a dummy that equals 1 if the worker is married and 0 otherwise.
- High = a dummy variable that equals 1 if the worker is in a "highly" unionized industry and 0 otherwise.

Let's upload the data set. There are multiple sheets in this data set. Thus I create a function to read automatically these 2 sheets.

```
multiplesheets <- function(fname) {

  # getting info about all excel sheets
  sheets <- readxl::excel_sheets(fname)
  tibble <- lapply(sheets, function(x) readxl::read_excel(fname, sheet = x))
  data_frame <- lapply(tibble, as.data.frame)

  # assigning names to data frames
  names(data_frame) <- sheets

  # print data frame
  print(data_frame)
}
```

After that, I executed the function to the excel file. I hide the code as it will appear in the knitted file. It only specifies the path and then executing the code into the path and save the value to a variable called "tibble".

```
df_prob_1 <- as.data.frame(tibble$Sheet1)
df_prob_2 <- as.data.frame(tibble$Shhet2)
```

```
head(df_prob_1)
```

```
##   exp2 grade ind1 married potexp union high
## 1  484          8       0     22     0    1
## 2    4          14       0      2     0    0
## 3  484          16       1     22     0    0
## 4 1156           8       1     34     1    1
## 5 2209           9       1     47     0    1
## 6 1024           9       1     32     0    0
```

```
head(df_prob_2)
```

```
##   age exp2 grade ind1 ind1 married   lnwage occ1 parttime potexp union   weight
## 1  35  484          8    4       0 2.331172    7        0     22     0 12061.39
## 2  21    4         14    8       0 1.504077    5        0      2     0  5161.85
## 3  43  484         16    9       1 3.911523    5        0     22     0  6085.48
## 4  47 1156          8    2       1 2.197225    7        0     34     1  8713.88
## 5  61 2209          9    4       1 2.788093    8        0     47     0  8022.82
## 6  46 1024          9    7       1 2.351375    8        0     32     0  3589.65
##   high
## 1    1
## 2    0
## 3    0
## 4    1
## 5    1
## 6    0
```

The column names in the data set is are not tidy. Let's convert the name to the appropriate format that could be readable by R.

```
colnames(df_prob_1) <- c("potexpsq", "grade", "married", "potexp", "union", "high")
colnames(df_prob_2) <- c("age", "potexpsq", "grade", "ind", "married", "lwage", "occ", "parttime", "pote
```

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 : Sheet 1

Using sheet1,

(1) Describe the model and determine the expected sign of the parameters.

I expected all of the parameters will positively influence the probability of the person being in the union worker, except for the grade. In my intuition, the higher the person getting into higher education, the higher their probability to get more decent jobs and be more individualistic. While in the reality, most of the workers that are involved in the union come from the blue collars worker.

(2) Estimate the linear probability model and evaluate your estimated results.

```
formula <- union ~ potexp + potexpsq + grade + married + high
data_1 <- df_prob_1

p1_lpm <- lm(formula, data_1)
summary(p1_lpm)
```

```
##
## Call:
## lm(formula = formula, data = data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49138 -0.26809 -0.15694  0.01529  1.00311
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.021e-01  7.493e-02   1.363   0.1732
## potexp       2.004e-02  3.897e-03   5.142 3.27e-07 ***
## potexpsq    -3.706e-04  8.188e-05  -4.526 6.75e-06 ***
## grade       -1.246e-02  5.100e-03  -2.444   0.0147 *
## married      1.334e-02  3.000e-02   0.445   0.6566
## high         1.439e-01  2.568e-02   5.605 2.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3951 on 994 degrees of freedom
## Multiple R-squared:  0.08373,    Adjusted R-squared:  0.07912
## F-statistic: 18.17 on 5 and 994 DF,  p-value: < 2.2e-16
```

$Union = 0.021 + 0.020 Potexp - 0.000 Potexp^2 - 0.013 Grade + 0.013 Married + 0.144 High$

From this equation, I found that number of year schooling and worker that work in highly unionized industry have a positive correlation to the probability that worker joins a worker union. I also found that marriage is not a good variable to explain the dependent variable.

```
p1coef_intercept <- p1_lpm[["coefficients"]][["(Intercept)"]]
p1coef_potexp <- p1_lpm[["coefficients"]][["potexp"]]
p1coef_potexpsq <- p1_lpm[["coefficients"]][["potexpsq"]]

det_1 <- (-p1coef_potexp / (2*p1coef_potexpsq))
focmaxima_1 <- p1coef_intercept*det_1 + p1coef_potexp*det_1 + p1coef_potexpsq*(det_1^2)

print(focmaxima_1)
```

```
## [1] 3.032365
```

For the potential experience, I am trying to interpret it by looking for the focal maxima value of the parabolic curve as the equation is in second degree polynomial. With the assumption that all of other values hold at the same level, the probability of the worker getting in the union will diminish after worker have a more than 3 years' experience.

(3) Try to estimate the logit and probit model. Compare the estimated results with that of the linear probability model.

```
p1_logit <- glm(formula, data_1, family = binomial(link = "logit"))

summary(p1_logit) # summary of logit regression
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3111  -0.7477  -0.5317  -0.2883   2.4990
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.5814359  0.5186858  -4.977 6.46e-07 ***
## potexp       0.1474021  0.0280970   5.246 1.55e-07 ***
## potexpsq    -0.0026869  0.0005654  -4.752 2.01e-06 ***
## grade       -0.0703209  0.0321420  -2.188   0.0287 *
## married      0.1154630  0.1967790   0.587   0.5574
## high         0.9801411  0.1800490   5.444 5.22e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1043.60  on 999  degrees of freedom
## Residual deviance:  951.11  on 994  degrees of freedom
## AIC: 963.11
##
## Number of Fisher Scoring iterations: 5
```

```
logitmfx(formula, data_1) # look for the marginal effects
```

```
## Call:
## logitmfx(formula = formula, data = data_1)
##
## Marginal Effects:
##                dF/dx   Std. Err.       z     P>|z|
## potexp    2.2275e-02  4.0626e-03  5.4830 4.182e-08 ***
## potexpsq -4.0605e-04  8.2573e-05 -4.9174 8.769e-07 ***
## grade    -1.0627e-02  4.8540e-03 -2.1893   0.02857 *
## married   1.7271e-02  2.9112e-02  0.5933   0.55300
## high      1.4263e-01  2.4415e-02  5.8421 5.156e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "married" "high"
```

```
p1_probit <- glm(formula, data_1, family = binomial(link = "probit"))

summary(p1_probit) # summary of probit regression
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "probit"), data = data_1)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2876  -0.7562  -0.5381  -0.2609   2.5611
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.468410   0.291029  -5.046 4.52e-07 ***
## potexp       0.083509   0.015548   5.371 7.83e-08 ***
## potexpsq    -0.001531   0.000317  -4.828 1.38e-06 ***
## grade       -0.042078   0.018658  -2.255   0.0241 *
## married      0.062252   0.112379   0.554   0.5796
## high         0.561295   0.099697   5.630 1.80e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1043.6  on 999  degrees of freedom
## Residual deviance:  950.5  on 994  degrees of freedom
## AIC: 962.5
##
## Number of Fisher Scoring iterations: 5
```

```
probitmfx(formula, data_1) # look for the marginal effects
```

```
## Call:
## probitmfx(formula = formula, data = data_1)
##
## Marginal Effects:
##                dF/dx   Std. Err.       z     P>|z|
## potexp    0.02269634  0.00413886  5.4837 4.165e-08 ***
## potexpsq -0.00041604  0.00008481 -4.9056 9.314e-07 ***
## grade    -0.01143610  0.00507078 -2.2553   0.02411 *
## married   0.01678813  0.03005886  0.5585   0.57650
## high      0.14709867  0.02472988  5.9482 2.711e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "married" "high"
```

Logit and Probit model produce a similar result with only slight difference in the coefficient (magnitude of each variable). Compared to linear probability model, I found that three of them also produce a similar result, in terms of the sign parameter and the significance of the variable. But they have huge difference in the coefficient and the standard error of the estimate.

Problem 2: Sheet 2

Using the same data (sheet2) calculate the following linear regression model for log wages (lwage):

$lwage = \beta_0 + \beta_1(Potexp) + \beta_2(Potexp)^2 + \beta_3(Grade) + \beta_4(Married) + \beta_5(High) + e$

(1) Perform OLS for this equation

```r
formula_2 <- lwage ~ potexp + potexpsq + grade + married + high
data_2 <- df_prob_2

p2_lpm <- lm(formula_2, data_2)
summary(p2_lpm)
```

```
##
## Call:
## lm(formula = formula_2, data = data_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81658 -0.29028  0.00997  0.30614  1.76230
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.704e-01  8.640e-02   5.444 6.56e-08 ***
## potexp       4.191e-02  4.493e-03   9.327  < 2e-16 ***
## potexpsq    -5.759e-04  9.441e-05  -6.100 1.52e-09 ***
## grade        9.252e-02  5.881e-03  15.732  < 2e-16 ***
## married      9.428e-02  3.459e-02   2.726  0.00653 **
## high         8.572e-02  2.961e-02   2.895  0.00387 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4555 on 994 degrees of freedom
## Multiple R-squared:  0.3497, Adjusted R-squared:  0.3464
## F-statistic: 106.9 on 5 and 994 DF,  p-value: < 2.2e-16
```

$lwage = 0.470 + 0.042 Potexp - 0.001 Potexp^2 + 0.093 Grade + 0.094 Married + 0.0857 High$

From this equation, I found that number of year schooling, marriage status, and worker that work in highly unionized industry have a positive correlation to the probability that worker joins a worker union.

```r
p2coef_intercept <- p2_lpm[["coefficients"]][["(Intercept)"]]
p2coef_potexp <- p2_lpm[["coefficients"]][["potexp"]]
p2coef_potexpsq <- p2_lpm[["coefficients"]][["potexpsq"]]

det_2 <- (-p2coef_potexp / (2*p2coef_potexpsq))
focmaxima_2 <- p2coef_intercept*det_1 + p2coef_potexp*det_1 + p2coef_potexpsq*(det_1^2)

print(focmaxima_2)
```

```
## [1] 13.42896
```

Using similar method that I used in the previous number, the worker will get higher wage until they work for 13 years. After that, the worker will get a lower wage. It seems doesn’t make sense, but as the problem states and build the model like that, I will ignore this.

2) Next, generate a new variable, say clwage such that:

if lwage >= 1.87, then clwage = lwage, otherwise clwage = 0

```
df_prob_2["clwage"] = ifelse(df_prob_2$lwage >= 1.87, df_prob_2$lwage, 0)
```

Now, perform a Tobit on the same model, replacing lwage with clwage. How do your estimates of the relevant coefficients compare?

In R, we can use two-ways. Either using tobit in vglm package by censoring the lower bound to 1.87, or manually transform the Dependent Variable first, and then conducting Tobit regression without defining any lower bound. But the latter way involves redundant steps.

In this case, as I already got the filtered DV, I will conduct the regression with the clwage variable.

```
formula_3 <- clwage ~ potexp + potexpsq + grade + married + high

p2_tobit <- vglm(formula_3, tobit(Lower = 1.87), data = df_prob_2)
```

```
## Warning in eval(slot(family, "initialize")): replacing response values less than
## ’Lower’ by ’Lower’
```

```
summary(p2_tobit)
```

```
##
## Call:
## vglm(formula = formula_3, family = tobit(Lower = 1.87), data = df_prob_2)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.3643325  0.0958909   3.799 0.000145 ***
## (Intercept):2 -0.7788439  0.0261178 -29.820  < 2e-16 ***
## potexp         0.0435822  0.0048280   9.027  < 2e-16 ***
## potexpsq      -0.0005795  0.0001010  -5.740 9.47e-09 ***
## grade          0.0988641  0.0063259  15.628  < 2e-16 ***
## married        0.0825541  0.0362573   2.277 0.022792 *
## high           0.0782053  0.0311785   2.508 0.012131 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Names of linear predictors: mu, loglink(sd)
##
## Log-likelihood: -675.2737 on 1993 degrees of freedom
##
## Number of Fisher scoring iterations: 8
##
## No Hauck-Donner effect found in any of the estimates
```

Like OLS regression coefficients, Tobit regression coefficients are interpreted in a similar way; however, the linear influence is on the uncensored latent variable rather than the observed outcome.

After censoring value below 1.87, we found that all the variables have similar sign of parameter and are significant at 95% confidence interval although, married and high variables have weaker p-value.

Another thing that we can notice is that there are 2 intercepts in this model. The first intercept is the usual intercept of the tobit model. The second intercept is the log-standard deviation of the latent variable.

```
p3coef_intercept <- p2_tobit@coefficients[["(Intercept):1"]]
p3coef_potexp <- p2_tobit@coefficients[["potexp"]]
p3coef_potexpsq <- p2_tobit@coefficients[["potexpsq"]]

det_3 <- (-p3coef_potexp / (2*p3coef_potexpsq))
focmaxima_3 <- p3coef_intercept*det_1 + p3coef_potexp*det_1 + p3coef_potexpsq*(det_1^2)

print(focmaxima_3)
```

```
## [1] 10.60518
```

Again, with same the same way, holding all the variables at constant value, I found that the worker will get higher wage until they work for 10 years. After that, the worker will get a lower wage.