Assignment 5: Binary, Multinomial, Tobit, Selection (b13)

A. Colin Cameron U.C.-Davis

Data are at http://cameron.econ.ucdavis.edu/bgpe2013

1. Logit and probit. Use data in file `mus14data.dta`
We will do analysis similar to that in the slides, but with a new regressor `linc`.

(a) Generate a new variable `linc = ln(hhincome)`
You will see that for nine observations a missing value is created. Explain why.

(b) Give command `scatter ins linc`
What does this graph suggest is the relationship between insurance and household income?

(c) Give command `scatter ins linc, jitter(5) msize(tiny) || lfit ins linc`
Is this more helpful in explaining the relationship between the two variables?

(d) From your answer in part (c), can you see problems with OLS estimation?

(e) Perform logit regression of `ins` on `linc`
Is there a statistically significant relationship?

(f) Use command `predict` to compute variable `plogit`, the logit model prediction of the probability of someone holding insurance. Then give commands

```
sort linc
scatter ins linc, jitter(5) msize(tiny) || line plogit linc, clstyle(p1)
```

Comment on the resulting graph.

(g) Manually compute the marginal effect on `ins` of a change in `linc` computed at the sample mean value of `linc`. (Use command `display`).
Compare your answer with that obtained from command `margins, dydx(*)`.

(h) Manually compute the average marginal effect on `ins` of a change in `linc` computed at the ample mean value of `linc`. (Use command `generate` and then command `summarize`).
Compare your answer with that obtained from command `margins, dydx(*) atmean.`

(i) Given your answer in the previous part, if `hhincome` increases by 10 percent what is the sample average increase in the probability of having insurance?

(j) Now perform probit regression of `ins` on `linc`.
Compare the logit and probit estimates on the basis of (1) estimated coefficient; (2) statistical significance; (3) likelihood; (4) average predicted probability of having insurance; (5) average marginal effect. Is there much difference between the two?

(k) Now perform logit regression of `ins` on `linc retire age hstatusg educyear married hisp`.
Does `linc` remain an important explanator of having private health insurance?

2. Logit model. Consider the logit model with $y_i = 1$ with probability $\Lambda(\mathbf{x}_i'\boldsymbol{\beta})$ and $y_i = 0$ with probability $1-\Lambda(\mathbf{x}_i'\boldsymbol{\beta})$, where $\Lambda(z) = e^z/(1+e^z)$ and $\Lambda'(z) = \Lambda(z)(1-\Lambda(z))$. Data are independent over i.

(a) Show that $\ln L = \sum_{i=1}^N y_i \ln \Lambda(\mathbf{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln(1-\Lambda(\mathbf{x}_i'\boldsymbol{\beta}))$.

(b) Show after some algebra the first-order conditions for the MLE $\widehat{\boldsymbol{\beta}}$ are

$$\textstyle\sum_{i=1}^N (y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}.$$

(c) Hence state the essential condition for $\widehat{\boldsymbol{\beta}}$ to be consistent.

(d) Give the asymptotic distribution for $\widehat{\boldsymbol{\beta}}$ using the result that the variance matrix of the MLE is minus the inverse of the expected value of the second derivatives of the log-likelihood.

3. Probit simulation
(a) Generate the following data
- Sample size is 400
- Seed is set to 10101
- Regressor x is uniform on (0,1) - use function `runiform`
- Latent variable $y^* = -2.5 + 4x + \varepsilon$ where ε is standard normal - use `rnormal(0,1)`
- Observed variable $y = 1$ if $y^* > 0$ and $y = 0$ if $y^* \leq 0$.

(b) Check that the generate data is as expected, using command `summarize`.

(c) Show that for this d.g.p. $\Pr[y = 1|\mathbf{x}] = \Phi(-2.5 + 4x)$.

(d) Perform probit regression of y on x.
Are the estimates what you expect? Explain.

4. Multinomial logit. Use data in file `mus15data.dta`
We will do analysis similar to that in the slides, but with one less alternative.
Specifically, drop all individuals who fish from the pier, leading to three alternatives.
To find the alternatives use `tabulate mode` and `tabulate mode, nolabel`

(a) Estimate a multinomial logit model with regressors an intercept and income, with charter boat fishing the base category.

(b) What is the effect on charter fishing of an increase in income? Give both the AME and the MEM. This uses option `predict(outcome())` and you need to choose the correct outcome.

(c) Consider the discrete random variable y_i that takes value 1 with probability $p_{1i} = F_1(\mathbf{x}_i'\boldsymbol{\beta})$; value 2 with probability $p_{1i} = F_1(\mathbf{x}_i'\boldsymbol{\beta})$; and value 3 with probability $p_{1i} = F_1(\mathbf{x}_i'\boldsymbol{\beta})$.
Define three binary variables $y_{1i} = 1$ if $y_i = 1$ and 0 otherwise; $y_{3i} = 1$ if $y_i = 2$ and 0 otherwise; and $y_{3i} = 1$ if $y_i = 3$ and 0 otherwise. Verify that

$$f(y_i) = p_{1i}^{y_{1i}} p_{2i}^{y_{2i}} p_{3i}^{y_{3i}}.$$

(d) Hence show that

$$\ln L(\boldsymbol{\beta}) = \textstyle\sum_{i=1}^{N} y_{1i} \ln F_1(\mathbf{x}_i'\boldsymbol{\beta}) + y_{2i} \ln F_2(\mathbf{x}_i'\boldsymbol{\beta}) + y_{3i} \ln F_3(\mathbf{x}_i'\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{3} y_{ji} \ln F_j(\mathbf{x}_i'\boldsymbol{\beta}).$$

(e) Show that the first-order conditions for the MLE $\widehat{\boldsymbol{\beta}}$ are $\sum_{i=1}^{N} \sum_{j=1}^{3} y_{ji} \frac{F_j'(\mathbf{x}_i'\boldsymbol{\beta})}{F_j(\mathbf{x}_i'\boldsymbol{\beta})} \mathbf{x}_i = \mathbf{0}$.

5. Tobit Use data in file `mus16data.dta`
We will do analysis similar to that in the slides, but with fewer regressors.

(a) Estimate a Tobit model of `ambexp` regressed on `totchr`.

(b) Compare your results to those from OLS. Are you surprised? Explain.

(c) Estimate a sample selection model of `lambexp` (this is $\ln y$ for $y > 0$ and missing otherwise) regressed on `totchr`, where `totchr` appears in both the selection and the participation equations. Use the MLE option? Does there appear to be selection on unobservables? Is the error correlation what you expect?

(d) Repeat (c) using Heckman's two-step estimator.

(e) Suppose $y_i^* \sim \mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta},\, \sigma^2]$ and we observe $y_i = y_i^*$ only if $y_i^* > \mathbf{z}_i'\boldsymbol{\gamma}$. Obtain $\mathrm{E}[y_i|\mathbf{x}_i, \mathbf{z}_i]$.