ECOM40006/90013 Econometrics 3

Week 11 Extras

Question 1: GLM, Maximum Likelihood, and You

Before delving into the more nitty-gritty aspects of generalized linear models, let's start off with the basics. Consider, for example, the classic linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $y_i \in \{0, 1\}$ is a binary variable and for now, u_i is an i.i.d. mean zero disturbance, with its distribution being symmetric around zero.

(a) The first way that we're going to interpret these models is via straight OLS. The resulting model is often called the *linear probability model.*

(i.) Give an interpretation of the coefficient $\hat{\beta}_1$ estimated by OLS.

(ii.) Describe the advantages of using a linear probability model.

(iii.) Correspondingly, what are the drawbacks?

(b) The *Bernoulli* probability mass function forms the foundation for an alternative type of model: the *generalized linear model*, or GLM. Its probability mass function is

$$p(y = y_i; \theta) = \theta^{y_i}(1 - \theta)^{1-y_i}$$

where $\theta \in [0, 1]$ and $y_i = \{0, 1\}$.

(i.) The parameter θ can be interpreted as a probability. To see why, substitute in $y_i = 1$ and $y_i = 0$. What do you get?

(ii.) Obtain the log-likelihood and find the MLE $\hat{\theta}$.

(c) The mean of a Bernoulli random variable is $\mathbb{E}(y_i) = \theta = \mathbf{P}(y_i = 1)$. If we assume this mean depends on x_i, we naturally get the conditional mean $\mathbb{E}(y_i|x_i) = \mathbf{P}(y_i = 1|x_i)$. Suppose that the true data generating process for y_i proceeds as follows:

$$y_i = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 x_i + u_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

(i.) Explain why this type of data generating process can be said to fall into the category of *latent*, or *unobservable* variables models.

(ii.) The source of randomness in this model comes from the disturbance u_i. Denote the CDF of u_i as $F(u_i)$. Write the probability that $y_i = 1$ in terms of $F(\cdot)$, the regressor x_i and the parameters β_0 and β_1.

(d) One way to get around the drawbacks of the linear probability model is to assume that the probability $\mathbf{P}(y_i = 1)$ depends on the regressors in some way. That is, we can replace $\theta = \theta_i$, where one possibility for θ_i is the *probit link function*

$$\theta_i = \Phi(\beta_0 + \beta_1 x_i) = \int_{\infty}^{\beta_0+\beta_1 x_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. In this sense, we can write the distribution of an individual y_i as

$$p(y_i; \theta_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}.$$

(i.) In terms of the predicted values of θ_i, why might this potentially be more preferable to using a linear probability model?

(ii.) Obtain expressions for the partial derivatives

$$\frac{\partial \theta_i}{\partial \beta_0}, \qquad \frac{\partial \theta_i}{\partial \beta_1}$$

using the *Fundamental Theorem of Calculus*[1] (and the Chain Rule). Use $\phi(\cdot)$ to denote the standard normal PDF:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

It might help to define $\phi_i = \phi(\beta_0 + \beta_1 x_i)$.

(iii.) Suppose $\beta = (\beta_0, \beta_1)$. Noting that θ_i is a function of β, the log-likelihood for this *probit* model can be written

$$\log L(\beta; y, X) = \sum_{i=1}^{n} [y_i \log(\theta_i) + (1-y_i)\log(1-\theta_i)]$$

(this can be obtained from part (b) as well by making appropriate substitutions.) Derive the score function. Do the first-order conditions have an analytic solution?

[1] Let f be continuous on $[a, b]$. If $F(y) = \int_a^y f(x)\,dx$ for $a \le y \le b$, then F is differentiable:

$$\frac{dF(y)}{dy} = \frac{d}{dy}\left(\int_a^y f(x)\,dx\right) = f(y).$$

(iv.) A natural (albeit painful) progression of the concepts from here is to transition to matrix notation. In preparation for this, consider adding another regressor so we have

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_i. \tag{1}$$

i. Obtain an expression for $\dfrac{\partial \theta_i}{\partial \beta_2}$.

ii. With the support of your previous answers (particularly from (d)-(ii) and your partial derivative just above), infer the form of the partial derivative

$$\frac{\partial \theta_i}{\partial \beta_j}.$$

iii. Hence, infer an expression for the *gradient vector*

$$\frac{\partial \theta_i}{\partial \beta}$$

in matrix form, for an arbitrary number of parameters k.

(e) Demonstrate the following properties of the probit model:

(i.) The marginal effects are not constant.

(ii.) Relative marginal effects are constant.

(iii.) The sign of the coefficient estimate tells you what effect it has on $\mathbf{P}(y_i = 1)$.

(f) Describe three ways in which you could potentially interpret the output from a probit regression model.

Question 2: GLM Formalities

If you're feeling more familiar with the foundations of binary response models, then this question will take things a bit further and formalize everything in matrix notation. It's recommended you give question 1 a look before you try this stuff out!

Consider a sample of n observations on a binary variable $y_i \in \{0, 1\}$. You have available k regressors, represented in the $1 \times k$ row vector x_i' for a single observation i. You are interested in the latent variables model

$$y_i = x_i'\beta + u_i,$$

where u_i is a mean zero disturbance and a distribution function $F(u_i)$ (with associated PDF $f(u_i)$) which is symmetric around zero. As alluded to in Question 1, the expression $F(x_i'\beta)$ represents the probability that y_i equals 1, conditional on the regressors x_i.

(a) Suppose that in the sample of size n, you observe that

- $y_i = 1$ occurs n_1 times

- $y_i = 0$ occurs n_2 times

so that $n_1 + n_2 = n$. Use this to obtain an expression for the joint density of the sample y, conditional on the regressors X. Carefully explain your steps. (*Remember: this is just the fancy way of saying "go find the likelihood function".*)

(b) Obtain an expression for the log-likelihood function. If you observe any similarities between this question and previous questions, you may skip the relevant working with justification.

(c) Show that the score function can be written in the form

$$S(\beta) = \sum_{i=1}^{n} x_i \nu_i(\beta), \qquad \text{where} \qquad \nu_i(\beta) = \frac{[y_i - F(x_i'\beta)]f(x_i'\beta)}{F(x_i'\beta)[1 - F(x_i'\beta)]}$$

is called the *generalized residual* when evaluated at the MLE $\hat{\beta}$.

- *Hint*: it is always a good idea to remember that both $x_i'\beta$ and $F(x_i'\beta)$ are scalars. The Chain Rule is used liberally here to get the desired answer.
- The notation can easily get out of hand. Shorthand like $F_i = F(x_i'\beta)$ and $f_i = f(x_i'\beta)$ may help reduce notational burden.

(d) Show that the score function has zero expected value. *Hint:* $\mathbb{E}(y_i|x_i) = F(x_i'\beta)$.

(e) (i.) Show that the score multiplied by its transpose gives

$$\sum_{i=1}^{n} x_i x_i' \nu_i(\beta)^2 + \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j' \nu_i(\beta) \nu_j(\beta).$$

(ii.) Hence derive an expression for the conditional variance of the score.

(f) Let $\gamma_i = (y_i - F_i)f_i$ and $\theta_i = F_i(1 - F_i)$, with F_i and f_i defined as in part (c) above. Consider the expression

$$\nu_i \equiv \nu_i(\beta) = \frac{(y_i - F_i)f_i}{F_i(1 - F_i)} = \frac{\gamma_i}{\theta_i}.$$

(i.) Show that for any $j = 1, \ldots, k$,

$$\Gamma_{ij} = \frac{\partial \gamma_i}{\partial \beta_j} = \left([y_i - F_i] \frac{\partial f_i}{\partial x_i'\beta} - f_i^2 \right) x_{ij},$$
$$\Theta_{ij} = \frac{\partial \theta_i}{\partial \beta_j} = (1 - 2F_i) f_i x_{ij}.$$

and hence derive an expression for $\dfrac{\partial \nu_i}{\partial \beta_j}$.

(ii.) Using your answers above, derive an expression for the Hessian $H(\beta)$.

(g) Calculate the expected value of the Hessian given the regressors x_i. How does this compare to the conditional variance of the score?