Using the Delta Method to Construct Confidence Intervals for Predicted Probabilities, Rates, and Discrete Changes

Jun Xu and J. Scott Long

Indiana University

August 22, 2005

The paper provides technical details on the methods described in Jun Xu and J. Scott Long (forthcoming) "Confidence Intervals for Predicted Outcomes in Regression Models for Categorical Outcomes" *The Stata Journal*. These formula were incorporated into `prvalue`. See www.indiana.edu/~jslsoc/spost.htm for further details.

1 General Formula

The delta method is a general approach for computing confidence intervals for functions of maximum likelihood estimates. The delta method takes a function that is too complex for analytically computing the variance, creates a linear approximation of that function, and then computes the variance of the simpler linear function that can be used for large sample inference.

We begin with a general result for maximum likelihood theory. Under standard regularity conditions, if $\widehat{\boldsymbol{\beta}}$ is a vector of ML estimates, then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left[\mathbf{0},Var\left(\widehat{\boldsymbol{\beta}}\right)\right]\ . \tag{1}$$

Let $G\left(\boldsymbol{\beta}\right)$ be some function, such as predicted probabilities from a logit or ordinal logit model. The Taylor series expansion of $G(\widehat{\boldsymbol{\beta}})$ is

$$\begin{aligned} G(\widehat{\boldsymbol{\beta}}) &= G(\boldsymbol{\beta})+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'G'(\boldsymbol{\beta})+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'G''(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})/2 \\ &\approx G(\boldsymbol{\beta})+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'G'(\boldsymbol{\beta})\ , \end{aligned} \tag{2}$$

where $G'(\boldsymbol{\beta})$ and $G''(\boldsymbol{\beta})$ are matrices of first and second partial derivatives with respect to $\boldsymbol{\beta}$, $\boldsymbol{\beta}^*$ is some value between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Then,

$$\sqrt{n}\left[G(\widehat{\boldsymbol{\beta}})-G(\boldsymbol{\beta})\right]\approx\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'G'(\boldsymbol{\beta})\ . \tag{3}$$

This leads to leads to $G(\widehat{\boldsymbol{\beta}})\rightarrow N\left(G(\boldsymbol{\beta}),\frac{\partial G(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}Var(\widehat{\boldsymbol{\beta}})\frac{\partial G(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)$ (Greene 2000; Agresti 2002). To estimate the variance, we evaluate the partials at the ML estimates, $\left.\frac{\partial G(\boldsymbol{\beta}|\mathbf{x})}{\partial\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$, which leads to

$$Var\left(G(\widehat{\boldsymbol{\beta}})\right)=\frac{\partial G(\widehat{\boldsymbol{\beta}})}{\partial\widehat{\boldsymbol{\beta}}'}Var(\widehat{\boldsymbol{\beta}})\frac{\partial G(\widehat{\boldsymbol{\beta}})}{\partial\widehat{\boldsymbol{\beta}}}\ . \tag{4}$$

For example, consider the logit model with

$$G(\boldsymbol{\beta}) = \Pr\left(y = 1 \mid \mathbf{x}\right) = \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) \ . \tag{5}$$

To compute the confidence interval, we need the gradient vector

$$\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left[\begin{array}{cccc} \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_0} & \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_K} \end{array}\right]' \ . \tag{6}$$

Since Λ is a cdf, $\frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial \Lambda(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}'\boldsymbol{\beta}} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k} = \lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) x_k$. Then

$$\frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left[\begin{array}{cccc} \lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) & \lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) x_1 & \cdots & \lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) x_K \end{array}\right]' \ . \tag{7}$$

To compute the confidence interval for a change in the probability as the independent variables change from $\mathbf{x}_a$ to $\mathbf{x}_b$, we use the function

$$G\left(\boldsymbol{\beta}\right) = \Lambda(\boldsymbol{\beta}|\mathbf{x}_a) - \Lambda\left(\boldsymbol{\beta}|\mathbf{x}_b\right) \ , \tag{8}$$

where

$$\begin{aligned} \frac{\partial G(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial\left[\Lambda(\boldsymbol{\beta}|\mathbf{x}_a) - \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)\right]}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}} - \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}} \ . \end{aligned} \tag{9}$$

Substituting this result into equation 4,

$$\begin{aligned} Var\left(G(\boldsymbol{\beta})\right) &= \left[\frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}}\right] - \left[\frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}}\right] \\ &\quad - \left[\frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_a)}{\partial \boldsymbol{\beta}}\right] + \left[\frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}'} Var(\widehat{\boldsymbol{\beta}}) \frac{\partial \Lambda(\boldsymbol{\beta}|\mathbf{x}_b)}{\partial \boldsymbol{\beta}}\right] \ . \end{aligned} \tag{10}$$

We now apply these formula to the models for which `prvalue` computes confidence intervals.

2 Binary Models

In binary models, $G(\boldsymbol{\beta}) = \Pr(y = 1 \mid \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})$ where F is the cdf for the logistic, normal, or cloglog function. The gradient is

$$\frac{\partial F(\mathbf{x}'\boldsymbol{\beta})}{\partial \beta_k} = \frac{\partial F(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}'\boldsymbol{\beta}} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k} = f(\mathbf{x}'\boldsymbol{\beta}) x_k \ , \tag{11}$$

where f is the pdf corresponding to F. For the vector $\mathbf{x}$ it follows that

$$\frac{\partial F(\mathbf{x}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = f(\mathbf{x}'\boldsymbol{\beta})\mathbf{x} \ . \tag{12}$$

From equation 4,

$$\begin{aligned} Var\left[\Pr(y=1 \mid \mathbf{x})\right] &= f(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}'Var(\widehat{\boldsymbol{\beta}})\mathbf{x}f(\mathbf{x}'\boldsymbol{\beta}) \\ &= f(\mathbf{x}'\boldsymbol{\beta})^2\mathbf{x}'Var(\widehat{\boldsymbol{\beta}})\mathbf{x} \; . \end{aligned} \tag{13}$$

The variances of $\Pr(y=0 \mid \mathbf{x})$ and $\Pr(y=0 \mid \mathbf{x})$ are the equal since

$$\frac{\partial\left[1-F(\mathbf{x}'\boldsymbol{\beta})\right]}{\partial\boldsymbol{\beta}} = -f(\mathbf{x}'\boldsymbol{\beta})\mathbf{x} \tag{14}$$

and

$$Var\left[\Pr(y=0 \mid \mathbf{x})\right] = \left[-f(\mathbf{x}'\boldsymbol{\beta})\right]^2\mathbf{x}'Var(\widehat{\boldsymbol{\beta}})\mathbf{x} \; . \tag{15}$$

3 Ordered Logit and Probit

Assume that there are $m = 1, J$ outcome categories, where

$$\Pr\left(y=m \mid \mathbf{x}\right) = F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right) - F\left(\tau_{m-1} - \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J \; . \tag{16}$$

Since we assume that $\tau_0 = -\infty$ and $\tau_J = \infty$, $F\left(\tau_0 - \mathbf{x}'\boldsymbol{\beta}\right) = 0$ and $F\left(\tau_J - \mathbf{x}'\boldsymbol{\beta}\right) = 1$. To compute the gradient,

$$\frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\beta_k} = \frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}\frac{\partial\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\beta_k} \tag{17}$$

$$= f\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)\left(-x_k\right) \tag{18}$$

and

$$\frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} = \frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}\frac{\partial\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} \; . \tag{19}$$

It follows that

$$\frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} = f\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right) \text{ if } j = m \tag{20}$$

and

$$\frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} = 0 \text{ if } j \neq m \; . \tag{21}$$

Using these results with equation 16,

$$\begin{aligned} \frac{\partial\Pr\left(y_i = m \mid \mathbf{x}_i\right)}{\partial\beta_k} &= \left[f\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)\left(-x_k\right)\right] - \left[f\left(\tau_{m-1} - \mathbf{x}'\boldsymbol{\beta}\right)\left(-x_k\right)\right] \\ &= -x_k f\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right) - \left[f\left(\tau_{m-1} - \mathbf{x}'\boldsymbol{\beta}\right)\right] \end{aligned} \tag{22}$$

and

$$\begin{aligned} \frac{\partial\Pr\left(y_i = m \mid \mathbf{x}_i\right)}{\partial\tau_j} &= \frac{\partial F\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} - \frac{\partial F\left(\tau_{m-1} - \mathbf{x}'\boldsymbol{\beta}\right)}{\partial\tau_j} \\ &= f\left(\tau_m - \mathbf{x}'\boldsymbol{\beta}\right) \text{ if } j = m \\ &= -f\left(\tau_{m-1} - \mathbf{x}'\boldsymbol{\beta}\right) \text{ if } j = m-1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{23}$$

For example, with three categories:

$$\Pr(y=1 \mid \mathbf{x}) = F(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) - 0 \tag{24}$$
$$\Pr(y=2 \mid \mathbf{x}) = F(\tau_2 - \mathbf{x}'\boldsymbol{\beta}) - F(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) \tag{25}$$
$$\Pr(y=3 \mid \mathbf{x}) = 1 - F(\tau_2 - \mathbf{x}'\boldsymbol{\beta}) \ , \tag{26}$$

then

$$\frac{\partial \Pr(y_i=1 \mid \mathbf{x}_i)}{\partial \beta_k} = -x_k \left[f(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) \right] \tag{27}$$
$$\frac{\partial \Pr(y_i=2 \mid \mathbf{x}_i)}{\partial \beta_k} = -x_k \left[f(\tau_2 - \mathbf{x}'\boldsymbol{\beta}) - f(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) \right] \tag{28}$$
$$\frac{\partial \Pr(y_i=3 \mid \mathbf{x}_i)}{\partial \beta_k} = -x_k \left[-f(\tau_2 - \mathbf{x}'\boldsymbol{\beta}) \right] \ . \tag{29}$$

With respect to τ,

$$\frac{\partial \Pr(y_i=1 \mid \mathbf{x}_i)}{\partial \tau_1} = f(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) \tag{30}$$
$$\frac{\partial \Pr(y_i=1 \mid \mathbf{x}_i)}{\partial \tau_2} = 0 \tag{31}$$
$$\frac{\partial \Pr(y_i=2 \mid \mathbf{x}_i)}{\partial \tau_1} = -f(\tau_1 - \mathbf{x}'\boldsymbol{\beta}) \tag{32}$$
$$\frac{\partial \Pr(y_i=2 \mid \mathbf{x}_i)}{\partial \tau_2} = f(\tau_2 - \mathbf{x}'\boldsymbol{\beta}) \tag{33}$$
$$\frac{\partial \Pr(y_i=3 \mid \mathbf{x}_i)}{\partial \tau_1} = 0 \tag{34}$$
$$\frac{\partial \Pr(y_i=3 \mid \mathbf{x}_i)}{\partial \tau_2} = 0 \ . \tag{35}$$

To implement these procedures in Stata, we create the augmented matrices:

$$\boldsymbol{\beta}^* = \left[\begin{array}{cccc} \boldsymbol{\beta}' & \tau_1 & \cdots & \tau_{J-1} \end{array} \right]' \tag{36}$$

and

$$\begin{aligned} \mathbf{x}_1^* &= \left[\begin{array}{ccccc} -\mathbf{x}' & 1 & 0 & \cdots & 0 \end{array} \right]' \\ \mathbf{x}_2^* &= \left[\begin{array}{ccccc} -\mathbf{x}' & 0 & 1 & \cdots & 0 \end{array} \right]' \\ &\vdots \\ \mathbf{x}_{J-1}^* &= \left[\begin{array}{ccccc} -\mathbf{x}' & 0 & 0 & \cdots & 1 \end{array} \right]' \ , \end{aligned} \tag{37}$$

such that

$$\mathbf{x}_j^{*\prime} \boldsymbol{\beta}^* = \tau_j - \mathbf{x}\boldsymbol{\beta} \ . \tag{38}$$

We then create the gradients described above.

4 Generalized Ordered Logit[1]

The generalized ordered logit model is identical to the ordinal logit model except that the coefficients associated with $\mathbf{x}$ differ for each outcome. Since there is an intercept for each outcome, the τ's are fixed to zero and $\boldsymbol{\beta}_J = \mathbf{0}$ for identification. Then,

$$\begin{aligned} \Pr(y_i = 1 \mid \mathbf{x}_i) &= F(-\mathbf{x}'\boldsymbol{\beta}_m) && (39) \\ \Pr(y_i = m \mid \mathbf{x}_i) &= F(-\mathbf{x}'\boldsymbol{\beta}_m) - F\left(-\mathbf{x}'\boldsymbol{\beta}_{m-1}\right) \text{ for } m = 2, J-1 && (40) \\ \Pr(y_i = J \mid \mathbf{x}_i) &= -F\left(-\mathbf{x}'\boldsymbol{\beta}_{m-1}\right) . && (41) \end{aligned}$$

The gradient with respect to the β's is

$$\frac{\partial F(-\mathbf{x}'\boldsymbol{\beta}_m)}{\partial \beta_{m,k}} = f(-\mathbf{x}'\boldsymbol{\beta}_m)(-x_k) \ , \tag{42}$$

while no gradient for thresholds is needed. Then,

$$\frac{\partial \Pr(y_i = m \mid \mathbf{x}_i)}{\partial \beta_{m,k}} = -x_k f(-\mathbf{x}'\boldsymbol{\beta}_m) - \left[f\left(-\mathbf{x}'\boldsymbol{\beta}_{m-1}\right)\right] \ . \tag{43}$$

5 Multinomial Logit

Assuming outcomes 1 through J,

$$\Pr(y = m|\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta}_m)}{\sum_{j=1}^{J} \exp\left(\mathbf{x}\boldsymbol{\beta}_j\right)} \ , \tag{44}$$

where without loss of generality we assume that $\boldsymbol{\beta}_1 = \mathbf{0}$ to identify the model (and accordingly, the derivatives below do not apply to the partial with respect to $\boldsymbol{\beta}_1$). To simplify notation, let $\Delta = \sum \exp\left(\mathbf{x}\boldsymbol{\beta}_j\right)$. The derivative of the probability of m with respect to $\boldsymbol{\beta}_n$ is

$$\frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_n} = \frac{\partial \exp(\mathbf{x}\boldsymbol{\beta}_m)\,\Delta^{-1}}{\partial \boldsymbol{\beta}_n} \ . \tag{45}$$

Using the quotient rule,

$$\frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_n} = \left[\Delta \frac{\partial \exp(\mathbf{x}\boldsymbol{\beta}_m)}{\partial \boldsymbol{\beta}_n} - \exp(\mathbf{x}\boldsymbol{\beta}_m) \frac{\partial \Delta}{\partial \boldsymbol{\beta}_n}\right] \Delta^{-2} \ . \tag{46}$$

Examining each partial in turn.

$$\begin{aligned} \frac{\partial \exp(\mathbf{x}\boldsymbol{\beta}_m)}{\partial \boldsymbol{\beta}_n} &= \frac{\partial \exp(\mathbf{x}\boldsymbol{\beta}_m)}{\partial \mathbf{x}\boldsymbol{\beta}_m} \frac{\partial \mathbf{x}\boldsymbol{\beta}_m}{\partial \boldsymbol{\beta}_n} && (47) \\ &= \exp(\mathbf{x}\boldsymbol{\beta}_m)\,\mathbf{x} \text{ if } m = n \\ &= 0 \text{ if } m \neq n \end{aligned}$$

[1] Confidence intervals for the generalized ordered logit model are not supported in `prvalue`.

and

$$\frac{\partial \sum_{j=1}^{J} \exp\left(\mathbf{x}\boldsymbol{\beta}_j\right)}{\partial \boldsymbol{\beta}_n} = \frac{\sum_{j=1}^{J} \partial \exp\left(\mathbf{x}\boldsymbol{\beta}_j\right)}{\partial \boldsymbol{\beta}_n} = \exp\left(\mathbf{x}\boldsymbol{\beta}_n\right)\mathbf{x} \ . \tag{48}$$

The last equality follows since the partial of $\exp\left(\mathbf{x}\boldsymbol{\beta}_j\right)$ with respect to $\boldsymbol{\beta}_n$ is 0 unless $j = n$. Combining these results. If $m = n$,

$$\begin{aligned} \frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_m} &= \left[\Delta \exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)\mathbf{x} - \exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)^2 \mathbf{x}\right]\Delta^{-2} \\ &= \left[\Delta \exp\left(\mathbf{x}\boldsymbol{\beta}_m\right) - \exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)^2\right]\Delta^{-2}\mathbf{x} \\ &= \left[\frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)}{\Delta} - \frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)}{\Delta}\frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)}{\Delta}\right]\mathbf{x} \\ &= \left[\Pr\left(y = m\right) - \Pr\left(y = m\right)\Pr\left(y = m\right)\right]\mathbf{x} \\ &= \Pr\left(y = m\right)\left[1 - \Pr\left(y = m\right)\right]\mathbf{x} \ . \end{aligned} \tag{49}$$

For $m \neq n$,

$$\begin{aligned} \frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_n} &= \left[0 - \exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)\exp\left(\mathbf{x}\boldsymbol{\beta}_n\right)\mathbf{x}\right]\Delta^{-2} \\ &= -\frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_m\right)}{\Delta}\frac{\exp\left(\mathbf{x}\boldsymbol{\beta}_n\right)}{\Delta}\mathbf{x} \\ &= \Pr\left(y = m\right)\Pr\left(y = n\right)\mathbf{x} \ . \end{aligned} \tag{50}$$

For example, for two x's and $m = 1$:

$$\frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_m} = \left[\begin{array}{ccc} p_m\left(1 - p_m\right)x_1 & p_m\left(1 - p_m\right)x_2 & p_m\left(1 - p_m\right) \end{array}\right]' \tag{51}$$

$$\frac{\partial \Pr(y = m|\mathbf{x})}{\partial \boldsymbol{\beta}_{n\neq m}} = \left[\begin{array}{ccc} -p_m p_n x_1 & -p_m p_n x_2 & -p_m p_n \end{array}\right]' \ . \tag{52}$$

6 Poisson and Negative Binomial Regression

In the Poisson regression model,

$$\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \ , \tag{53}$$

so that

$$\begin{aligned} \frac{\partial \mu}{\partial \beta_k} &= \frac{\partial \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}{\partial \beta_k} \\ &= \frac{\partial \exp(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}'\boldsymbol{\beta}}\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \beta_k} \\ &= \exp(\mathbf{x}'\boldsymbol{\beta})x_k \\ &= \mu x_k \end{aligned} \tag{54}$$

Using matrices,

$$\frac{\partial \mu}{\partial \boldsymbol{\beta}} = \frac{\partial \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = \frac{\partial \exp(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}'\boldsymbol{\beta}}\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \mu\mathbf{x} \ . \tag{55}$$

The probability of a given count is

$$\Pr\left(y|\mathbf{x}\right) = \frac{\exp\left(-\mu\right)\mu^{y}}{y!} \ , \tag{56}$$

so we can compute the gradient as:

$$\frac{\partial \exp\left(-\mu\right)\mu^{y}/y!}{\partial \beta_k} = \frac{1}{y!}\frac{\partial \exp\left(-\mu\right)\mu^{y}}{\partial \mu}\frac{\partial \mu}{\partial \beta_k} \ . \tag{57}$$

Since the last term was computed above, we only need to derive

$$\begin{aligned} \frac{\partial \exp\left(-\mu\right)\mu^{y}}{\partial \mu} &= \exp\left(-\mu\right)\frac{\partial \mu^{y}}{\partial \mu} + \mu^{y}\frac{\partial \exp\left(-\mu\right)}{\partial \mu} \\ &= \exp\left(-\mu\right)y\mu^{y-1} - \mu^{y}\exp\left(-\mu\right) \ . \end{aligned} \tag{58}$$

This leads to

$$\begin{aligned} \frac{\partial \Pr\left(y|\mathbf{x}\right)}{\partial \beta_k} &= \frac{1}{y!}\mu\left[\exp\left(-\mu\right)y\mu^{y-1} - \mu^{y}\exp\left(-\mu\right)\right]x_k \\ &= \frac{\exp\left(-\mu\right)y\mu^{y} - \mu^{y+1}\exp\left(-\mu\right)}{y!}x_k \\ &= \frac{y\mu^{y} - \mu^{y+1}}{\exp\left(\mu\right)y!}x_k \ . \end{aligned} \tag{59}$$

Using matrices,

$$\begin{aligned} \frac{\partial \Pr\left(y|\mathbf{x}\right)}{\partial \boldsymbol{\beta}} &= \frac{\exp\left(-\mu\right)y\mu^{y} - \mu^{y+1}\exp\left(-\mu\right)}{y!}\mathbf{x} \\ &= \frac{y\mu^{y} - \mu^{y+1}}{\exp\left(\mu\right)y!}\mathbf{x} \ . \end{aligned} \tag{60}$$

The negative binomial model is specified as

$$\begin{aligned} \mu &= \exp\left(\mathbf{x}'\boldsymbol{\beta} + \varepsilon\right) \\ &= \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)\exp\left(\varepsilon\right) \ , \end{aligned} \tag{61}$$

where ε has a gamma distribution with variance α. The counts have a negative binomial distribution

$$\Pr\left(y_i \mid \mathbf{x}_i\right) = \frac{\Gamma(y_i + \nu)}{y_i!\,\Gamma(\nu)}\left(\frac{\nu}{\nu + \mu_i}\right)^{\nu}\left(\frac{\mu_i}{\nu + \mu_i}\right)^{y_i} \ , \tag{62}$$

where $\nu = \alpha^{-1}$. The derivatives of the log-likelihood are given in Stata Reference, Version 8, page 10. To simplify notation, we define $\tau = \ln \alpha$, $m = 1/\alpha$, $p = 1/(1 + \alpha\mu)$, and $\mu = \exp(\mathbf{x}\boldsymbol{\beta})$. With $\psi(z)$ being the digamma function evaluated at z,

$$\frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}} = p(y - \mu) \tag{63}$$

$$\frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \tau} = -m\left[\frac{\alpha(\mu - y)}{1 + \alpha\mu} - \ln(1 + \alpha\mu) + \psi(y + m) - \psi(m)\right]. \tag{64}$$

Then by the chain rule,

$$\begin{aligned} \frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}} &= \frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \Pr(y|\mathbf{x})}\frac{\partial \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}} \\ &= \Pr(y|\mathbf{x})^{-1}\frac{\partial \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}}, \end{aligned} \tag{65}$$

so that

$$\frac{\partial \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}} = \frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \mathbf{x}\boldsymbol{\beta}} \Pr(y|\mathbf{x}). \tag{66}$$

Similarly for τ,

$$\frac{\partial \Pr(y|\mathbf{x})}{\partial \tau} = \frac{\partial \ln \Pr(y|\mathbf{x})}{\partial \tau} \Pr(y|\mathbf{x}). \tag{67}$$

7 References

Agresti, Alan. 2002. Categorical Data Analysis. 2nd Edition. New York: Wiley.

Greene, William H. 2000. Econometric Analysis, 4th Ed. New York: Prentice Hall.

File: spost_deltaci-22Aug2005.tex - 22Aug2005