Derivation of Integrated Heckman-Conway-Maxwell-Poisson Model

LEONARD KING'ORA THUO[1], DR. JOSEPH OUNO OMONDI[2], DR. BONIFACE KWACH[3]
[1, 3] *Kibabii University-Kenya*
[2] *Masai Mara University-Kenya*

Abstract- ***The main objective of this paper was to illustrate derivation of Heckman-Conway-Maxwell-Poisson 'HeckCOMPoisson' model. HeckCOMPoisson model is an integrated model from Heckman and Conway-Maxwell-Poisson model. HeckCOMPoisson was developed to handle count model with selection. The model perform perfectly in terms of Goodness-of-Fit (GOF) and in prediction of count data with selection and handling under-dispersion and over-dispersion experienced in count data. HeckCOMPoisson distribution is flexible and gives robust models for dispersed counts.***

Indexed Terms- ***COMPoisson, HeckCOMPoisson, Integrated, Heckman***

I. INTRODUCTION

This paper explicitly surveyed basic concepts on Heckman selection model and COMPoisson models then developed an integrated 'HeckCOMPoisson' model.

1.1 Heckman Selection Model

Heckman's 1979 two-step estimator model is a decision model that corrects bias from non-randomly selected samples. Heckman has two separate equations; one focusing on selection into the sample and the second is the main equation linking the covariates of interest to the outcome.

1.1.1 Heckman's Stage One Probit Function

In Heckman's stage one, a Probit function is as shown in the following equation.

$$\Pr(Y_k Observed|\, W_k\alpha) = \Phi\big(h(W_k\alpha)\big) + \varepsilon_{k1} \quad (1)$$

Such that, we only observe the binary outcome given by the probit model when,

$$Y_k^{probit} = (\,y > 0) \quad (2)$$

Therefore,

$$Y_k = \Phi\big(h(W_k\alpha)\big) + \varepsilon_{k1} \quad (3)$$

1.1.2 Heckman Select Equation

The dependent variable is observed only if k is observed in the selection equation. The select equation is as shown in the equation below.

$$Y_k^{select} = \mathrm{X}_k\delta + \varepsilon_{k2} > 0 \quad (4)$$

1.2 COM-Poisson Model

COMPoisson model was introduced by Conway and Maxwell in 1962. COMPoisson distribution consists of an extra parameter, denoted by ν, and which governs the rate of decay of successive ratios of probabilities. COMPoisson distribution's structure allows for a variety of generalizations such as zero-inflated data. COMPoisson is flexible and would fit equi-dispersed, over-dispersed and under-dispersed data.

1.2.1 COMPoisson Probability Mass Function (PMF)

COMPoisson is dependent on probability mass function (PMF) which is formulated as shown in (5) and (6).

$$P(Y = k) = \frac{1}{Z(\lambda, v)}\frac{\lambda^k}{(k!)^v} \qquad k = 0,1,2,\ldots, \quad (5)$$

Where,

$$Z(\lambda, v) = \sum_{n=0}^{\infty}\frac{\lambda^n}{(n!)^v} \qquad v \geq 0 \quad (6)$$

II. INTEGRATED HECKCOMPOISSON MODEL

Unlike the usual Heckman model, integrated HeckCOMPoisson is an appropriate estimating model that fits a COMPoisson regression model into the Heckman model and hence eliminates endogenous sample selection bias. Heckman model is not appropriate for count outcomes because its linear model for the outcome often produces negative

predicted values and does not restrict the predicted values to integers, but the integrated HeckCOMPoisson solves this problem. Integrated HeckCOMPoisson is an integrated model for estimating parameters of a count-data model with endogenous sample selection.

In addition, drawing the correct inference about selection effects depends on collinearity between the inverse Mills ratio and the other predictors in the second stage equation. When the error terms from the selection and the outcome equations are correlated (that is, $\rho \neq 0$), the standard probit techniques yield biased results, Breslow [6]. However, there are few methods to correct the standard errors in the second stage. The two documented methods are manual matrix manipulation and automatic correction using statistical packages. Endogeneity issue is the major limitation in Heckmam model. Endogeneity technically rises from correlation between the variables in the probit equation and select equation. Endogeneity biases parameter estimates. This research study added COMPoisson model in the second stage as an endogeneity correction approach. COMPoisson introduced a parameter denoted by ν that governed the rate of decay of successive ratios of probabilities and was not correlated with variables in the probit model. That is, Corr[probit(x),COMP(x')] = 0

2.1 Integrated HeckCOMPoisson Stage One Probit Function

The integrated HeckCOMPoisson model has two equations, one equation for the count outcome, *y*, and another equation for a binary selection indicator, *z*. The indicator z takes values of 0 or 1. In summary, the count variable Y_i is assumed to have a COMPoisson distribution, conditional on the covariates X_i, with conditional mean given by the equation below.

$$\mathrm{E}(Y_i \mid X_i, \varepsilon_{1i}) = \lambda \frac{\partial \log Z(\lambda,\nu)}{\partial \lambda} \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu} + \varepsilon_{1i} \quad (7)$$

2.2 Integrated HeckCOMPoisson Stage Two Outcome Function

We only observe outcome Y_i for observation *i* if z_i, the selection outcome, which is the binary outcome from a latent-variable model with covariates Θ_i is equal to 1, that is;

$$z_i = \begin{cases} 1, & if\ \Theta_i\gamma + \varepsilon_{2i} > 0 \\ 0, & elsewhere \end{cases} \quad (8)$$

The error terms ε_{1i} from (7) and ε_{2i} from (8) have bivariate normal distribution with zero mean and covariance matrix, $\begin{bmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{bmatrix}$

where σ and ρ have their usual interpretation for the bivariate normal distribution. A nonzero ρ implies that the selected sample is not representative of the whole population and hence inference based on standard Poisson regression using the observed sample is incorrect.

To compute the matrix above, HeckCOMPoisson uses the truncated value of ρ. This enables the estimate of σ to be made consistent with the truncated estimate of ρ and therefore,

$$\sigma' = \beta_i \dot{\rho} \quad (9)$$

Where β_i is the coefficient of the truncated rho.

Both the truncated ρ and the new estimate of σ' are used in all computations to estimate the two-step covariance matrix. The truncated rho lie in the range [-1;1]. If the two-step estimate for ρ is less than -1 then ρ is set to -1 and if the two-step estimate is greater than 1, ρ is set to 1.

In order to retain ρ within the valid limits described above and for numerical stability during optimization, the integrated HeckCOMPoisson estimates the inverse hyperbolic tangent of ρ using the equation below.

$$\operatorname{atanh}\rho = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) \quad (10)$$

In Stata, the Integrated HeckCOMPoisson automatically computes the inverse hyperbolic tangent of ρ. Similarly, the Integrated HeckCOMPoisson does not directly estimate σ, for numerical stability it estimates $ln\,\sigma$. Estimation of ρ and σ in the forms atanh ρ and $ln\ \sigma$ extends the range of these parameters to infinity in both directions. Additionally, ρ and σ were used to compute λ which represented selection effect and was computed as

$$\lambda = \sigma\rho \quad (11)$$

The standard error of λ was computed using the delta method (that is, propagation of error method) as shown in (12) below.

$$Var(\lambda) \approx D\ var\ (\operatorname{atanh}\rho\ ln\sigma)\ \check{D} \quad (12)$$

Where D is the Jacobian of λ with respect to atanh ρ and ln σ

REFERENCES

[1] Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Am. Statist. Ass.*, 85, 565–571