

---

Reading: • Diggle, Heagerty, Liang & Zeger, Chapters 2, & 3  
          ◦ Verbeke & Molenberghs, Chapters 1, 2, & 3

---

## Overdispersion and Relative Efficiency

1. In this exercise we will consider the relative efficiency of count data regression estimators that adopt either the Poisson (scale) variance model, or weighted estimators that adopt the negative binomial variance model. The key issue that we focus on is whether using the correct variance model leads to substantially more efficient regression estimators? First we will evaluate analytical forms for the regression estimators and then verify our calculations using simulation.

A Poisson regression estimator solves the estimating equation:

$$0 = U_0(\beta) = \sum_{i=1}^n \left( \mathbf{X}_i^T \mu_i \right) (1/\mu_i)(Y_i - \mu_i)$$

while an estimator that adopts the negative-binomial variance solves:

$$0 = U_2(\beta) = \sum_{i=1}^n \left( \mathbf{X}_i^T \mu_i \right) \{1/[\mu_i \cdot (1 + \alpha \cdot \mu_i)]\}(Y_i - \mu_i)$$

Let  $\hat{\beta}^{(0)}$  denote the standard Poisson estimator (solution to  $U_0$ ) and let  $\hat{\beta}^{(2)}$  denote the solution to the negative-binomial variance weighted estimator (solution to  $U_2$ ). Let  $n = 200$  denote the number of observations and

$$\mu_i = \exp(1.0 + 1.5 \cdot X_{1,i} + 1.0 \cdot X_{2,i})$$

$$X_{1,i} = (i - n/2)/n$$

$$X_{2,i} = \mathbf{1}(i > 100)$$

(a) First, assume that the negative-binomial variance is the correct form. For  $\alpha = 0.25, 0.50, 1.0$  compare the relative efficiency of  $\hat{\beta}^{(0)}$  to  $\hat{\beta}^{(2)}$  by deriving and computing the variance for each of these regression estimators in this situation. Comment on apparent patterns.

(b) For  $\alpha = 1.0$  compute the relative efficiency for different values of  $\beta$ . Specifically, let  $\beta_1 = 0, 0.5, 1.0$ . Comment on apparent patterns.

(c) Second, assume that the correct model is now the scale model where  $\text{var}(y_i) = \phi \cdot \mu_i$ . For  $\phi = 2.0, 3.0, 4.0$  compare the relative efficiency of  $\hat{\beta}^{(0)}$  to  $\hat{\beta}^{(2)}$  by deriving and computing the variance for each of these regression estimators in this situation. Comment on apparent patterns.

(d) Use simulations to verify the calculations in (a) for  $\alpha = 1$  and in (c) for  $\phi = 4.0$ . To generate the data use unobserved heterogeneity with scaled gamma random effects (ie.  $E(Y_i | \nu_i) = \mu_i \cdot \nu_i$ . (see lecture notes pp. 171–175)

(e)[**optional**] Finally, in our computations in 1(a) we treated  $\alpha$  as if it were fixed and known. In practice we actually estimate  $\alpha$ . Comment on the impact of the estimation of  $\alpha$  on the variance of  $\hat{\beta}^{(2)}$  (either mathematically or via simulation).

## Continuous Correlated Data

2. In this exercise we will generate correlated continuous data with different forms for the correlation (covariance) structure. Let the number of observations per subject be  $n_i = 10$ , evaluated at times  $t_{ij} = j$  for  $j = 1, 2, \dots, 10$ . Use the mean model:

$$E(Y_{ij} | \mathbf{X}_i) = \beta_0 + \beta_1 \cdot t_{ij} \quad (1)$$

For each of the scenarios below generate response vectors,  $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{i10})$  with the specified covariance structure. Generate data for  $m = 25$  subjects for each scenario. Use the parameter value  $\beta = (70, 10)$  for each scenario.

(a) **Random Intercepts:** To introduce correlation we assume that each subject has their own intercept. The complete model is given by equation (1) and:

$$Y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + b_{0,i} + \epsilon_{ij}$$

$$b_{0,i} \sim \mathcal{N}(0, \tau^2)$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where  $b_{0,i}$  and  $\epsilon_{ij}$  are mutually independent.

- Give the general form for the covariance matrix  $\Sigma = \text{cov}(\mathbf{Y}_i)$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $t_i$  for  $m = 25$  using  $\sigma = 20.0, \tau = 5.0$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $t_i$  for  $m = 25$  using  $\sigma = \sqrt{(425/2)}, \tau = \sqrt{(425/2)}$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $t_i$  for  $m = 25$  using  $\sigma = 5.0, \tau = 20.0$ .

(b) **Random Intercepts and Slopes:** To introduce correlation we assume that each subject has their own intercept and their own slope. The complete model is given by equation (1) and:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 \cdot t_{ij} + b_{0,i} + b_{1,i}t_{ij} + \epsilon_{ij} \\ \mathbf{b}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where  $\mathbf{b}_i = (b_{0,i}, b_{1,i})$  and  $\epsilon_{ij}$  are mutually independent. Let  $\sigma = 1$ .

◦ Give the general form for the covariance matrix  $\mathbf{\Sigma} = \text{cov}(\mathbf{Y}_i)$ .

- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 1.0$ ,  $\mathbf{D} = \begin{bmatrix} 100.0 & 0 \\ 0 & 2 \end{bmatrix}$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 1.0$ ,  $\mathbf{D} = \begin{bmatrix} 100 & -2 \\ -2 & 2 \end{bmatrix}$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 1.0$ ,  $\mathbf{D} = \begin{bmatrix} 50 & 0 \\ 0 & 4 \end{bmatrix}$ .

(c) **Serial Correlation:** To introduce correlation we assume that each subject has their own “process” that is serially correlated. The complete model is given by equation (1) and:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 \cdot t_{ij} + W_i(t_{ij}) + \epsilon_{ij} \\ \mathbf{W}_i &\sim \mathcal{N}(\mathbf{0}, \mathbf{G}) \\ \text{var}[W_i(t_{ij})] &= \tau^2 \quad \text{diagonal elements of } \mathbf{G} \\ \text{cov}[W_i(t_{ij}), W_i(t_{ik})] &= \tau^2 \rho^{|t_{ij} - t_{ik}|} \quad \text{off-diagonal elements of } \mathbf{G} \end{aligned}$$

where  $\mathbf{W}_i$  and  $\epsilon_{ij}$  are mutually independent.

◦ Give the general form for the covariance matrix  $\mathbf{\Sigma} = \text{cov}(\mathbf{Y}_i)$ .

- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 10.0, \tau = 20.0, \rho = 0.7$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 10.0, \tau = 20.0, \rho = 0.9$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 5, \tau = 20.0, \rho = 0.9$ .

(d) **Non-normal random effects:** A standard assumption in linear mixed models is the normality of random effects. In this section we will generate non-normal random effects and plot the data. Is the

non-normality apparent in the data?

We return to the assumption that each subject has their own intercept. The complete model is now given by:

$$Y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + b_{0,i} + \epsilon_{ij}$$

$$b_{0,i} \sim F_b$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

where  $b_{0,i}$  and  $\epsilon_{ij}$  are mutually independent.

- If the random effects distribution has mean 0 and variance  $\tau^2$  then give the general form for the covariance matrix  $\Sigma = \text{cov}(\mathbf{Y}_i)$ .
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 5.0, \tau = 10.0$ , and exponential random effects (centered to have mean 0:  $b_{0,i} = \tau \cdot (z_i - 1)$ , where  $z_i \sim \text{exponential}$ ).
- Generate  $\mathbf{Y}_i$  and plot (lines) versus  $\mathbf{t}_i$  for  $m = 25$  using  $\sigma = 5.0, \tau = 10.0$ , and  $\chi^2(4)$  random effects (centered and scaled to have mean 0 and standard deviation  $\tau$ :  $b_{0,i} = \tau \cdot (z_i - 4)/\sqrt{8}$ , where  $z_i \sim \chi^2(4)$ ).
- If you had 10 observations on each of 100 subjects then can you suggest a method for evaluating assumptions regarding the distribution of random intercepts and random slopes? What might you do to see if the standard assumption of normality appears satisfied?