A generalized confidence interval for the mean response in log-regression models

Miguel Fonseca Thomas Mathew João Tiago Mexia

CompStat'2010
Paris, France

Model

- $\mathbf{y} = \left((\log(w_1), \ldots, \log(w_n)\right)'$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\tau} + \mathbf{e}$$

- $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices of dimensions $n \times p$ and $n \times s$
- $\boldsymbol{\tau} \sim N(\mathbf{0}, \sigma_\tau^2 I_s)$
- $\mathbf{e} \sim N(\mathbf{0}, \sigma_\tau^2 I_n)$

$$\mathbf{y} \sim N(\mathbf{X}\beta, \sigma_\tau^2 \mathbf{Z}\mathbf{Z}' + \sigma_e^2 I_n)$$

Problem

$$Y_0 \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma_\tau^2 \mathbf{z}_0'\mathbf{z}_0 + \sigma_e^2).$$

The mean of W_0 is then given by

$$E(W_0) = E\left(\exp(Y_0)\right) = \exp\left(\mathbf{x}_0'\beta, \frac{\sigma_\tau^2 \mathbf{z}_0'\mathbf{z}_0 + \sigma_e^2}{2}\right).$$

Thus the interval estimation of $E(W_0)$ is equivalent to the interval estimation of

$$\theta = \mathbf{x}_0'\beta + \frac{\sigma_\tau^2 \mathbf{z}_0'\mathbf{z}_0 + \sigma_e^2}{2}.$$

Generalized Pivotal Quantities

$$G(\mathbf{y}, \mathbf{y}_{\rm obs}; \theta, \eta)$$

(i) given the observed value $\mathbf{y}_{\rm obs}$, the distribution of $G(\mathbf{y}, \mathbf{y}_{\rm obs}; \theta, \eta)$ is free of unknown parameters,
(ii) the observed value of $G(\mathbf{y}, \mathbf{y}_{\rm obs}; \theta, \eta)$ is free of the nuisance parameter η.

Generalized Pivotal Quantities

$$G(\mathbf{y}, \mathbf{y}_{\mathrm{obs}}; \theta, \eta)$$

(i) given the observed value $\mathbf{y}_{\mathrm{obs}}$, the distribution of $G(\mathbf{y}, \mathbf{y}_{\mathrm{obs}}; \theta, \eta)$ is free of unknown parameters,
(ii) the observed value of $G(\mathbf{y}, \mathbf{y}_{\mathrm{obs}}; \theta, \eta)$ is free of the nuisance parameter η.

When the above conditions hold,

$$G_{1-\alpha} = \{\theta : \ G(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{obs}}; \theta, \eta) \leq G_{1-\alpha}\}$$

is a $100(1-\alpha)\%$ one-sided generalized confidence interval for θ. A two-sided interval can be similarly defined.

Generalized Pivotal Quantity

$$G_\theta = G_{\mathbf{x}_0'\beta} + \frac{G_{\sigma_\tau^2} \times \mathbf{z}_0'\mathbf{z}_0 + G_{\sigma_e^2}}{2}$$

- $G_{\mathbf{x}_0'\beta} \longrightarrow \mathbf{x}_0'\beta$
- $G_{\sigma_\tau^2} \longrightarrow \sigma_\tau^2$
- $G_{\sigma_e^2} \longrightarrow \sigma_e^2$

Parameters

- σ_e^2

$$G_{\sigma_e^2} = \frac{\sigma_e^2}{SS_e} ss_e = \frac{ss_e}{U_e^2}$$

- $U_e^2 = \frac{SS_e}{\sigma_e^2} \sim \chi^2_{n-r}$
- $SS_e = \mathbf{y}' \left[\mathbf{I}_n - P_{(\mathbf{X},\mathbf{Z})} \right] \mathbf{y}$
- $P_{(\mathbf{X},\mathbf{Z})} = (\mathbf{X},\mathbf{Z}) \left[(\mathbf{X},\mathbf{Z})'(\mathbf{X},\mathbf{Z}) \right]^{-} (\mathbf{X},\mathbf{Z})' = \mathbf{Q}\mathbf{Q}'$

Parameters

- σ_τ^2
  - $\mathbf{V}_G = G_{\sigma_\tau^2}\mathbf{Q}'\mathbf{Z}\mathbf{Z}'\mathbf{Q} + G_{\sigma_e^2}\mathbf{I}_r$
  - $\mathbf{Q}\mathbf{Q}' = \mathbf{P}_{(\mathbf{X},\mathbf{Z})},\ \mathbf{Q}'\mathbf{Q} = \mathbf{I_r}$
  - $U_0^2 = \mathbf{y}'_{\mathrm{obs}}\mathbf{Q}\left[\mathbf{V}_G^{-1} - \mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X}(\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\right]\mathbf{Q}'\mathbf{y}_{\mathrm{obs}}$
    - $U_0^2 \sim \chi^2_{r-p}$
- $\mathbf{x}_0'\beta$

$$
\begin{aligned}
G_{\mathbf{x}_0'\boldsymbol{\beta}} &= \mathbf{x}_0'(\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{y}_{\mathrm{obs}} \\
&\quad - \frac{\mathbf{x}_0'\hat{\boldsymbol{\beta}}_{\mathbf{V}} - \mathbf{x}_0'\boldsymbol{\beta}}{\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{Q}\mathbf{V}^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{x}_0}} \times \sqrt{\left[\mathbf{x}_0'(\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{x}_0\right]_+} \\
&= \mathbf{x}_0'(\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{y}_{\mathrm{obs}} - Z\sqrt{\left[\mathbf{x}_0'(\mathbf{X}'\mathbf{Q}\mathbf{V}_G^{-1}\mathbf{Q}'\mathbf{X})^{-1}\mathbf{x}_0\right]_+}
\end{aligned}
$$

  - $Z \sim \mathcal{N}(0,1)$

Application

The data were obtained from 34 licensed rural nursing facilities and 18 urban nursing facilities in the State of New Mexico.

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \tau_i + e_{ij}$$

- x_1 number of beds
- x_2 medical in-patient days
- τ_i Rural/non-rural
- W total patient-care revenue

$\mathbf{x}_0 = (1, 0.8368, 1.8476)' \rightarrow$ 90% confidence interval: $[9.3241, 9.5419]$

Simulations

$$\beta = (1,1,1)', (2,2,2)', (3,3,3)'$$
$$\sigma_\tau^2 = 0.1, 0.25, 0.5, 1, 2, 5$$
$$\sigma_e^2 = 1$$

1000 runs were performed, each with a pseudo-sample of size 1000. The confidence was 90%.

Table: Coverage Probability

$\beta \backslash \sigma_1^2$	0.1	0.25	0.5	1	2	5
$(1,1,1)'$	0.953	0.905	0.888	0.835	0.837	0.619
$(2,2,2)'$	0.952	0.922	0.877	0.836	0.822	0.630
$(3,3,3)'$	0.943	0.919	0.878	0.829	0.831	0.633

Simulations

$$\beta = (1,1,1)', (2,2,2)', (3,3,3)'$$
$$\sigma_\tau^2 = 0.1, 0.25, 0.5, 1, 2, 5$$
$$\sigma_e^2 = 1$$

1000 runs were performed, each with a pseudo-sample of size 1000. The confidence was 90%.

Table: Average Length

$\beta \backslash \sigma_1^2$	0.1	0.25	0.5	1	2	5
$(1,1,1)'$	2.306	2.584	2.891	3.160	3.487	3.733
$(2,2,2)'$	2.309	2.610	2.844	3.179	3.441	3.741
$(3,3,3)'$	2.338	2.622	2.847	3.146	3.480	3.765

Restricted Model

Let $\mathbf{P_X} = \mathbf{A_X}\mathbf{A}'_\mathbf{X}$ be the orthogonal projection matrix (OPM) on $\mathrm{R}(\mathbf{X})$ an $\mathbf{I} - \mathbf{P} = \mathbf{A}^o_\mathbf{X}\mathbf{A}^{o\prime}_\mathbf{X}$ the OPM on the orthogonal complement of $\mathrm{R}(\mathbf{X})$. Then

$$\mathbf{y}_0 = \mathbf{A}^{o}_\mathbf{X}{}'\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_\tau \mathbf{A}^{o}_\mathbf{X}{}'\mathbf{Z}\mathbf{Z}'\mathbf{A}^o_\mathbf{X} + \sigma^2_e \mathbf{I}\right)$$

Matrices $\sigma^2_1 \mathbf{A}^{o}_\mathbf{X}{}'\mathbf{Z}\mathbf{Z}'\mathbf{A}^o_\mathbf{X}$ and $\sigma^2_e \mathbf{I}$ span a commutative Jordan algebra (CJA) with principal basis

$$\{\mathbf{Q}_1, \ldots, \mathbf{Q}_w, \mathbf{Q}_{w+1}\},$$

where $\mathbf{Q}_{w+1} = \mathbf{I} - \sum_{j=1}^{w} \mathbf{Q}_j$. Then,

$$\sigma^2_1 \mathbf{A}^{o}_\mathbf{X}{}'\mathbf{Z}\mathbf{Z}'\mathbf{A}^o_\mathbf{X} + \sigma^2_e \mathbf{I} = \sum_{j=1}^{w+1} c_j \mathbf{Q}_j,$$

with $c_j = \lambda_j \sigma^2_\tau + \sigma^2_e$ for $i = 1, \ldots, w$ and c_{w+1}.

Generalized Pivotal Quantities

- σ_τ^2 and σ_e^2
  - $S_j = \mathbf{y}_0'\mathbf{Q}_j\mathbf{y}_0 \sim c_j\chi^2_{g_j}$
  - GPQ – $\dot{c}_j = \frac{S_j}{U_j}$
  - GPQs – $(\dot{\sigma}_\tau^2, \dot{\sigma}_e^2)' = \mathbf{F}^+\dot{\mathbf{c}}$,
  - $U_j \sim \chi^2_{g_j}$
- $\mathbf{x}_0'\boldsymbol{\beta}$
  - $\dot{\mathbf{V}} = \dot{\sigma}_\tau^2\mathbf{Z}\mathbf{Z}' + \dot{\sigma}_e^2\mathbf{I}$

$$G_{\mathbf{x}_0'\boldsymbol{\beta}} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} - Z\sqrt{\left(\mathbf{x}_0'\left(\mathbf{X}'\dot{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{x}_0\right)_+}$$

  - $Z \sim \mathcal{N}(0,1)$

Application

The data were obtained from 34 licensed rural nursing facilities and 18 urban nursing facilities in the State of New Mexico.

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \tau_i + e_{ij}$$

- x_1 number of beds
- x_2 medical in-patient days
- τ_i Rural/non-rural
- W total patient-care revenue

$\mathbf{x}_0 = (1, 0.8368, 1.8476)' \rightarrow$ 90% confidence interval: $[9.2527, 9.9517]$

Simulations

$$\boldsymbol{\beta} = (1,1,1)', (2,2,2)', (3,3,3)'$$
$$\sigma_\tau^2 = 0.1, 0.25, 0.5, 1, 2, 5$$
$$\sigma_e^2 = 1$$

1000 runs were performed, each with a pseudo-sample of size 1000. The confidence was 90%.

Table: Coverage Probability

$\boldsymbol{\beta} \backslash \sigma_1^2$	0.1	0.25	0.5	1	2	5
$(1,1,1)'$	0.909	0.893	0.887	0.887	0.887	0.898
$(2,2,2)'$	0.913	0.895	0.902	0.875	0.895	0.905
$(3,3,3)'$	0.907	0.913	0.884	0.877	0.885	0.886

Simulations

$$\beta = (1,1,1)', (2,2,2)', (3,3,3)'$$
$$\sigma_\tau^2 = 0.1, 0.25, 0.5, 1, 2, 5$$
$$\sigma_e^2 = 1$$

1000 runs were performed, each with a pseudo-sample of size 1000. The confidence was 90%.

Table: Average Length

$\beta \backslash \sigma_1^2$	0.1	0.25	0.5	1	2	5
$(1,1,1)'$	23.970	44.0767	75.893	153.4390	295.4439	727.9971
$(2,2,2)'$	29.256	47.990	84.800	149.483	304.313	693.920
$(3,3,3)'$	25.384	45.100	78.443	161.064	357.139	730.132

Random Model

$$Y_{ijk} = \mu + \tau_i + \beta_{j(i)} + e_{k(ij)},$$

with

- μ and τ_i, $i = 1, ..., a$ are fixed,
- $\beta_{j(i)} \sim N(0, \sigma_\beta^2)$, $i = 1, ..., a$, $j = 1, ..., b_i$,
- $e_{k(ij)} \sim N(0, \sigma_e^2)$,
- all random variables are independent.

Upper Tolerance Limit

- Distributions of the quantiles
  - $T_{3p} \sim N(\mu, \sigma_\tau^2 + \sigma_\beta^2 + \sigma_e^2)$
  - $T_{4p} \sim N(\mu, \sigma_\tau^2 + \sigma_\beta^2)$
- γ confidence bound for the p quantile of the unknown distribution
- Confidence bound for the parametric functions:
  - $\mu + z_p\sqrt{\sigma_\tau^2 + \sigma_\beta^2 + \sigma_e^2}$
  - $\mu + z_p\sqrt{\sigma_\tau^2 + \sigma_\beta^2}$,

  where z_p is the $100p\%$ quantile of a standard normal distribution

Statistics

$$\bar{Y}_{\dots} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}$$

$$SS_{\tau} = \frac{1}{bn}\sum_{i=1}^{a}(Y_{i..} - \bar{Y}_{\dots})^2$$

$$SS_{\beta} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij.} - \bar{Y}_{i..})^2$$

$$SS_{e} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij.})^2$$

Distribution of Statistics

$$
\begin{aligned}
Z &= \frac{\sqrt{abn}(\bar{Y}_{\dots} - \mu)}{\sqrt{bn\sigma_\tau^2 + n\sigma_\beta^2 + \sigma_e^2}} \sim N(0,1) \\
U_\tau &= \frac{SS_\tau}{bn\sigma_\tau^2 + n\sigma_\beta^2 + \sigma_e^2} \sim \chi^2_{a-1} \\
U_\beta &= \frac{SS_\beta}{n\sigma_\beta^2 + \sigma_e^2} \sim \chi^2_{a(b-1)} \\
U_e &= \frac{SS_e}{\sigma_e^2} \sim \chi^2_{ab(n-1)}
\end{aligned}
$$

Upper Tolerance Limit – $N(\mu, \sigma_\tau^2 + \sigma_\beta^2 + \sigma_e^2)$

- Generalized Pivot Statistic

$$
\begin{aligned}
T_{3p} &= \bar{y}_{\dots} - \frac{\sqrt{abn}(\bar{Y}_{\dots} - \mu)}{\sqrt{SS_\tau}} \times \frac{\sqrt{ss_\tau}}{abn} \\
&\quad + z_p \left[\frac{\sigma_e^2}{SS_e} \times ss_e + \frac{1}{n} \left(\frac{n\sigma_\beta^2 + \sigma_e^2}{SS_\beta} \times ss_\beta - \frac{\sigma_e^2}{SS_e} \times ss_e \right) \right. \\
&\quad \left. + \frac{1}{bn} \left(\frac{bn\sigma_\tau^2 + n\sigma_\beta^2 + \sigma_e^2}{SS_\tau} \times ss_\tau - \frac{n\sigma_\beta^2 + \sigma_e^2}{SS_\beta} \times ss_\beta \right) \right]^{1/2} \\
&= \bar{y}_{\dots} - \frac{Z}{\sqrt{U_\tau}} \times \frac{\sqrt{ss_\tau}}{\sqrt{abn}} \\
&\quad + \frac{z_p}{\sqrt{bn}} \left[\frac{ss_\tau}{U_\tau} + (b-1)\frac{ss_\beta}{U_\beta} + b(n-1)\frac{ss_e}{U_e} \right]^{1/2}
\end{aligned}
$$

- $100\gamma\%$ upper bound for T_{3p} obtained through Monte Carlo

Upper Tolerance Limit – $N(\mu, \sigma_\tau^2 + \sigma_\beta^2)$

- Generalized Pivot Statistic

$$
\begin{aligned}
T_{4p} &= \bar{y}_{\ldots} - \frac{\sqrt{abn}(\bar{Y}_{\ldots} - \mu)}{\sqrt{SS_\tau}} \times \frac{\sqrt{ss_\tau}}{abn} \\
&+ z_p \left[\frac{1}{n} \left(\frac{n\sigma_\beta^2 + \sigma_e^2}{SS_\beta} \times ss_\beta - \frac{\sigma_e^2}{SS_e} \times ss_e \right) \right. \\
&+ \left. \frac{1}{bn} \left(\frac{bn\sigma_\tau^2 + n\sigma_\beta^2 + \sigma_e^2}{SS_\tau} \times ss_\tau - \frac{n\sigma_\beta^2 + \sigma_e^2}{SS_\beta} \times ss_\beta \right) \right]_+^{1/2} \\
&= \bar{y}_{\ldots} - \frac{Z}{\sqrt{U_\tau}} \times \frac{\sqrt{ss_\tau}}{\sqrt{abn}} \\
&+ \frac{z_p}{\sqrt{bn}} \left[\frac{ss_\tau}{U_\tau} + (b-1)\frac{ss_\beta}{U_\beta} - b\frac{ss_e}{U_e} \right]_+^{1/2}
\end{aligned}
$$

- $100\gamma\%$ upper bound for T_{4p} obtained through Monte Carlo

Numerical Results

(0.90,0.95)**upper tolerance limit for** $N(\mu, \sigma_\tau^2 + \sigma_\beta^2 + \sigma_e^2)$

$a = 5, b = 5$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9738	0.9703	0.9653	0.9594	0.9523

$a = 5, b = 20$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9694	0.9660	0.9611	0.9568	0.9518

$a = 20, b = 5$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9716	0.9683	0.9668	0.9647	0.9581

$a = 20, b = 20$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9634	0.9611	0.9598	0.9592	0.9573

(0.90,0.95)**upper tolerance limit for** $N(\mu, \sigma_\tau^2 + \sigma_\beta^2)$

$a = 5, b = 5$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9766	0.9723	0.9668	0.9593	0.9521

$a = 5, b = 20$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9715	0.9661	0.9600	0.9555	0.9512

$a = 20, b = 5$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9730	0.9717	0.9672	0.9624	0.9571

$a = 20, b = 20$

ρ	0.1	0.3	0.5	0.7	0.9
Monte Carlo	0.9642	0.9628	0.9602	0.9577	0.9553

Breeding Experiment I

$$y_{1..} = 2.67,\ y_{2..} = 2.53,\ y_{3..} = 2.63,\ y_{4..} = 2.47,\ y_{5..} = 2.57,$$
$$ss_\beta = 0.56,\ ss_e = 0.39$$

(0.9, 0.95) upper tolerance limit for $N(\mu_i, \sigma_\beta^2 + \sigma_e^2)$

i	Monte Carlo method	Approximation
1	3.52	3.51
2	3.37	3.38
3	3.49	3.48
4	3.33	3.32
5	3.43	3.42

Breeding Experiment II

(0.9, 0.95) upper tolerance limit for $N(\mu_i, \sigma^2_\beta)$

i	Monte Carlo method	Approximation
1	3.46	3.47
2	3.32	3.34
3	3.42	3.44
4	3.26	3.28
5	3.36	3.38

(0.9, 0.95) upper tolerance limits

Distribution	Monte Carlo method
$N(\sigma^2_\tau + \sigma^2_\beta + \sigma^2_e)$	3.08
$N(\sigma^2_\tau + \sigma^2_\beta)$	2.99

Final Remarks

- Method 1 produces short intervals, but with low coverage probabilities
- Method 2 produces intervals with good coverage probabilities, but with very large length
- $\frac{1}{\chi_1^2}$ random variables have no moments
- Other methodologies...

References

19 March 2007 Fonseca, M., Mathew, T., Mexia, J.T. and Zmyślony, R. (2007). Tolerance intervals in a two-way nested model with mixed or random effects, *Statistics*, 41:4, 289 - 300

Tian, L. and Wu, J. (2007). Inferences on the mean response in a log-regression model: The generalized variable approach. *Statistics in Medicine*, 26, 5180-5188

THANK YOU!