

## Sample Questions: Log-Normal Regression

STA312 Fall 2023. Copyright information is at the end of the last page.

1. Let the continuous random variable  $T$  have median  $m$ . Let  $Y = g(T)$ , where  $g(x)$  is an increasing function. Show that the median of  $Y$  is  $g(m)$ . This is why the median of a log-normal is  $e^\mu$ .

$$\begin{aligned} \frac{1}{2} &= P(T \leq m) = P(g(T) \leq g(m)) \\ &= P(Y \leq g(m)) \end{aligned}$$

$\uparrow$   
 median of  $Y$

2. Show that the expected value of a log-normal is  $e^{\mu + \frac{1}{2}\sigma^2}$ . Hint: the moment-generating function of a normal random variable is  $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ .

$$X \sim LN(\mu, \sigma^2) \text{ means } Y = \log(X) \sim N(\mu, \sigma^2)$$

$$E(X) = E(e^Y) = E(e^{Yt}) \text{ with } t=1$$

"  
 $M_Y(t)$

$$M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \text{ at } t=1,$$

$$= e^{\mu + \frac{1}{2}\sigma^2} = E(X) \text{ Easier than}$$

$$\int_{-\infty}^{\infty} e^{yt} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy$$

$w = g(u)$   
 $\downarrow$   
 $\int w f_w(w) dw$

3. Write the log-normal regression model in multiplicative form.

$$y_i = x_i^T \beta + \varepsilon_i^*, \quad \varepsilon_i^* \sim N(0, \sigma^2) \quad \log(t_i) = \mu_i$$

$$t_i = e^{\mu_i} = e^{x_i^T \beta + \varepsilon_i^*} = e^{x_i^T \beta} e^{\varepsilon_i^*} = e^{x_i^T \beta} \varepsilon_i$$

$\varepsilon_i \sim LN(0, \sigma^2)$

4. For a log-normal regression model, show that if  $x_{i,k}$  is increased by  $c$  units,  $E(t_i)$  is multiplied by  $e^{c\beta_k}$ .

$$E(t_i) = e^{\mu_i + \frac{1}{2}\sigma^2} = e^{x_i^T \beta + \frac{1}{2}\sigma^2}$$

$$\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k (x_k + c) + \dots + \beta_{p-1} x_{p-1} + \frac{1}{2}\sigma^2}}{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \dots + \beta_{p-1} x_{p-1} + \frac{1}{2}\sigma^2}} = \frac{e^{\beta_k (x_k + c)}}{e^{\beta_k x_k}} = \frac{e^{\beta_k x_k} e^{c\beta_k}}{e^{\beta_k x_k}} = e^{c\beta_k}$$

5. If  $x_{i,k}$  is increased by one unit, the median of  $t_i$  is multiplied by  $e^{\beta_k}$ .

$$\text{Median} = e^{x^T \beta}$$

6. If  $x_{i,k}$  is increased by one unit, the value of  $t_i$  is multiplied by  $e^{\beta_k}$ .

$$T = e^Y, Y \sim N(x^T \beta, \sigma^2)$$

7. Write the hazard function of a log-normal regression model in terms of  $\Phi(x)$ , the cumulative distribution function of a standard normal. Is this a proportional hazards model?

$$h(t) = \frac{f_T(t)}{S_T(t)}$$

$$\left( \begin{aligned} f_T(t) &= \frac{d}{dt} P(T \leq t) = \frac{d}{dt} P(e^Y \leq t) \\ &= \frac{d}{dt} P(Y \leq \log t) = \frac{d}{dt} F_Y(\log t) \\ &= f_Y(\log t) \cdot \frac{1}{t} \end{aligned} \right)$$

$$= \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - F_T(t)} = \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - P(T \leq t)} = \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - P(e^Y \leq t)}$$

$$= \frac{f_T(\log t) \cdot \frac{1}{t}}{1 - P(Y \leq \log t)} = \frac{f_T(\log t) \cdot \frac{1}{t}}{1 - P\left(\frac{Y - x^T \beta}{\sigma} \leq \frac{\log t - x^T \beta}{\sigma}\right)}$$

$Z \sim N(0,1)$

$$= \frac{f_T(\log t) \cdot \frac{1}{t}}{1 - \Phi\left(\frac{\log t - x^T \beta}{\sigma}\right)}$$

$$= \frac{\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (\log t - x^T \beta)^2} \cdot \frac{1}{t}}{1 - \Phi\left(\frac{\log t - x^T \beta}{\sigma}\right)}$$

Not proportional hazards

8. Show that in general, if  $\hat{\theta}_n \sim N_k(\theta, V_n)$  and  $\mathbf{a}$  is a non-zero  $k \times 1$  vector of constants, then  $W = \mathbf{a}^T \hat{\theta}_n \sim N(\mathbf{a}^T \theta, \mathbf{a}^T V_n \mathbf{a})$ .

$$W_n = \mathbf{a}^T \hat{\theta}_n = g(\hat{\theta}_n), \text{ where}$$

$$g(\theta) = a_1 \theta_1 + a_2 \theta_2 + \dots + a_k \theta_k$$

$$\dot{g}(\theta) = (a_1, a_2, \dots, a_k) = \mathbf{a}^T$$

So by delta,

$$W_n = g(\hat{\theta}_n) \sim N(g(\theta), \dot{g}(\theta) V_n \dot{g}(\theta)^T)$$

$$= N(\mathbf{a}^T \theta, \mathbf{a}^T V_n \mathbf{a})$$

---

9. What is the parameter vector  $\theta$  for a log-normal regression model with  $p-1$  explanatory variables?

$$\theta = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2)$$

10. For a log-normal regression model, let  $\mathbf{x}_{n+1}$  be a  $p \times 1$  vector of explanatory variable values, maybe starting with a 1 for the intercept. A new observation (log failure time) could be written  $y_{n+1} = \mathbf{x}_{n+1}^T \beta + \epsilon_{n+1}$ , where  $\epsilon_{n+1} \sim N(0, \sigma^2)$ , and  $\epsilon_{n+1}$  is independent of  $\epsilon_1, \dots, \epsilon_n$ . It is natural to predict the value of  $y_{n+1}$  with the estimated expected value, so  $\hat{y}_{n+1} = \mathbf{x}_{n+1}^T \hat{\beta}$ .

Let  $\mathbf{V}_n$  denote the  $(p+1) \times (p+1)$  asymptotic covariance matrix of the parameter vector. What is the asymptotic distribution of  $\hat{y}_{n+1}$ ?

$$\text{Let } \mathbf{a}^T = (1, x_1, x_2, \dots, x_{p-1}, 0)$$

$$\mathbf{a}^T \theta = \mathbf{x}_{n+1}^T \beta$$

By problem 8,

$$\hat{y}_{n+1} = \mathbf{x}_{n+1}^T \hat{\beta} \approx N(\mathbf{x}_{n+1}^T \beta, \mathbf{a}^T \mathbf{V}_n \mathbf{a})$$

$$\mathbf{x}_{n+1}^T C_n \mathbf{x}_{n+1}$$

where  $C_n$  is the asymptotic covariance matrix of  $\hat{\beta}$ .

11. What is the asymptotic distribution of the error in prediction  $y_{n+1} - \hat{y}_{n+1}$ ? Justify your answer; include calculation of the expected value and variance.

Have  $\hat{y}_{n+1} \sim N(x_{n+1}^T \beta, x_{n+1}^T C_n x_{n+1})$

$y_{n+1}$  &  $\hat{y}_{n+1}$  are independent. So

$$\text{Var}(y_{n+1} - \hat{y}_{n+1}) = \text{Var}(y_{n+1}) + \text{Var}(\hat{y}_{n+1})$$

$y_{n+1}$  is normal,  $\hat{y}_{n+1}$  is asymptotically normal  
so difference is asymptotically normal.

$$y_{n+1} - \hat{y}_{n+1} \sim N(0, \sigma^2 + x_{n+1}^T C_n x_{n+1})$$

$$\begin{aligned} E(y_{n+1} - \hat{y}_{n+1}) &= E(y_{n+1}) - E(\hat{y}_{n+1}) = x_{n+1}^T \beta - E(x_{n+1}^T \hat{\beta}) \\ &= x_{n+1}^T \beta - x_{n+1}^T \beta = 0 \end{aligned}$$

12. What is the standard error of  $y_{n+1} - \hat{y}_{n+1}$ . Remember, a standard error is an *estimated* standard deviation, something that can be computed from sample data.

$$se = \sqrt{\sigma^2 + x_{n+1}^T \hat{C}_n x_{n+1}}$$

13. Dividing  $y_{n+1} - \hat{y}_{n+1}$  by its standard error, obtain a  $Z$  statistic. What is the asymptotic distribution of  $Z$ ?

$$Z = \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\hat{\sigma}^2 + x_{n+1}^T \hat{C}_n x_{n+1}}} \sim \mathcal{N}(0, 1)$$

14. Use the  $Z$  statistic to obtain a 95% prediction interval for  $y_{n+1}$ .

$$\begin{aligned} 0.95 &\approx P(-1.96 < Z < 1.96) \\ &= P\left(-1.96 < \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\text{stuff}}} < 1.96\right) \\ &: \end{aligned}$$

$$= P\left(\hat{y}_{n+1} - 1.96 \sqrt{\hat{\sigma}^2 + x_{n+1}^T \hat{C}_n x_{n+1}} < y_{n+1} < \hat{y}_{n+1} + 1.96 \sqrt{\hat{\sigma}^2 + x_{n+1}^T \hat{C}_n x_{n+1}}\right)$$

This assignment was prepared by Jerry Brunner, Department of Mathematical and Computational Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The L<sup>A</sup>T<sub>E</sub>X source code is available from the course website:

<http://www.utstat.toronto.edu/brunner/oldclass/312f23>