Sample Questions: Log-Normal Regression

STA312 Spring 2019. 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^{μ}.

$\frac{1}{2} = P(T \leq m) = P(g(T) \leq g(m))$

$= P(Y \leq g(m))$

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 $e^{\mu t+\frac{1}{2}\sigma^2 t^2}$.

$y = \log(t) \iff t = e^y \quad E(T) = E(e^Y)$

$= M_Y(t=1)$ Bec $M_Y(t) = E(e^{Yt})$

$= e^{\mu t+\frac{1}{2}\sigma^2 t^2} = e^{\mu+\frac{1}{2}\sigma^2}$

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

$t_i = e^{x_i^T \beta} \varepsilon_i$, where $\varepsilon_i \sim \text{lognormal}(0, \sigma^2)$

$\log t_i = \log(t_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \log \varepsilon_i$, $\log \varepsilon_i \sim N(0, \sigma)$

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

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

$$\text{ratio} = \frac{e^{\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k (x_{ik} + c) + \cdots + \beta_{p-1} x_{i,p-1}} e^{\frac{1}{2}\sigma^2}}{e^{\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{ik} + \cdots + \beta_{p-1} x_{i,p-1}} e^{\frac{1}{2}\sigma^2}}$$

$$= \frac{e^{\beta_k x_{ik} + \beta_k c}}{e^{\beta_k x_{ik}}} = e^{c\beta_k}$$

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

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

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?

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

$$f_T(t) = \frac{d}{dt} F_T(t) = \frac{d}{dt} P(T \le t) = \frac{d}{dt} P(e^Y \le t)$$

$$= \frac{d}{dt} P(Y \le \log(t))$$

$$= \frac{d}{dt} F_Y(\log t) = f_Y(\log t) \cdot \frac{1}{t}$$

$$= \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - F_T(t)} = \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - P(T \le t)}$$

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

$$= \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - P\left(\frac{Y-\mu}{\sigma} \le \frac{\log t - \mu}{\sigma}\right)} = \frac{f_Y(\log t) \cdot \frac{1}{t}}{1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)}$$

$Z \sim N(0,1)$

$\mu_i = x_i^T \beta$

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

NO

8. Show that in general, if $\widehat{\boldsymbol{\theta}}_n \dot{\sim} N_k(\boldsymbol{\theta}, \mathbf{V}_n)$ and $\mathbf{a}$ is a non-zero $k \times 1$ vector of constants, then $W_n = \mathbf{a}^\top \widehat{\boldsymbol{\theta}}_n \dot{\sim} N\left(\mathbf{a}^\top \boldsymbol{\theta},\, \mathbf{a}^\top \mathbf{V}_n \mathbf{a}\right)$.

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

$g(\Theta) = a_1\Theta_1 + a_2\Theta_2 + \cdots + a_k\Theta_k$

$\dot{g}(\Theta) = \left(\frac{\partial g}{\partial \Theta_1}, \frac{\partial g}{\partial \Theta_2}, \cdots, \frac{\partial g}{\partial \Theta_k}\right)$

$= (a_1, a_2 \cdots, a_k) = a^T$

So by the delta method,

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

$= N(a^T\Theta, a^T V_n a)$

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

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

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

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

Let $a^\top = (1, x_1, x_2, \cdots x_{p-1}, 0)$

$x_{n+1}^\top$

$a^\top \theta = x_{n+1}^\top \beta$, by Problem 8

$\hat{y}_{n+1} = x_{n+1}^\top \hat{\beta} \dot{\sim} N(x_{n+1}^\top \beta, a^\top V_n a)$

$x_{n+1}^\top C_n x_{n+1}$

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

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

$y_{n+1} - \hat{y}_{n+1}$ is asymptotically normal, because linear combinations of normals are normal.

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

$$\mathrm{Var}(y_{n+1} - \hat{y}_{n+1}) \overset{ind}{=} \mathrm{Var}(y_{n+1}) + \mathrm{Var}(\hat{y}_{n+1})$$
$$= \sigma^2 + a^T V_n a$$

So $y_{n+1} - \hat{y}_{n+1} \dot{\sim} N(0, \sigma^2 + a^T V_n a)$

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

$$\sqrt{\hat{\sigma}^2 + a^T \hat{V}_n a}$$

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

$$Z = \frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\hat{\sigma}^2 + a^T \hat{V}_n a}}$$

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

$$0.95 \approx P(-1.96 < Z < 1.96)$$

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

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

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