Log-Normal Regression[1]

STA312 Fall 2023

[1]See last slide for copyright information.

Overview

1. The Model
2. Prediction Intervals

The Log-Normal Distribution

- Failure time $t \sim$ Log-Normal(μ, σ^2) means $y = \log(t) \sim N(\mu, \sigma^2)$.
- $y = \log(t) \Leftrightarrow t = e^y$.
- The log-normal density is

$$f(t|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \left\{ \frac{(\log(t) - \mu)^2}{2\sigma^2} \right\} \frac{1}{t}$$

- $P(T > 0) = 1$, right-skewed.
- Median $= e^{\mu}$, expected value is $e^{\mu + \frac{1}{2}\sigma^2}$.

Regression

- In normal regression, $\mu_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}$ e
- So just take logs of the failure times and do normal regression.
- Lots of things are familiar, except for censoring.
- Because of censoring, formulas like $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ do not apply.
- F and t distributions do not apply.
- Everything is large-sample maximum likelihood.

Interpretation in Terms of Failure Time

- People think in terms of time, not log time.
- Don't talk about log failure time, except to statisticians.
- $\mu_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} = \mathbf{x}_i^\top \boldsymbol{\beta}$.
- The quantity μ_i has meaning on the time scale. The median failure time for a log-normal is e^{μ_i}. Mean failure time is $e^{\mu_i + \frac{1}{2}\sigma^2}$.
- Anything that makes $\mathbf{x}_i^\top \boldsymbol{\beta}$ larger makes average failure time larger.
- Ideas like positive and negative relationship, "controlling for," etc. carry over directly.

Prediction intervals

- You have a good log-normal regression analysis of a set of data.
- Want to predict the value of a future observation, given the explanatory variable values.
- That is, you have $\mathbf{x}_{n+1}$ and you want to predict t_{n+1}. This is a very practical goal.
- A natural prediction would be the estimated median for those $\mathbf{x}_{n+1}$ values: $e^{\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}}$.

- Estimates and predictions are more valuable when they come with a margin of error, or interval of likely values.

Prediction versus Estimation

- In statistics, we estimate parameters – or functions of parameters.
- These are fixed constants.
- With increasing sample size, the confidence interval shrinks to zero.
- Prediction tries to “estimate” the value of a random variable.
- There is uncertainty about the average value, and further uncertainty that comes from variation of random variables around the average.
- Prediction intervals are always wider than confidence intervals.

Prediction for the Normal Model

- Prediction intervals for normal regression are straightforward.
- In survival analysis, the distinction between confidence intervals and prediction intervals is largely ignored.
- This is probably because the distribution theory for prediction intervals is so hard.
- Except for log-normal regression ...
- So the following is "new," and based on the derivation for ordinary regression.

Derivation of the Prediction Interval

Details will be covered in the sample problems

- Get a point prediction and interval for $y_{n+1} = \log(t_{n+1})$, and then take the exponential function.

$$\begin{aligned} 0.95 &\approx P(A < y_{n+1} < B) \\ &= P(e^A < e^{y_{n+1}} < e^B) \\ &= P(e^A < t_{n+1} < e^B) \end{aligned}$$

- $\widehat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, \mathbf{C}_n)$.
- $\widehat{y}_{n+1} = \mathbf{x}_{n+1}^\top \widehat{\boldsymbol{\beta}} \dot{\sim} N\left(\mathbf{x}_{n+1}^\top \boldsymbol{\beta},\, \mathbf{x}_{n+1}^\top \mathbf{C}_n \mathbf{x}_{n+1}\right)$.
- $y_{n+1} \sim N\left(\mathbf{x}_{n+1}^\top \boldsymbol{\beta}, \sigma^2\right)$.
- y_{n+1} and $\widehat{y}_{n+1}$ are independent.
- $y_{n+1} - \widehat{y}_{n+1} \dot{\sim} N\left(0,\, \sigma^2 + \mathbf{x}_{n+1}^\top \mathbf{C}_n \mathbf{x}_{n+1}\right)$.

Derivation of the Prediction Interval Continued

- $y_{n+1} - \widehat{y}_{n+1} \dot{\sim} N\left(0,\, \sigma^2 + \mathbf{x}_{n+1}^\top \mathbf{C}_n \mathbf{x}_{n+1}\right)$
- $Z = \frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{\widehat{\sigma}^2 + \mathbf{x}_{n+1}^\top \widehat{\mathbf{C}}_n \mathbf{x}_{n+1}}} \dot{\sim} N(0,1)$

$$
\begin{aligned}
0.95 &\approx P(-1.96 < Z < 1.96) \\
&= P\left(-1.96 < \frac{y_{n+1} - \widehat{y}_{n+1}}{\sqrt{\widehat{\sigma}^2 + \mathbf{x}_{n+1}^\top \widehat{\mathbf{C}}_n \mathbf{x}_{n+1}}} < 1.96\right)
\end{aligned}
$$

- Isolate y_{n+1}.
- Prediction interval is $\widehat{y}_{n+1} \pm 1.96\sqrt{\widehat{\sigma}^2 + \mathbf{x}_{n+1}^\top \widehat{\mathbf{C}}_n \mathbf{x}_{n+1}}$.
- Exponential function of the endpoints gives prediction interval for t_{n+1}.

Derivation of the Prediction Interval Concluded

Prediction interval for t_{n+1} is from

$$\exp\left(\mathbf{x}_{n+1}^\top\widehat{\boldsymbol{\beta}} - 1.96\sqrt{\widehat{\sigma}^2 + \mathbf{x}_{n+1}^\top\widehat{\mathbf{C}}_n\mathbf{x}_{n+1}}\right)$$

to

$$\exp\left(\mathbf{x}_{n+1}^\top\widehat{\boldsymbol{\beta}} + 1.96\sqrt{\widehat{\sigma}^2 + \mathbf{x}_{n+1}^\top\widehat{\mathbf{C}}_n\mathbf{x}_{n+1}}\right)$$

where $\mathbf{C}_n$ is the estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$, obtained from the inverse of the Hessian.

Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistics, 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 LaTeX source code is available from the course website:

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