STA 312s19 Assignment Nine[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 8 on March 18th. The R part may be handed in as part of the quiz. **Bring hard copy of your printout to the quiz**. Do not write anything on your printout in advance except possibly your name and student number.

1. Let T be a log-normal random variable with parameters zero and one. That is, the log of T is standard normal. Let $Y = e^{\mu}T^{\sigma}$, where $\sigma > 0$. Show that the distribution of Y is log-normal, and give the parameters.

2. Prove that the median of a log-normal(μ, σ^2) is e^{μ}.

3. Show that the expected value of a log-normal(μ, σ^2) is $e^{\mu+\frac{1}{2}\sigma^2}$.

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

5. For a log-normal regression model, show that if $x_{i,k}$ is increased by c units, the median is multiplied by $e^{c\beta_k}$.

6. 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?

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

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

9. 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 $y_{n+1} = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon_{n+1}$, 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 $\widehat{y}_{n+1}$?

10. 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.

[1] 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 LaTeX source code is available from the course website: `http://www.utstat.toronto.edu/~brunner/oldclass/312s19`

11. 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.

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

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

14. For a particular form of cancer, the standard treatment is a combination of chemotherapy and radiation therapy. Both chemotherapy and radiation have serious side effects. Some patients may be so weakened by the treatment that they die from other things (such as infections) that are apparently unrelated to the cancer.

    Volunteer patents who were considering no treatment at all were randomly assigned to one of three experimental conditions. They received either Chemotherapy only, Radiation only, or Both treatments. The response variable is survival time, which in some cases will be right-censored. Age is an important predictor of survival, and is used as a covariate.

    (a) Write the (multiplicative) log-normal regression equation, denoting the length of time between diagnosis and death (call it survival time) for patient i by t_i. Denote age by x_i. There should be *no interactions* in the model. You do not need to say how your dummy variables are defined. You will do that in the next part. Complete the equation below.

    $t_i =$

    (b) In the table below, make columns showing how your dummy variables are defined. In the last column, write the expected survival time, using the notation of your model from Question 14a above. If *symbols* for your dummy variables appear in the last column, the answer is wrong.

		Expected Survival Time
Chemotherapy		
Radiation		
Both		

    (c) In the notation of your model, what is the expected survival time for a 25-year-old patient receiving both radiation and chemotherapy?

(d) You want to produce a large-sample confidence interval for expected survival time, for a 25-year-old patient receiving both radiation and chemotherapy. You need to use the delta method.

i. What is the parameter vector $\boldsymbol{\theta}$? Give a general answer for your model.

ii. What is $\dot{g}(\boldsymbol{\theta})$?

(e) For a 60-year-old patient receiving radiation only, the median survival time is __________ times as great as the expected survival time for a 60-year-old receiving both radiation and chemotherapy. Answer in terms of the Greek letters from your model.

(f) For a 47-year-old patient receiving radiation only, the median survival time is __________ times as great as the expected survival time for a 47-year-old receiving chemotherapy only. Answer in terms of the Greek letters from your model.

(g) You want to know whether, controlling for age, experimental treatment (Chemotherapy, Radiation, or Both) has any effect on average survival time. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(h) That last question could be answered with either a large-sample likelihood ratio test, or a Wald test.

i. Suppose you decided on a likelihood ratio test. Write the multiplicative Weibull regression equation for the restricted model.

$t_i =$

ii. Suppose you decided on a Wald test. Write the null hypothesis $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{0}$ in terms of specific matrices.

(i) You want to know whether it is better for patients to get both radiation and chemotherapy, or just radiation. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(j) You want to know whether it is better for patients to get both radiation and chemotherapy, or just chemotherapy. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(k) You want to know whether it is better for patients to get just radiation or just chemotherapy. What is the null hypothesis? Answer in terms of the Greek letters from your model.

15. The `survival` package has a built-in data set on patients with advanced lung cancer. Type `help(cancer)` for details. This is the same data set you used in Assignment 8.

    (a) Last week, we used Weibull regression and arrived at a model with just ECOG rating (from 0=good to 5=dead) and sex. Fit a log-normal model with all the explanatory variables. Notice that only sex and ECOG rating are significant. Test all the other variables simultaneously, controlling for sex and ECOG rating. Use a Wald test. Does it appear safe to drop all these variables?

    (b) Now fit a model with just sex and ECOG rating (so the answer to the last question must have been Yes), and display the `summary`.

        i. Controlling for ECOG rating, median survival time for females is estimated to be __________ times as great as median survival time for males. This is something you could do either with R or with a calculator.

        ii. Predict the survival time in days for a new female patient with an ECOG rating of one. This number is the same as the estimated median.

        iii. Give a 95% prediction interval for the survival time (in days) for a new female patient with an ECOG rating of one. Your answer is a pair of numbers. My upper prediction limit is around 8.7 years.

        iv. Compare the 95% confidence interval for median survival time. Your answer is another pair of numbers.

Please bring your printout to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.