STA 312f23 Assignment Nine[1]

The paper and pencil part of this assignment is not to be handed in. It is practice for Quiz 9 on November 17th. 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 $e^{\mu+\frac{1}{2}\sigma^2}$. Hint: If $Y \sim N(\mu, \sigma^2)$, the moment-generating function of Y is $E(e^{Yt}) = e^{\mu t+\frac{1}{2}\sigma^2 t^2}$.

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, $E(t_i)$ is multiplied by $e^{c\beta_k}$.

6. Write the Survival function of a log-normal regression model in terms of $\Phi(x)$, the cumulative distribution function of a standard normal.

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) Let $\mathbf{V}_n$ denote the $(p+1) \times (p+1)$ asymptotic covariance matrix of the parameter vector, and let $\mathbf{C}_n$ be the upper $p \times p$ portion of $\mathbf{V}_n$, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$. What is the asymptotic distribution of $\widehat{y}_{n+1}$?

   (b) Derive an approximate $(1-\alpha)100\%$ confidence interval for $\mathbf{x}_{n+1}^\top \boldsymbol{\beta}$. It's "approximate" because it's based on the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

[1] This assignment was prepared by Jerry Brunner, Department of Mathematical and Computational Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/312f23

10. Again 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 log failure time could be written $y_{n+1} = \mathbf{x}_{n+1}^\top \boldsymbol{\beta} + \epsilon_{n+1}$, 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 $\widehat{y}_{n+1} = \mathbf{x}_{n+1}^\top \widehat{\boldsymbol{\beta}}$.

    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.

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. In a study of religion and marriage, participants were married male-female couples, and both partners had the same religious affiliation. The study was further limited to just three common religions (called A, B and C), and a None category. The response variable was elapsed time between marriage and divorce. The response values for participants who were still married at the end of the study were censored, as were data for those who died during the course of the study. The explanatory variables were family income, and dummy variables for Religion. For purposes of the data analysis, None is a religion.

    (a) Write the (multiplicative) log-normal regression equation, denoting the length of marriage (a kind of survival time) for couple i by t_i. Denote income by x_i, in thousands of dollars per year. 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 median length of marriage, 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.

		Median Length of Marriage
Religion A		
Religion B		
Religion C		
None		

(c) In the notation of your model, what is the median length of marriage for couples from Religion B, with an annual family income of $75,000?

(d) You want to produce a large-sample *confidence interval* (not prediction interval) for the expected (not median) length of marriage. You need to use the delta method. Give general answers to the questions below for your model.

i. What is the parameter vector $\boldsymbol{\theta}$?

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

(e) The median length of marriage for a couple from Religion A making $120,000 per year is __________ times as great as the median length of marriage for a couple with no religion making $120,000 per year. Answer in terms of the Greek letters from your model.

(f) The expected length of marriage for a couple from Religion B making $72,000 per year is __________ times as great as the expected length of marriage for a couple from Religion C making $72,000 per year. Answer in terms of the Greek letters from your model.

(g) You want to know whether, controlling for income, Religious affiliation (including None) is related to average length of marriage. 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 Log-normal 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, controlling for income, average length of marriage is greater for Religion A, or No Religion. What is the null hypothesis? Answer in terms of the Greek letters from your model.

(j) You want to know whether, controlling for income, average length of marriage is greater for Religion A, or Religion C. What is the null hypothesis? Answer in terms of the Greek letters from your model.

15. The file `ColonCancer.data.txt` is a subset of the `colon` data set from the survival package. A sample of advanced colon cancer patients had surgery that removed all detectable cancer. Then, patients were randomly assigned to one of three drug treatments:

- `Obs`: Just observed, without any drug.
- `Lev`: Levamisole, a low-toxicity compound previously used to treat worm infestations in animals
- `Lev+5FU`: A combination of levamisole and 5-FU, a "moderately" toxic chemotherapy agent.

The variables we will use for this question are

- `rx`: Drug treatment group
- `sex`: 0 = Female, 1 = Male
- `age`: Age in years
- `nodes`: Number of lymph nodes affected
- `status`: 0 = Right censored, 1 = Uncensored
- `time`: Time until censoring or recurrence of the cancer

(a) Take a look at the data. Are there any missing values? What proportion of the data are censored?

(b) Fit a log-normal regression model in which the reference category for treatment group is `Obs`. Look at `summary`. The output includes a chi-squared test with five degrees of freedom. What is it telling you?

(c) Holding age, sex and number of affected nodes constant, the median time to recurrence is estimated to be __________ times as great for patients in the combination Levamisole and 5-FU condition, compared to patients receiving no drug treatment.

(d) You want to know, controlling for experimental treatment, age and number of lymph nodes affected, whether males and females have different median times to recurrence.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the test statistic (z or chi-squared). The answer is a number from your printout.
iii. What is the p-value? The answer is a number from your printout.
iv. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
v. In plain, non-statistical language, what do you conclude?

(e) You want to know, allowing for experimental treatment, age and sex, whether the number of lymph nodes affected is related to time until recurrence.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the test statistic (z or chi-squared). The answer is a number from your printout.
iii. What is the p-value? The answer is a number from your printout.
iv. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
v. In plain, non-statistical language, what do you conclude?

(f) You want to know whether, correcting for age, sex and number of lymph nodes affected, experimental treatment had any effect on recurrence time.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the likelihood ratio test statistic? What is the p-value? Do you reject H_0 at $\alpha = 0.05$?
iii. What is the value of the Wald test statistic? What is the p-value? Do you reject H_0 at $\alpha = 0.05$?
iv. In plain, non-statistical language, what do you conclude from these tests?

(g) You want to know whether, controlling for age, sex and number of lymph nodes affected, median recurrence time was different for patients receiving Levamisole alone versus patients receiving both Levamisole and 5-FU.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the test statistic (z or chi-squared). The answer is a number from your printout.
iii. What is the p-value? The answer is a number from your printout.
iv. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
v. In plain, non-statistical language, what do you conclude?

(h) Controlling for age, sex and number of lymph nodes affected, you want to compare expected time to recurrence for (1) patients receiving Levamisole alone and (2) patients receiving no drug treatment.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the test statistic (z or chi-squared). The answer is a number from your printout.
iii. What is the p-value? The answer is a number from your printout.
iv. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
v. In plain, non-statistical language, what do you conclude?

(i) Controlling for age, sex and number of lymph nodes affected, you want to compare median time to recurrence for (1) patients receiving the combination of Levamisole and 5-FU, and (2) patients receiving no drug treatment.

i. Using the order of variables in your fitted model, what is the null hypothesis?
ii. What is the value of the test statistic (z or chi-squared). The answer is a number from your printout.
iii. What is the p-value? The answer is a number from your printout.
iv. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.
v. In plain, non-statistical language, what do you conclude?

(j) Summarizing the results from earlier questions, is there any evidence that the de-worming drug Levamisole helped slow down recurrence of colon cancer?

(k) Since age and sex were not significantly related to recurrence time, we will drop them from the model, and base predictions on a model with just drug treatment and number of affected lymph nodes. Using a prediction based on median recurrence time, give an estimated recurrence time and a 95% prediction interval for a patient in the Levamisole-only condition, with 6 affected lymph nodes. Display the standard error as well as the point estimate and the lower and upper prediction limits. Give the answer in years. How would you advise a doctor to express the upper limit to a patient?