Name Jerry

Student Number

STA 312 f2023 Quiz 9

1. In Question 1~~6~~5 of Assignment 9, you analyzed the `ColonCancer` data. Time in days until recurrence of the cancer is the response variable. You began with a full model in which the explanatory variables were drug treatment condition (including Observation only), sex, age and number of lymph nodes affected. You want to know whether, taking into account sex, number of lymph nodes affected and drug treatment, the patient's age is related to time until recurrence.

   (a) (1 point) Using the order of variables in your fitted model, what is the null hypothesis?

   $H_0: \beta_3 = 0$

   (b) (1 point) What is the value of the test statistic (z or chi-squared)? The answer is a number from your printout. Write the number in the space below.

   $z = 0.87$

   (c) (1 point) What is the p-value? The answer is a number from your printout. Write the number in the space below.

   $p = 0.39$

   (d) (1 point) Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

   No

   (e) (2 points) In plain, non-statistical language, what do you conclude?

   Allowing for sex, number of lymph nodes affected and drug treatment, there is no evidence that age is related to time until recurrence. (Minus one for "not related")

   On your printout, circle the requested numbers, and write "Question 1" beside them.

2. At the end of Question 15 from Assignment 9, you calculated a predicted recurrence time and a 95% prediction interval for a patient in the Levamisole-only condition, with 6 affected lymph nodes.

   (a) (1 point) Give the predicted recurrence time *in years*. Write the number in the space below.

   2.94

   (b) (3 points) Give the prediction interval, *in years*. Write the two numbers in the space below.

   (0.081, 106.47)

   On your printout, circle the numbers, and write "Question 2" beside them. It is okay if the numbers on the printout are in days.

Please attach your printout, showing all input and output required to answer the questions above. **Make sure your name and student number are written on the printout.**

```
R version 4.2.3 (2023-03-15) -- "Shortstop Beagle"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.79 (8198) x86_64-apple-darwin17.0]

[Workspace restored from /Users/brunner/.RData]
[History restored from /Users/brunner/.Rapp.history]

> # Assignment 9, Question 15
>
> rm(list=ls());  options(scipen=999)
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival) # Do this every time
>
> # (a)
> ColonCancer = read.table("https://www.utstat.toronto.edu/brunner/data/legal/
ColonCancer.data.txt")
>
> head(ColonCancer); dim(ColonCancer)
        rx sex age nodes status time
2  Lev+5FU   1  43     5      1  968
4  Lev+5FU   1  63     1      0 3087
6      Obs   0  71     7      1  542
8  Lev+5FU   0  66     6      1  245
10     Obs   1  69    22      1  523
12 Lev+5FU   0  57     9      1  904
[1] 929   6
> # Make Obs the reference category for rx
> ColonCancer = within(ColonCancer,{
+ rx = factor(rx)
+ contrasts(rx) = contr.treatment(3,base=3)
+ colnames(contrasts(rx)) = c("Lev","Lev+5FU")
+ })
> summary(ColonCancer)
       rx           sex              age            nodes            status      
 Lev    :310   Min.   :0.000   Min.   :18.00   Min.   : 0.00   Min.   :0.0000  
 Lev+5FU:304   1st Qu.:0.000   1st Qu.:53.00   1st Qu.: 1.00   1st Qu.:0.0000  
 Obs    :315   Median :1.000   Median :61.00   Median : 2.00   Median :1.0000  
```

```
              Mean   :0.521   Mean   :59.75   Mean   : 3.66   Mean   :0.5038
              3rd Qu.:1.000   3rd Qu.:69.00   3rd Qu.: 5.00   3rd Qu.:1.0000
              Max.   :1.000   Max.   :85.00   Max.   :33.00   Max.   :1.0000
                                              NA's   :18
      time
 Min.   :   8
 1st Qu.: 370
 Median :1548
 Mean   :1405
 3rd Qu.:2289
 Max.   :3329

>
>
> # (b)
> # Full model
> full = survreg(Surv(time,status) ~ rx + sex + age + nodes,
+                dist="lognormal", data=ColonCancer)
> summary(full)

Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + nodes,
    data = ColonCancer, dist = "lognormal")
               Value Std. Error     z                  p
(Intercept)  7.47451    0.37240 20.07 < 0.0000000000000002
rxLev        0.03024    0.15991  0.19                 0.85
rxLev+5FU    0.75633    0.16838  4.49            0.0000071
sex          0.18520    0.13517  1.37                 0.17
age          0.00487    0.00563  0.87                 0.39
nodes       -0.15335    0.01804 -8.50 < 0.0000000000000002
Log(scale)   0.60148    0.03711 16.21 < 0.0000000000000002

Scale= 1.82

Log Normal distribution
Loglik(model)= -3933   Loglik(intercept only)= -3983.6
    Chisq= 101.24 on 5 degrees of freedom, p= 0.000000000000000000029
Number of Newton-Raphson Iterations: 3
n=911 (18 observations deleted due to missingness)

> # Something is going on. At least one variable matters.
>
> # (c)
> exp(0.75633) # Comparing Lev+5FU to nothing.
[1] 2.130443
>
> # (d) See z-test.
> # (e) See z-test.
>
> # (f)
> # LR test of rx
> norx = update(full, . ~ . - rx)
> # summary(norx) # n is correct
> anova(norx,full)
```

Question 1

```
                   Terms Resid. Df    -2*LL Test Df Deviance       Pr(>Chi)
1      sex + age + nodes       906 7891.115      NA       NA             NA
2 rx + sex + age + nodes       904 7865.945    =  2 25.16918 0.000003424373
>
> # Wald test of rx
> betahat = coef(full); betahat
 (Intercept)      rxLev   rxLev+5FU         sex         age       nodes
 7.474508107 0.030236057 0.756331612 0.185201833 0.004873333 -0.153352960
> V = vcov(full)[(1:6),(1:6)] # Omitting last row and col for log scale.
> round(V,5)
            (Intercept)    rxLev rxLev+5FU      sex      age    nodes
(Intercept)     0.13868 -0.01169  -0.01343 -0.00879 -0.00190 -0.00191
rxLev          -0.01169  0.02557   0.01251 -0.00060 -0.00001  0.00003
rxLev+5FU      -0.01343  0.01251   0.02835  0.00157  0.00001  0.00001
sex            -0.00879 -0.00060   0.00157  0.01827 -0.00002  0.00005
age            -0.00190 -0.00001   0.00001 -0.00002  0.00003  0.00001
nodes          -0.00191  0.00003   0.00001  0.00005  0.00001  0.00033
> Lrx = rbind(c(0,1,0,0,0,0),
+             c(0,0,1,0,0,0))
> colnames(Lrx) = names(coef(full))
> Lrx
     (Intercept) rxLev rxLev+5FU sex age nodes
[1,]           0     1         0   0   0     0
[2,]           0     0         1   0   0     0
> source("http://www.utstat.toronto.edu/brunner/Rfunctions/Wtest.txt")
> Wtest(Lrx,betahat,V)
             W              df          p-value
24.767674038180  2.000000000000  0.000004185698
>
> # (g) Levamisole alone versus patients receiving both Levamisole and 5-FU
> # Custom test.
>
> Lqf = cbind(0,1,-1,0,0,0)
> Wtest(Lqf,betahat,V)
            W             df         p-value
18.23931306793  1.00000000000  0.00001948159
>
>
> # (h) See z-test.
> # (i) See z-test.
>
> # (j) Nope.
>
> # (k) Prediction interval based on a model with just treatment and number of nodes.
>
> model2 = survreg(Surv(time,status) ~ rx + nodes, dist="lognormal", data=ColonCancer)
> summary(model2)

Call:
survreg(formula = Surv(time, status) ~ rx + nodes, data = ColonCancer,
    dist = "lognormal")
             Value Std. Error     z                  p
(Intercept) 7.8722     0.1385 56.86 < 0.0000000000000002
rxLev       0.0382     0.1601  0.24                0.81
```

```
rxLev+5FU    0.7409    0.1682  4.41            0.000011
nodes       -0.1556    0.0180 -8.64 < 0.0000000000000002
Log(scale)   0.6032    0.0371 16.25 < 0.0000000000000002

Scale= 1.83

Log Normal distribution
Loglik(model)= -3934.3   Loglik(intercept only)= -3983.6
    Chisq= 98.56 on 3 degrees of freedom, p= 0.000000000000000000000032
Number of Newton-Raphson Iterations: 3
n=911 (18 observations deleted due to missingness)

>
> new = data.frame(rx="Lev", nodes=6); new
   rx nodes
1 Lev     6
> pred = predict(model2,newdata=new,type='linear',se=TRUE) ; pred
$fit
       1
6.976977

$se.fit
        1
0.1222231

> yhat = pred$fit
> t_hat= exp(yhat)
> t_hat # Prediction = estimated median number of days
       1
1071.674
>
> # Prediction interval
> sigmasqhat = model2$scale^2
> se = sqrt(sigmasqhat+pred$se^2); se
       1
1.832001
> L = yhat - 1.96*se; U = yhat + 1.96*se
> lower95 = exp(L); upper95 = exp(U)
> predint = c(t_hat,lower95,upper95)
> names(predint) = c('t-hat','lower95','upper95')
> predint
      t-hat      lower95      upper95
 1071.67380     29.55505  38859.17061
> predint/365
       t-hat       lower95       upper95
  2.93609260  0.08097274  106.46348112
>
>
>
>
```

Question 2