Assignment 9

(1) $T \sim LN(0,1) \iff Z = \log T \sim N(0,1)$

$$Y = e^{\mu} T^{\sigma} = e^{\mu}(e^{z})^{\sigma} = e^{\mu} e^{\sigma z}$$

$$= \exp\{\sigma Z + \mu\} = \exp\{X\} \text{ where}$$

$$X = \sigma Z + \mu \sim N(\mu, \sigma^2)$$

Now $\log e^{X} = X \sim N(\mu, \sigma^2)$, so

$$e^{X} = e^{\mu} T^{\sigma} \sim LN(\mu, \sigma^2)$$

OR one could work with densities, but this is easier.

(2) Let $T \sim LN(\mu, \sigma^2)$. m is the median of T iff $\frac{1}{2} = P(T \le m) = P(\log T \le \log m)$

$$= P(X \le \log m) \text{ where } X \sim N(\mu, \sigma^2)$$

The median of X is μ, so

$$\mu = \log m \iff m = e^{\mu}$$

(3) $T \sim LN(\mu, \sigma^2)$, so

$\log T = X \sim N(\mu, \sigma^2)$ $\quad \log T = X \iff T = e^X$

$E(T) = E(e^X) = E(e^{Xt}) = M_X(t)\Big|_{t=1}$

$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$,

so

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

✓

(4) $t_i = e^{x_i^T \beta} \varepsilon_i^{\sigma}$ where $\varepsilon_i \sim LN(0,1)$

Take log to get normal regression model.

(5)
$$\frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k (x_k + c) + \cdots + \beta_{p-1} x_{p-1} + \frac{1}{2}\sigma^2}}{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_{p-1} x_{p-1} + \frac{1}{2}\sigma^2}}$$

$$= \frac{\cancel{e^{\beta_0}} \cancel{e^{\beta_1 x_1}} \cdots e^{\beta_k (x_k + c)} \cdots \cancel{e^{\beta_{p-1} x_{p-1}}} \cancel{e^{\frac{1}{2}\sigma^2}}}{\cancel{e^{\beta_0}} \cancel{e^{\beta_1 x_1}} \cdots e^{\beta_k x_k} \cdots \cancel{e^{\beta_{p-1} x_{p-1}}} \cancel{e^{\frac{1}{2}\sigma^2}}}$$

$$= \frac{\cancel{e^{\beta_k x_k}} e^{c\beta_k}}{\cancel{e^{\beta_k x_k}}} = e^{c\beta_k}$$

(6) For $t > 0$

$$S(t) = P(T > t) = 1 - F_T(t)$$

$$= 1 - P(T \le t) = 1 - P(e^X \le t)$$

where $X \sim N(x^T\beta, \sigma^2)$

$$= 1 - P(X \le \log t) = 1 - P\left(\frac{X - x^T\beta}{\sigma} \le \frac{\log t - x^T\beta}{\sigma}\right)$$

$$= 1 - P\left(Z \le \frac{\log t - x^T\beta}{\sigma}\right) \text{ where } Z \sim N(0,1)$$

$$= 1 - \Phi\left(\frac{\log t - x^T\beta}{\sigma}\right)$$

(7) Using the delta method,

$$g(\theta) = a_1\theta_1 + a_2\theta_2 + \cdots + a_k\theta_k, = a^T\theta$$

$$\dot{g}(\theta) = (a_1, a_2, \ldots, a_k) = a^T, \text{ so}$$

by delta,

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

(8) $\theta = (\beta_0, \beta_1, \ldots \beta_{p-1}, \sigma)$

(9) (a) $\widehat{x_{n+1}^T\beta} = x_{n+1}^T\hat{\beta}_n = \hat{y}_{n+1} \,\dot{\sim}\, N(x_{n+1}^T\beta, x_{n+1}^T C_n x_{n+1})$

where $\hat{\beta}_n \,\dot{\sim}\, N_k(\beta, C_n)$

(b)

so $$Z = \frac{\hat{y}_{n+1} - x_{n+1}^T\beta}{\sqrt{x_{n+1}^T\hat{C}_n x_{n+1}}} \,\dot{\sim}\, N(0,1), \text{ and}$$

$$1-\alpha \approx P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{\hat{y}_{n+1} - x^T\beta}{\sqrt{}} < z_{\alpha/2}\right)$$

$= \ldots$

$$= P\left\{ x_{n+1}^T\hat{\beta} - z_{\alpha/2}\sqrt{x_{n+1}^T C_n x_{n+1}} < x_{n+1}^T\beta < x_{n+1}^T\hat{\beta} + z_{\alpha/2}\sqrt{x_{n+1}^T C_n x_{n+1}} \right\}$$

~~10 From Q9, $\hat{y}_{n+1} \sim N(x^T_{n+1}\beta, x^T_{n+1} C_n x_{n+1})$~~

10 $E(y_{n+1} - \hat{y}_{n+1}) = E(y_{n+1}) - E(\hat{y}_{n+1}) \approx x^T_{n+1}\beta - x^T_{n+1}\beta = 0$

$Var(y_{n+1} - \hat{y}_{n+1}) \overset{ind}{=} Var(y_{n+1}) + Var(\hat{y}_{n+1})$

$\approx \sigma^2 + x^T_{n+1} C_n x_{n+1}$

Linear combination of normals is normal, so

$y_{n+1} - \hat{y}_{n+1} \overset{\cdot}{\sim} N(0, \sigma^2 + x^T_{n+1} C_n x_{n+1})$

11 SE of $y_{n+1} - \hat{y}_{n+1} = \sqrt{\hat{\sigma}^2 + x^T_{n+1} \hat{C}_n x_{n+1}}$

12 $Z_n = \dfrac{y_{n+1} - \hat{y}_{n+1}}{se} \overset{\cdot}{\sim} N(0,1)$

13 $0.95 \approx P(-1.96 < Z_n < 1.96)$

$= P(\hat{y}_{n+1} - 1.96 * se < y_{n+1} < \hat{y}_{n+1} + 1.96 * se)$

14 (a) $Y_i = e^{\beta_0+\beta_1 x+\beta_2 d_1+\beta_3 d_2+\beta_4 d_3} \times \varepsilon_i^{\sigma}$, $\varepsilon_i \sim LN(0,1)$

(b)

RELIGION	d_1	d_2	d_3	Median Length of Marriage
A	1	0	0	$e^{\beta_0+\beta_1 x}\ e^{\beta_2}$
B	0	1	0	$e^{\beta_0+\beta_1 x}\ e^{\beta_3}$
C	0	0	1	$e^{\beta_0+\beta_1 x}\ e^{\beta_4}$
None	0	0	0	$e^{\beta_0+\beta_1 x}$

(c) $e^{\beta_0+\beta_3+75\beta_1}$

(d)(i) $\theta = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4\ \sigma)$ I am parameterising by σ, not σ^2

(i) Expected value $= g(\theta) = e^{\beta_0+\beta_1 x+\beta_2 d_1+\beta_3 d_2+\beta_4 d_3+\frac{1}{2}\sigma^2}$
$= e^{x^T\beta+\frac{1}{2}\sigma^2}$

$\dot{g}(\theta)$

$= e^{x^T\beta}(1, x, d_1, d_2, d_3, \sigma) \cdot e^{\frac{1}{2}\sigma^2}$

(e) e^{β_2}

(f) $e^{\beta_3-\beta_4}$

(g) $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$

(1~~4~~h) (i) Reduced model is

$$Y_i = e^{\beta_0 + \beta_1 x_i} \times \varepsilon_i^{\sigma}$$

(ii)

$$\underbrace{\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \sigma \end{pmatrix}}_{\theta} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}}_{h}$$

(1~~4~~i) $H_0: \beta_2 = 0$

(j) $H_0: \beta_2 = \beta_3$

```
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)
```

```
                   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= 0.00000000000000000000032
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
>
>
>
>
```