ISYE 6420: Homework 6

aelhabr3

1. Cancer of Tongue.

Instructions

Sickle-Santanello et al (1988) provide data on 80 males diagnosed with cancer of the tongue. Data are provided in the file tongue.csv|dat|xlsx. The variables in the dataset are as follows:

- *Tumor DNA profile (1 - aneuploid tumor, 2 - diploid tumor);*
- *Time to death or on-study time (in weeks); and*
- *Censoring indicator (0=observed, 1=censored)*

Fit the regression with tumor profile as covariate. What is the 95% Credible Set for the slope β_1*?*

Response

```
library(tidyverse)
```

Below is the model code.

```
model {
  for(i in 1:n) {
    times[i] ~ dweib(v, lambda[i]) I(censor[i],)
    lambda[i] <- exp(beta0 + beta1 * type[i])
    S[i] <- exp(-lambda[i] * pow(times[i], v));
    f[i]  <- lambda[i] * v * pow(times[i], v-1) * S[i]
    h[i] <- f[i] / S[i]
    index[i] <- i
  }
  beta0 ~ dnorm(0.0, 0.0001)
  beta1 ~ dnorm(0.0, 0.0001)
  v ~ dexp(0.001)

  median0 <- pow(log(2) * exp(-beta0), 1 / v)
  median1 <- pow(log(2) * exp(-beta0 - beta1), 1 / v)
}

# data
list(
  n = 80,
  type = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),
  times = c(1, 3, 3, 4, 10, 13, 13, 16, 16, 24, 26, 27, 28, 30, 30, 32, 41, 51, 65, 67, 70, 72, 73, 77, 91,
93, 96, 100, 104, 157, 167, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, 1, 3, 4, 5, 5, 8, 12, 13, 18, 23, 26, 27, 30, 42, 56, 62, 69, 104, 104, 112, 129, 181, NA, NA, NA, NA, N
A, NA),
  censor = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6
1, 74, 79, 80, 81, 87, 87, 88, 89, 93, 97, 101, 104, 108, 109, 120, 131, 150, 231, 240, 400, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 67, 76, 104, 176, 231)
)

# inits
list(v = 1, beta0 = 0, beta1 = 0)
```

Note the following about the model code.

- The model code is very similar to that from the “gastric.odc” example from lecture. Notably, we use a Weibull prior for `times`. (Well, really, since we define the initial value `v = 1`, it is initially equivalent to an exponential distribution.)
- The regression here is a Poisson regression.
- The data is modified to be in a workable format for OpenBUGs. Specifically, the `times` values corresponding to censored observations (i.e. `censor = 1`) are re-defined as `NA`, and the `censor` values for these observations are re-defined to the `times` values (and, otherwise, left as 0).

Below is a summary of the output, truncated because there are lots of monitored parameters.

```
res_sim_summ_q1
```

```
## # A tibble: 246 x 9
##    var          mean        sd   MC_error  val2.5pc   median val97.5pc start sample
##    <chr>       <dbl>     <dbl>      <dbl>     <dbl>    <dbl>     <dbl> <dbl>  <dbl>
##  1 beta0     -4.810    0.5952    0.05698    -6.018    -4.819    -3.692   1001  10000
##  2 beta1      0.6133   0.2779    0.02384     0.08395   0.6006    1.203   1001  10000
##  3 median0  243.2    161.1      12.39       79.2     201.6     710.5     1001  10000
##  4 median1  101.7     24.31      1.567      64.54     98.28    160.5     1001  10000
##  5 v          0.8337   0.09158   0.007389    0.6639    0.8306    1.02    1001  10000
##  6 deviance 600.9      2.588     0.1723    598       600.3     608.1     1001  10000
##  7 S[1]       0.9835   0.007352  0.0006472   0.9654    0.9849    0.9942  1001  10000
##  8 S[2]       0.9608   0.01386   0.001242    0.9284    0.9627    0.9824  1001  10000
##  9 S[3]       0.9608   0.01386   0.001242    0.9284    0.9627    0.9824  1001  10000
## 10 S[4]       0.9508   0.01623   0.001458    0.9135    0.9527    0.9766  1001  10000
## # ... with 236 more rows
```

From the output shown above, we see that the posterior mean for `beta1` is 0.6133, **and we see that 95% CS is [0.0839, 1.2030].**

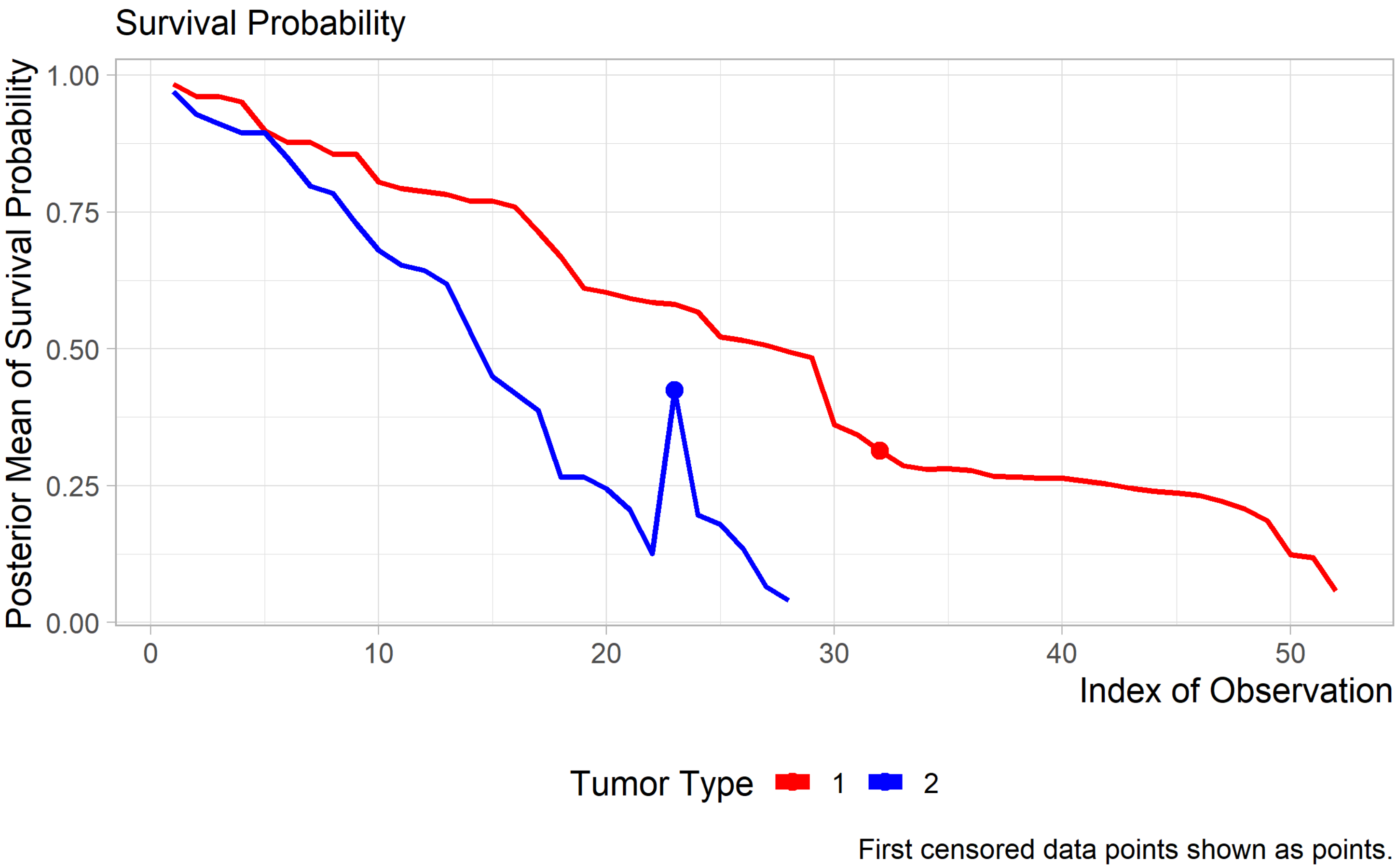
Aside: A Closer Evaluation

Below is a subset of the results from before. Specifically, the first 5 estimates of the `S` “array” of monitored parameters corresponding to each tumor `type` group are shown.

```
res_sim_summ_filt_q1
```

```
## # A tibble: 11 x 9
##    var     mean       sd  MC_error val2.5pc median val97.5pc start sample
##    <chr>  <dbl>    <dbl>     <dbl>    <dbl>  <dbl>     <dbl> <dbl>  <dbl>
##  1 S[1]  0.9835 0.007352 0.0006472   0.9654 0.9849    0.9942  1001  10000
##  2 S[2]  0.9608 0.01386  0.001242    0.9284 0.9627    0.9824  1001  10000
##  3 S[3]  0.9608 0.01386  0.001242    0.9284 0.9627    0.9824  1001  10000
##  4 S[4]  0.9508 0.01623  0.001458    0.9135 0.9527    0.9766  1001  10000
##  5 S[5]  0.8993 0.02601  0.002332    0.843  0.9017    0.9438  1001  10000
##  6 S[53] 0.9697 0.01431  0.001049    0.9344 0.9723    0.989   1001  10000
##  7 S[54] 0.9286 0.02646  0.001915    0.8647 0.9322    0.9673  1001  10000
##  8 S[55] 0.9108 0.03074  0.00221     0.8367 0.915     0.9568  1001  10000
##  9 S[56] 0.8942 0.03439  0.002457    0.8114 0.8987    0.9466  1001  10000
## 10 S[57] 0.8942 0.03439  0.002457    0.8114 0.8987    0.9466  1001  10000
## 11 S[58] 0.8491 0.04294  0.003013    0.7457 0.8541    0.917   1001  10000
```

Below is a plot of the survival curves corresponding to the two types of tumors (1 for aneuploid tumor, 2 for diploid tumor). (These are the `S` parameter estimates.)



We observe that there is a noticeable "jump" in the survival probability curve for the second tumor type group. This observation corresponds to the first censored value in the group. From my understanding, such a phenomenon is not necessarily impossible for an experiment where there is a relatively significant sample that is censored (here, 6 of 28 for the second group). Nonetheless, the "pseudo-discontinuity" suggests that there could be a better way of modeling the group of observations. But, since this is not asked of us, I leave this for another time.

2. Airfreight Breakage with Missing Data.

Instructions

A substance used in…

Response

The model code below is written so as to answer all parts of this question. (In particular, `lambdastar` corresponds to the expected average value asked for in (b) and `pred1` corresponds to the prediction asked for in (c).)

```
model {
  for(i in 1:n) {
    y[i] ~ dpois(lambda[i])
    lambda[i] <- exp(beta0 + beta1 * x[i])
  }
  for(i in 1:n) {
    x[i] ~ dpois(2)
  }
  beta0 ~ dnorm(0, 0.0001)
  beta1 ~ dnorm(0, 0.0001)
  lambdastar <- exp(beta0 + (4) * beta1)
  pred1 ~ dpois(lambdastar)
}

# data
list(
  n = 15,
  x = c(2, 1, 0, 2, NA, 3, 1, 0, 1, 2, 3, 0, 1, NA, NA),
  y = c(NA, 16, 9, 17, 12, 22, 13, 8, NA, 19, 17, 11, 10, 20, 2)
)

# inits
list(
  beta0 = 0,
  beta1 = 0
)
```

Note that we make the assumption that the probability of a value missing does not depend on the data that is missing, i.e. we assume that the data is missing at random (MAR). Thus, we simply specify missing X using a reasonable non-negative distribution, such as $\mathcal{Pois}(2)$, as suggested in the hint. (This corresponds to `x[i] ~ dpois(2)` in the model code.) If we did not assume this, then we would need to change the model code from its current form.

Below is a summary of the output from OpenBUGS.

```
## # A tibble: 10 x 9
##    var            mean      sd MC_error val2.5pc   median val97.5pc start sample
##    <chr>         <dbl>   <dbl>    <dbl>    <dbl>    <dbl>     <dbl> <dbl>  <dbl>
##  1 beta0        2.174  0.1346  0.007108    1.916    2.176     2.422   1001  10000
##  2 beta1        0.2884 0.07028 0.003705    0.1562   0.2874    0.4215  1001  10000
##  3 deviance   109.4    3.146   0.08923   105.5    108.5     117.2     1001  10000
##  4 lambdastar  28.37   5.349   0.2274     19.25    27.91     40.14    1001  10000
##  5 pred1       28.42   7.621   0.231      15       28        45       1001  10000
##  6 x[5]         1.26   0.7823  0.009606    0        1         3       1001  10000
##  7 x[14]        2.533  0.8165  0.01083     1        3         4       1001  10000
##  8 x[15]        0.1992 0.4333  0.00781     0        0         1       1001  10000
##  9 y[1]        15.65   4.134   0.04778     8       15        24       1001  10000
## 10 y[9]        11.74   3.531   0.05402     5       12        19       1001  10000
```

a

As shown in the output above, **the deviance of the fit is 109.4.**

b

The posterior estimate and 95% CS for $X = 4$ correspond to `lambdastar` from the output shown above.

From the output shown above, we see that the posterior mean (for `lambdastar`) is 28.37, **and we see that the 95% CS is [19.25, 40.14].**

c

The prediction for $X = 4$ is indicated by `pred1` in the output shown above (a). It has a posterior mean of 28.42. Also, note that this prediction has a CS of [15, 45], which is larger than that found in (b). (The standard deviation `sd` of the prediction here (for `pred1` is also larger than that in (b) (for `lambdastar`).) This is because the CS here accounts for uncertainty about the prediction, as well as the uncertainty about σ; on the other hand, the CS in (b) only accounts for uncertainty about the prediction `pred1`.

Drawing upon frequentist statistics, one might say that the interval here is analogous to a prediction interval, whereas the interval in (b) is analogous to a confidence interval. Prediction intervals are always equal to or greater than confidence intervals because they account for additional uncertainty.

d

As shown in the output above (a), **the estimates (of the posterior means) are as follows.**

- X_5 = 1.26
- X_{14} = 2.533
- X_{15} = 0.1992
- Y_1 = 15.65
- Y_9 = 11.74.