Gustavo Brusse

EDSD 2018/2019

Assignment – Event History

Problem 1.1

The way on how the flies` life span was measured in this study doesn't allow us to know the exact life span duration for each observation over time. We only can observe the time interval when the flies died. This incomplete dataset requires a particular approach in order to calculate the corresponding likelihood function. For each time interval we have different cases of censuring data and consequently different ways to define the likelihood function:

For $t_i \in (0, 7]$: Left-censoring

$$L(a,b) = \prod_{i=1}^{30} 1 - S(7)$$

For $t_i \in (7, 14]$: Interval-censoring

$$L(a,b) = \prod_{i=1}^{211} [S(7) - S(14)]$$

For $t_i \in (14, 21]$: Interval-censoring

$$L(a,b) = \prod_{i=1}^{355} [S(14) - S(21)]$$

For $t_i > 21$:: Right-censoring

$$L(a,b) = \prod_{i=1}^{369} S(21)$$

For the complete interval of the study, the corresponding likelihood function is given by:

$$L(a,b) = \prod_{i=1}^{30} 1 - S(7) * \prod_{i=1}^{211} [S(7) - S(14)] * \prod_{i=1}^{355} [S(14) - S(21)] * \prod_{i=1}^{369} S(21)$$

Thus, the log-likelihood is:

$$l(a,b) = \sum_{i=1}^{30} \ln(1 - S(7)) + \sum_{i=1}^{211} \ln(S(7) - S(14)) + \sum_{i=1}^{355} \ln\big(S(14) - S(21)\big) + \sum_{i=1}^{369} \ln S(21)$$

In order to apply in R, the log-likelihood can be written using terms of $S(t) = 1 - F(t)$:

$$l(a,b) = \sum_{i=1}^{30} \ln \mathrm{F}(7) + \sum_{i=1}^{211} \ln(S(7) - S(14)) + \sum_{i=1}^{355} \ln\big(S(14) - S(21)\big) + \sum_{i=1}^{369} \ln S(21)$$

$$l(a,b) = 30 * \ln F(7) + 211 * \ln(F(14) - F(7)) + 355 * \ln(F(21) - F(14)) + 369 * \ln(1 - F(21))$$

Assuming that the data is distributed by Weibull distribution, the maximum likelihood estimators are given by optimization the log-likelihood function above in relation to the parameters a and b. Using the function *optmin* in R, we have:

$$\hat{a}_{MLE} = 3.04$$
$$\hat{b}_{MLE} = 21.25$$

Problem 1.2

The 95% confidence Intervals are defined as:

$$[\hat{a}_{MLE} - z_{0.975} * s.e_{\hat{a}};\ \hat{a}_{MLE} + z_{0.975} * s.e_{\hat{a}}]$$
$$[\hat{b}_{MLE} - z_{0.975} * s.e_{\hat{b}};\ \hat{b}_{MLE} + z_{0.975} * s.e_{\hat{b}}]$$

Where $z_{0.975} = 1.96$ and the maximum likelihood estimators standard errors are:

$$S.E_{\hat{a}} = \sqrt{\frac{1}{-l"(a)}} = \sqrt{\frac{1}{-\boldsymbol{H}(1,1)}} = 0.12$$
$$S.E_{\hat{b}} = \sqrt{\frac{1}{-l"(b)}} = \sqrt{\frac{1}{-\boldsymbol{H}(2,2)}} = 0.29$$

Then, the estimated confidence intervals are:

$$CI[a, 0.95] = [3.04 - 1.96 * 0.12;\ 3.04 + 1.96 * 0.12]$$
$$CI[a, 0.95] = [2.8; 3.3]$$

$$CI[b, 0.95] = [21.25 - 1.96 * 0.29;\ 21.25 + 1.96 * 0.29]$$
$$CI[a, 0.95] = [20.7; 21.8]$$

Previously, using the full flies' dataset, we have the parameters of the Weibull distributions and the respective confidence intervals:

$$\hat{a}_{MLE.full} = 3.08$$
$$\hat{b}_{MLE.full} = 21.81$$

$$CI[a, 0.95] = [2.9; 3.23]$$
$$CI[a, 0.95] = [21.3; 22.3]$$

It can be noted that with the complete data we have different estimation for the parameters a and b with smaller range confidence intervals. This difference is observed because resuming the data in intervals yield a lack of information when compared to the complete single data. The variability (information) between one single observation and the other is lost when we group into intervals, making the standard errors estimation larger and the confidence interval range larger.

Problem 1.3

The proportion of flies still alive at age 30 days is given by:

$$P(t > 30) = S(30)$$

$$S(30) = 1 - F(30)$$

Using the function *pweibull* with a and b MLE estimators, we have:

$$S(30) = 0.058$$

$$prop.full\ dataset = \frac{71}{965} = 0.074$$

Using the δ-Method to estimate the standard error of $\widehat{S(30)}$:

$$S.E\left(\widehat{S(30)}\right) = \sqrt{V(S(30)}$$
$$S.E\left(\widehat{S(30)}\right) = V[g(a,b)] = \sqrt{g` * V(a,b) * g`}$$

$$S.E\left(\widehat{S(30)}\right) = 0.011$$

Then, the interval confidence for $\widehat{S(30)}$ is given by:

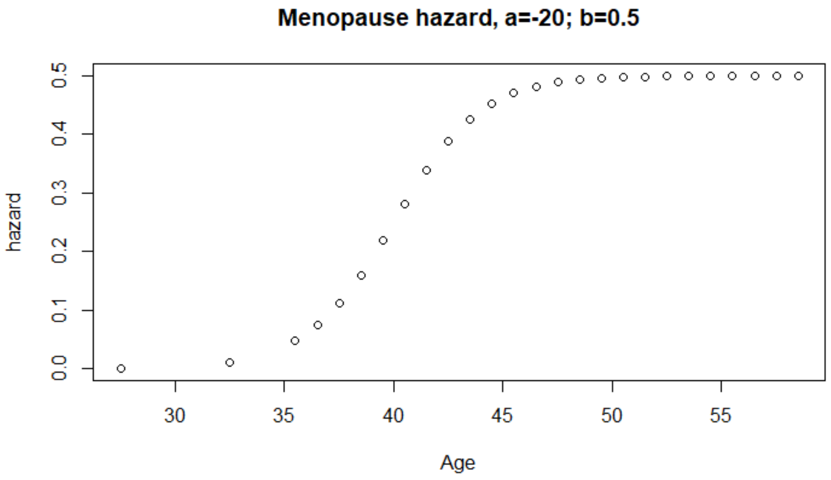
$$CI\left(\widehat{S(30)}, 0.95\right) = \left[\widehat{S(30)} - z_{0.975} * s.e_{\widehat{s(30)}};\ \widehat{S(30)} + z_{0.975} * s.e_{\widehat{s(30)}}\right]$$

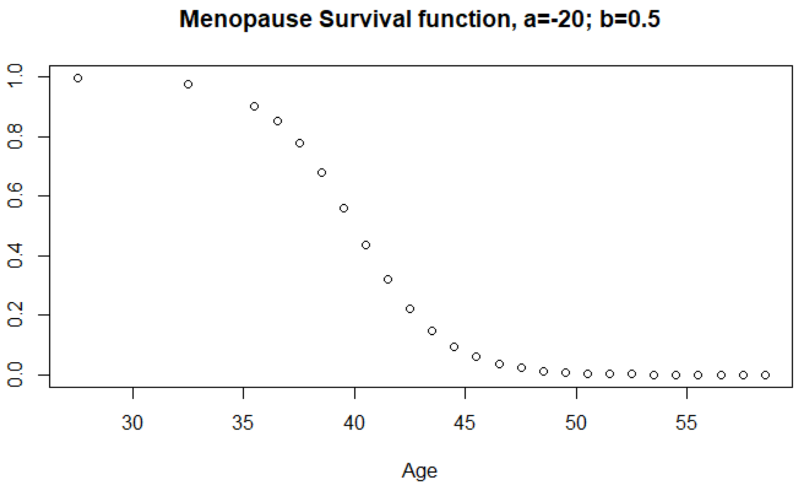
$$CI\left(\widehat{S(30)}, 0.95\right) = [0.058 - 1.96 * 0.011;\ 0.058 + 1.96 * 0.011]$$

$$CI\left(\widehat{S(30)}, 0.95\right) = [0.036; 0.079]$$

We can conclude that, even though the proportion of flies still alive at age 30 days is greater in the full dataset, the interval confidence includes 0.07, so it’s not statistically significant different.

Problem 2.1





Problem 2.2

As we don`t have any kind of censoring data in menopause study, we can apply the general approach (without censoring) for obtaining the likelihood function, assuming the data has a logistic hazard:

$$L(a,b) = \prod_{i=1}^{n=2076} F(t_i; a,b)^{\delta_i} \cdot \; S(t_i; a,b)^{(1-\delta_i)}$$

$$l(a,b) = \sum_{i=1}^{n=2076} \left[\delta_i \, log\, F(t_i) \; + \; (1-\delta_i) \;\, log\, S(t_i)\right] \; =$$

$$= \sum_{i=1}^{n=2076} \left[\delta_i \, log \left(1 - S(t_i)\right) + (1-\delta_i) \cdot \; (-H(t_i))\right]$$

Writing just in terms of $h(t)$ and $H(t)$, we have:

$$l(a,b) = \sum \left[\delta\text{i} \;\; log \left(1 - \left(\frac{1+e^{a+bt}}{1+e^{a}}\right)^{-1}\right) + \left((1-\delta i) \;\, log \left(\frac{1+e^{a+bt}}{1+e^{a}}\right)^{-1} \right) \right]$$

Problem 2.3

The R program used were (see in detail in R code attached):

```
a<--20
b<-0.5
t<-data.meno$age.interview
h.t<-(b*exp(a+(b*t))/(1+exp(a+(b*t))))

## Survival: -H(x)=ln(s(x)) -> exp(-H(x))=s(x) ##
H.t<-log((1+exp(a+(b*t)))/1+exp(a))
S.t<-exp((-1)*H.t)
log.like<-((1-S.t)^data.meno$menop)*(S.t)^(1-data.meno$menop)

# Log-Likelihood function
MLE<-function(par,data.meno){
  a=par[1]
  b=par[2]
  F.t<-1-(((1+exp(a+b*t))/(1+exp(a)))^(-1))
  MLE<-sum(data.meno$menop*log(F.t)+(1-data.meno$menop)*(-H.t))
  return(MLE)
}
```

Problem 2.4

Using the *optim* function, the maximum likelihood estimators for a and b are:

$$\hat{a}_{MLE} = -21.47$$
$$\hat{b}_{MLE} = 0.43$$

The standard errors are obtained by:

$$S.E_{\hat{a}} = \sqrt{\frac{1}{-l"(a)}} = \sqrt{\frac{1}{-\boldsymbol{H}(1,1)}} = 1.15$$
$$S.E_{\hat{b}} = \sqrt{\frac{1}{-l"(b)}} = \sqrt{\frac{1}{-\boldsymbol{H}(2,2)}} = 0.02$$

And the confidence intervals are:

$$[\hat{a}_{MLE} - z_{0.975} * s.e_{\hat{a}};\ \hat{a}_{MLE} + z_{0.975} * s.e_{\hat{a}}]$$
$$[\hat{b}_{MLE} - z_{0.975} * s.e_{\hat{b}};\ \hat{b}_{MLE} + z_{0.975} * s.e_{\hat{b}}]$$

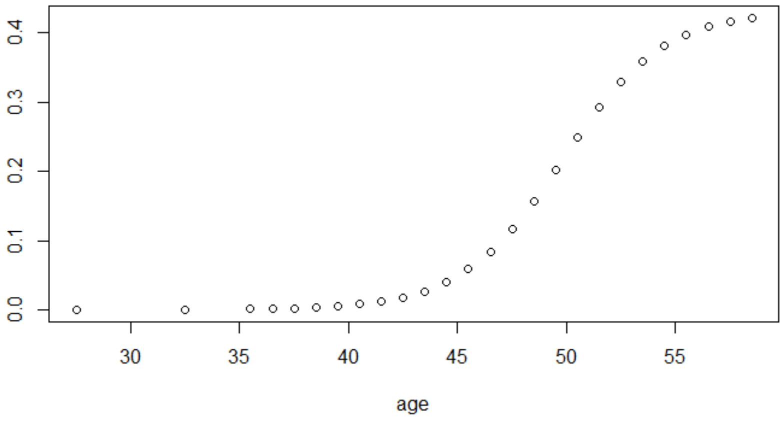
$$CI[a, 0.95] = [-21.47 - 1.96 * 1.15;\ -21.47 + 1.96 * 1.15]$$
$$CI[a, 0.95] = [-23.75; -19.21]$$

$$CI[b, 0.95] = [0.43 - 1.96 * 0.02;\ 0.43 + 1.96 * 0.02]$$
$$CI[b, 0.95] = [0.38; 0.47]$$

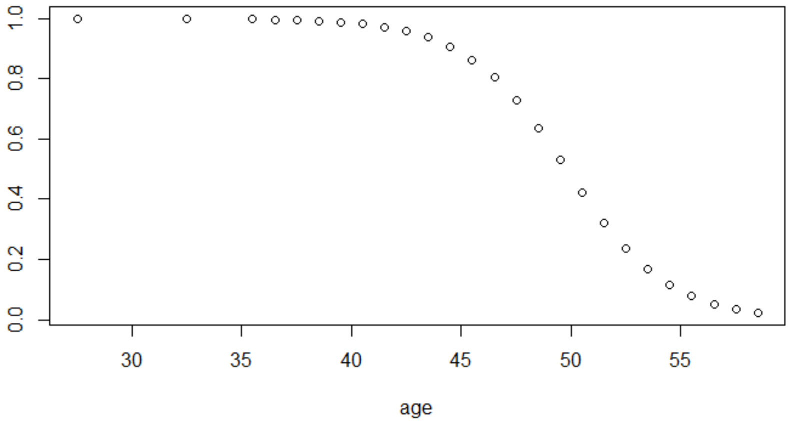
Problem 2.5

From the parameters estimated above, we have these respective graphs:

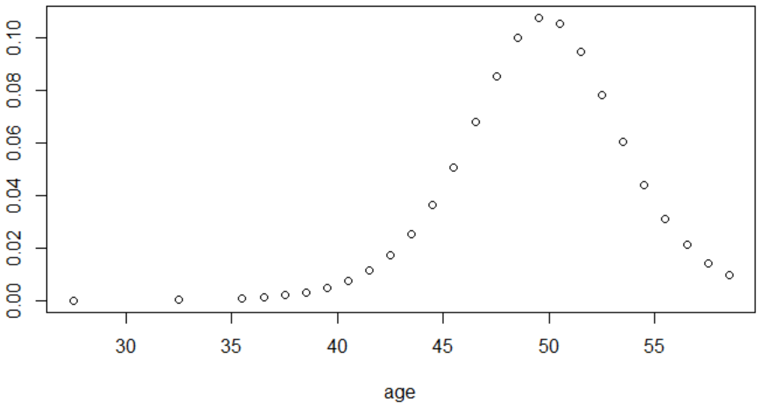
hazard



survival



density

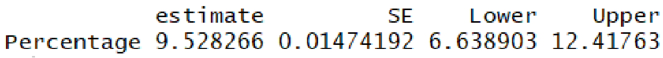


The hazard plot shows that the risk of a woman enter in menopause is very low below the age of 40 years, almost zero as should be biologic expected. After this age, there is a high increment of the hazard rate until the age of 55, where risk of entering in menopause start to be constant level, as almost every woman has already experienced the event.

The graph of the density shows that the modal age of entering in menopause is about 50 years old.

Problem 2.6

The estimation, the standard error and respect confidence interval for percentage of women that have not experienced menopause by age 55 is given below:



```
           estimate         SE    Lower    Upper
Percentage 9.528266 0.01474192 6.638903 12.41763
```

Problem 3

Assuming that a positive random variable T has the property $h(t) > 0$ for $t \in (0, t*]$ and $h(t) = 0$ for all $t > t*$,where $t* > 0$ implies that there is a certain point in time $t*$ when the occurrence of this event stop happening. It happens because with a $h(t) = 0$ the density function becomes:

$$f(t) = h(t) * S(t) = 0$$

And the survival function becomes constant, as no event occurs. So, we can conclude that all the events of this phenomenon happen before the time t*.