hw 6

The Child Mortality dataset (under LS Content, or the 'child' data frame in the "eha" package for R) contains data on child mortality in Sweden during the 1800's.

```
library(eha)
```

```
## Warning: package 'eha' was built under R version 4.1.3
```

```
library(survival)
child = eha::child
head(child)
```

```
##      id   m.id    sex socBranch  birthdate enter   exit event illeg  m.age
## 3     9 246606   male   farming 1853-05-23     0 15.000     0    no 35.009
## 42  150 377744   male   farming 1853-07-19     0 15.000     0    no 30.609
## 47  158 118277   male    worker 1861-11-17     0 15.000     0    no 29.320
## 54  178 715337   male   farming 1872-11-16     0 15.000     0    no 41.183
## 78  263 978617 female    worker 1855-07-19     0  0.559     1    no 42.138
## 102 342 282943   male   farming 1855-09-29     0  0.315     1    no 32.931
```

1. Evaluate the data in the dataset to see how well it follows a Weibull distribution. Give a 1-2 sentence explanation and description of plots or other diagnostics used.

From the plots below we see that the child data looks like it follows a scaled weibull distribution. We used the optim function to find the MLE for the parameters of the distribution.

```
logLikWeib2 <- function(par, tt, status) {
  mu <- par[1]
  sigma <- par[2]
  lambda.p <- exp(-mu)
  alpha.p <- 1/sigma
  uncens <- sum(
    dweibull(tt[status==1], alpha.p, 1/lambda.p, log=TRUE)
  )
  rcens <- sum(
    pweibull(tt[status==0], alpha.p, 1/lambda.p, log=TRUE,
             lower.tail = FALSE)
  )
  uncens + rcens
}

logLikWeib2(c(4,2), 10, 1)
```

```
## [1] -4.272407
```

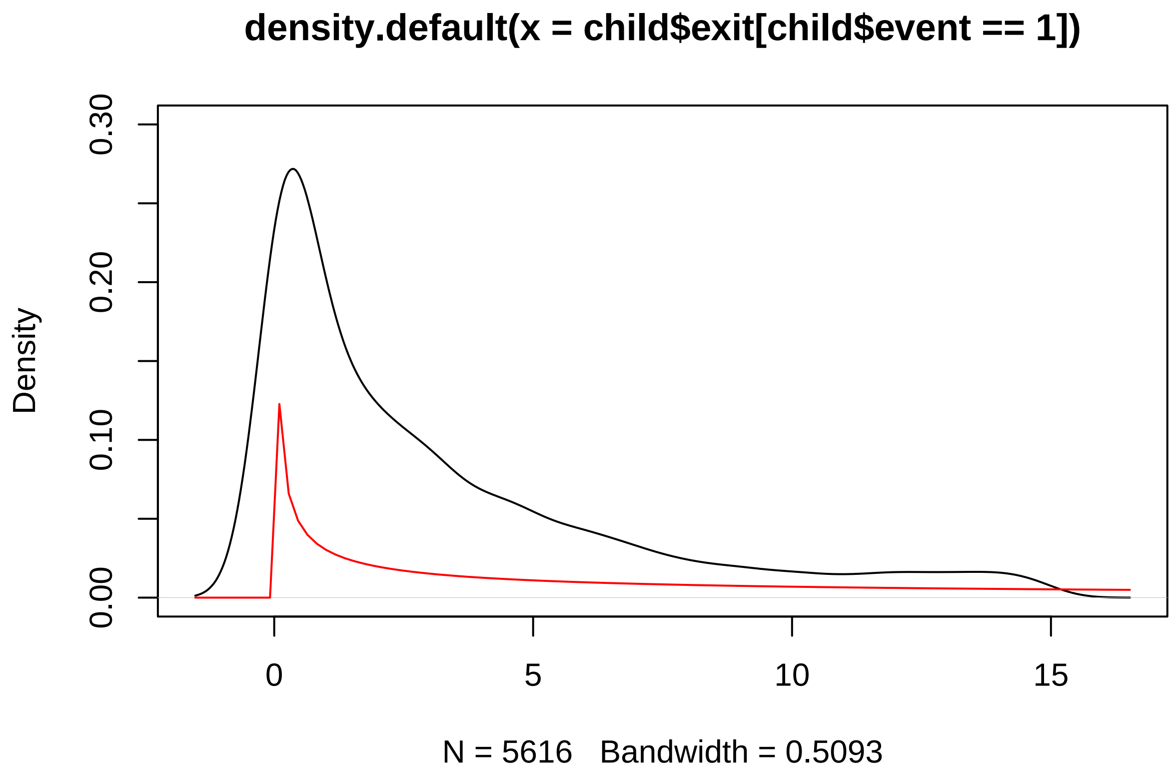
```r
o2 <- optim(c(4,2), fn=logLikWeib2,
            control=list(fnscale= -1),
            tt = child$exit,
            status = child$event
)

o2$par
```

```
## [1] 6.120841 2.413445
```

```r
survreg(Surv(exit, event) ~ 1,
  data=child,
  dist='weibull')
```

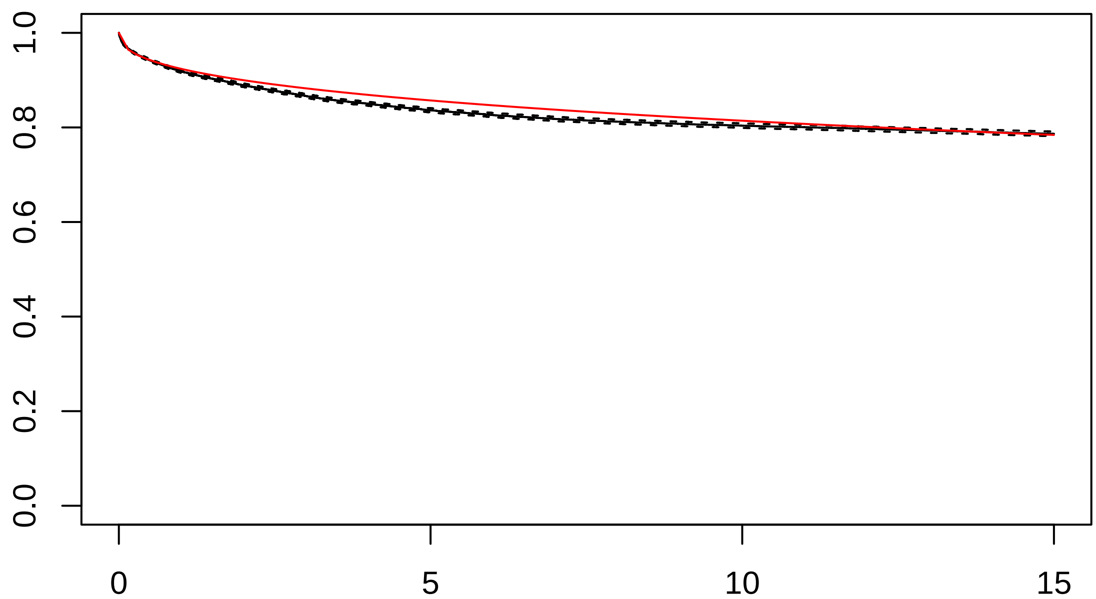
```
## Call:
## survreg(formula = Surv(exit, event) ~ 1, data = child, dist = "weibull")
##
## Coefficients:
## (Intercept)
##    6.120483
##
## Scale= 2.413115
##
## Loglik(model)= -25165   Loglik(intercept only)= -25165
## n= 26574
```

```r
plot(density(child$exit[child$event==1]),
     type='l', ylim=c(0,0.3))
curve(dweibull(x, 1/o2$par[2], exp(o2$par[1])),
      add=TRUE, col='red')
```




```
plot(survfit(Surv(exit,event)~1, data=child), main = "Data and Weibull")
curve(pweibull(x, 1/o2$par[2], exp(o2$par[1]),
               lower.tail = FALSE),
      add=TRUE, col='red')
```

Data and Weibull



2. Fit a parametric regression model with the Weibull distribution to the data (use exit and event variables, do not worry about enter) with sex of the child, mothers age, and the social group (socBranch) as predictors.

```
colnames(child)
```

```
##  [1] "id"         "m.id"       "sex"        "socBranch" "birthdate" "enter"
##  [7] "exit"       "event"      "illeg"      "m.age"
```

```
fit1 = survreg(Surv(exit,event) ~ sex + socBranch + m.age,
             data=child, dist = "weibull")

summary(fit1)
```

```
##
## Call:
## survreg(formula = Surv(exit, event) ~ sex + socBranch + m.age,
##     data = child, dist = "weibull")
##                        Value Std. Error     z       p
## (Intercept)          6.65062    0.28152 23.62 < 2e-16
## sexfemale            0.19949    0.06454  3.09 0.00199
## socBranchfarming     0.03944    0.22267  0.18 0.85941
## socBranchbusiness -0.82289    0.34028 -2.42 0.01560
## socBranchworker   -0.24225    0.22749 -1.06 0.28693
```

```
## m.age             -0.01800    0.00514 -3.50 0.00046
## Log(scale)         0.88025    0.01287 68.38 < 2e-16
##
## Scale= 2.41
##
## Weibull distribution
## Loglik(model)= -25142.2   Loglik(intercept only)= -25165
##  Chisq= 45.63 on 5 degrees of freedom, p= 1.1e-08
## Number of Newton-Raphson Iterations: 8
## n= 26574
```

3. Further explore the fit in part 2 to see if a non-linear effect of age gives a better fit and if any interactions are important.

A non-linear effect for age lowers the AIC and so we believe it improves the fit of the model. Adding in an interaction between sex and social branch does not improve the model fit so we leave it out.

```r
fit2 = survreg(Surv(exit,event) ~ sex + socBranch + pspline(m.age, df = 4),
             data=child, dist = "weibull")

summary(fit2)
```

```
##
## Call:
## survreg(formula = Surv(exit, event) ~ sex + socBranch + pspline(m.age,
##     df = 4), data = child, dist = "weibull")
##                     Value Std. Error     z       p
## (Intercept)        5.2926     0.7397  7.16 8.3e-13
## sexfemale          0.1998     0.0645  3.10   0.002
## socBranchfarming   0.0420     0.2226  0.19   0.850
## socBranchbusiness -0.8244     0.3401 -2.42   0.015
## socBranchworker   -0.2387     0.2274 -1.05   0.294
## ps(m.age)3         0.3702     0.3878  0.95   0.340
## ps(m.age)4         0.7362     0.6309  1.17   0.243
## ps(m.age)5         0.9535     0.7228  1.32   0.187
## ps(m.age)6         0.8567     0.7204  1.19   0.234
## ps(m.age)7         0.8512     0.7081  1.20   0.229
## ps(m.age)8         0.8543     0.7070  1.21   0.227
## ps(m.age)9         0.7396     0.7085  1.04   0.297
## ps(m.age)10        0.5332     0.7100  0.75   0.453
## ps(m.age)11        0.4436     0.7138  0.62   0.534
## ps(m.age)12        0.4128     0.7432  0.56   0.579
## ps(m.age)13        0.3992     0.8589  0.46   0.642
## ps(m.age)14        0.3889     1.0964  0.35   0.723
## Log(scale)         0.8797     0.0129 68.34 < 2e-16
##
## Scale= 2.41
##
## Weibull distribution
## Loglik(model)= -25137.2   Loglik(intercept only)= -25165
##  Chisq= 55.6 on 7.5 degrees of freedom, p= 1.9e-09
## Number of Newton-Raphson Iterations:  6 16
## n= 26574
```

```r
AIC(fit1)
```

```
## [1] 50298.35
```

```r
AIC(fit2)
```

```
## [1] 50293.32
```

4. Refit your model from part 3 using a distribution other than the Weibull. Briefly describe how this fit compares to the previous one.

We refit the model using a Gaussian distribution instead of the Weibull. This significantly increased AIC. However, fitting the data to a lognormal decreased AIC giving us the best fit of all the models yet. This holds true for BIC as well.

```r
fit3 = survreg(Surv(exit,event) ~ sex + socBranch + pspline(m.age, df = 4),
             data=child, dist = "gaussian")

fit4 = survreg(Surv(exit,event) ~ sex + socBranch + pspline(m.age, df = 4),
             data=child, dist = "lognormal")


AIC(fit1)
```

```
## [1] 50298.35
```

```r
AIC(fit2)
```

```
## [1] 50293.32
```

```r
AIC(fit3)
```

```
## [1] 64629.14
```

```r
AIC(fit4)
```

```
## [1] 49911.91
```

```r
# BIC
BIC(fit1)
```

```
## [1] 50355.67
```

```r
BIC(fit2)
```

```
## [1] 50370.85
```

```
BIC(fit3)
```

```
## [1] 64706.05
```

```
BIC(fit4)
```

```
## [1] 49989.37
```