Survival Assingment 2

Faizan Khalid Mohsin

February 1, 2019

Contents

1 Abstract

Background: Time to events are a very important class of random variables, for which the field of survival analysis has be developed. In survival analysis, the flexibility that the non-parametric coxproportional hazard model offers, has become very popular. However, if the data allows, using fully parametric models can be very powerfull as well. Especially, estimating the parameters of a model exactly can be a very powerful predictive tool. **Purpose:** In this study we will compare four parametric survival models: log-normal, log-logistic, exponential and weibull model. We will first use the Kaplan-Meier curves, gamma distribution, AIC and likelihood ratio test to see which model fits the data best. Using which, we will conduct a complete survival analysis. After that, we will further perform an additional survival analysis using the log-logistic model and present the results. **Methods:** We use the Melanoma data set to perform survival analysis with death the outcome variable, and the tumour stage, presence of skin ulcers, and tumour thickness as the main covariates. **Results:** The Weilbull model had the smallest AIC (551.9342) and LogLikelihood (-273.9671) of the four models. When we fit the complete model using the Weibull we get . . . . For the log-logistic survival model. **Conclusion:**

2 Introduction

3 Methods

3.1 Data Set

3.2 Statistical Analysis

4 Results

4.1 Descriptive Statisitcs

4.1.1 Question 1 a

```
data0 = fread("melanoma.csv", na.strings=c("","  ","NA"))
#str(data0)

# lOOKING at the number of missing data.
#sapply(data0, function(x) sum(is.na(x)))
# data01 = fread("melanoma1.csv")
# str(data01)
# datasas0 = read.sas7bdat("melanoma.sas7bdat")
# str(datasas0)

data1=data0
data1$biopsydate = as.Date(data0$biopsydate,format='%d%b%Y')
data1$vstatusdate = as.Date(data0$vstatusdate, format='%d%b%Y')
#str(data1)
#summary(data1)
#sapply(data1[, c("vstatus", "clarklevel")], function(x) unique(x))

data1$event = ifelse(data1$vstatus == "Dead", 1, 0)
#str(data1)
```

```
data1$days  = data1$vstatusdate - data1$biopsydate
#str(data1)
data1$vstatus = as.factor(data1$vstatus)
data1$clarklevel = as.factor(data1$clarklevel)

#data1$vstatusdate[1] - data1$biopsydate[1]

incorrect_date = filter(data1, days<=0)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
# some time differences are 0. Are these administrative errors? One of them had the event of Ulceration

# Therefore, for now we will exclude these patients.

data1$years = data1$days/365.2422
data1$years = as.numeric(data1$years)

#str(data1)
#table(data1$vstatus)
proportion_alive = sum(data1$event)/length(data1$event)

#proportion_alive
# approximately 8%

data_correct_people = filter(data1, days>0)
#str(data_correct_people)

# Drop all unnessisary variables to create final dataset.
datta = select(data_correct_people, -id, -days)
#str(datta)
```

```
kable(table1, caption = "Summary of Variables.")
```

Table 1: Summary of Variables.

	level	Overall
n		911
vstatus (%)	Alive	838 (92.0)
	Dead	73 (8.0)
clarklevel (%)	I	50 (5.6)
	II	157 (17.6)
	III	261 (29.3)
	IV	380 (42.6)
	V	43 (4.8)
ulceration (%)	0	451 (66.7)
	1	225 (33.3)
thickness (mean (sd))		1.94 (2.05)
years (mean (sd))		1.51 (1.24)

From Table 1 we can see that about 8% of the patients died at the end of their respective follow-up times.

4.1.2 Question 1 b

```
kable(table2, caption = "Summary of Variables by Clark Level.")
```

Table 2: Summary of Variables by Clark Level.

	level	I	II	III	IV	V	p	test
n		50	157	261	380	43		
vstatus (%)	Alive	49 (98.0)	155 (98.7)	247 (94.6)	336 (88.4)	34 (79.1)	<0.001	
	Dead	1 (2.0)	2 (1.3)	14 (5.4)	44 (11.6)	9 (20.9)		
ulceration (%)	0	26 (96.3)	108 (90.0)	144 (76.6)	158 (52.8)	13 (38.2)	<0.001	
	1	1 (3.7)	12 (10.0)	44 (23.4)	141 (47.2)	21 (61.8)		
thickness (mean (sd))		0.03 (0.11)	0.48 (0.34)	1.24 (0.98)	2.74 (1.78)	6.24 (3.50)	<0.001	
years (mean (sd))		1.20 (1.36)	1.39 (1.32)	1.69 (1.29)	1.50 (1.16)	1.38 (0.99)	0.032	

From Table 2 we can see that the percentage of people with ulceration and mean tumor thickness increase as the Clark level increases. We also see that the percentage of people dead increases as Clark level increases. So as the Clark level increases, the percentage of people with uleration is increasing and the average size of the tumor thickness that people have is also increasing and, as can be seen from Table 2, more people are dying.

```
# table(datta$clarklevel, datta$ulceration)
# m = cor(select(datta, as.factor(clarklevel), ulceration, thickness))
# corrplot()
```

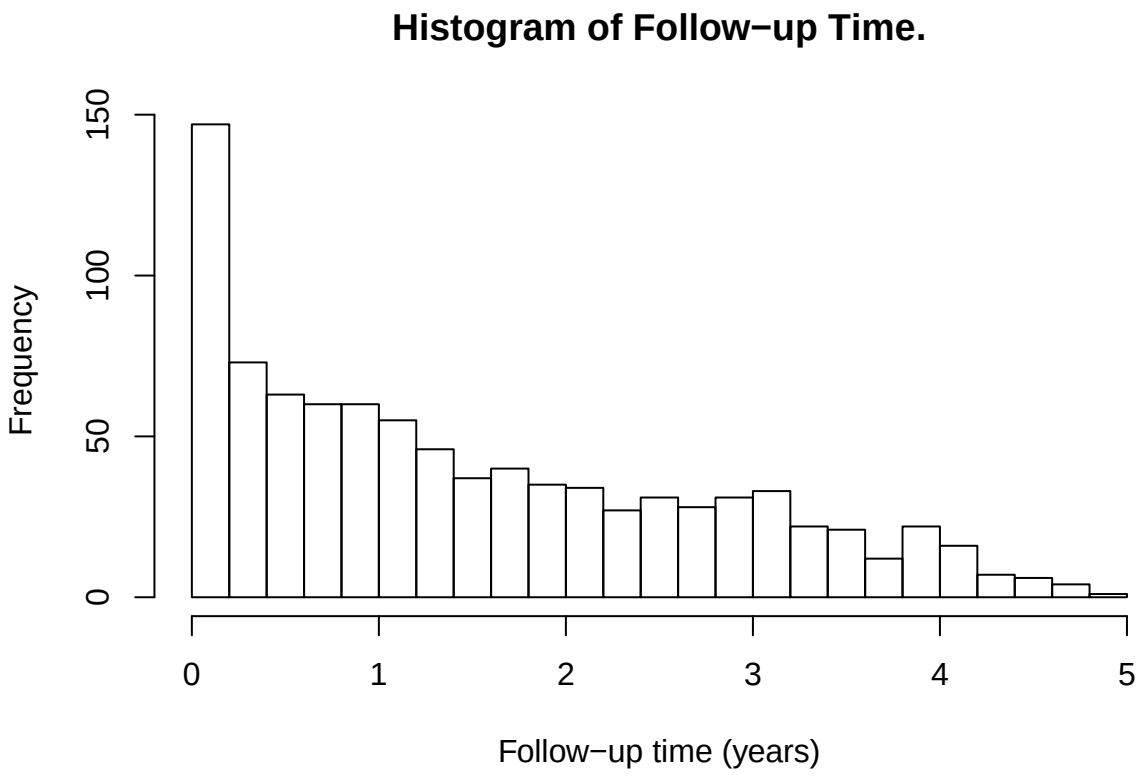
4.2 Parametric Survival Models

4.2.1 Question 1 c

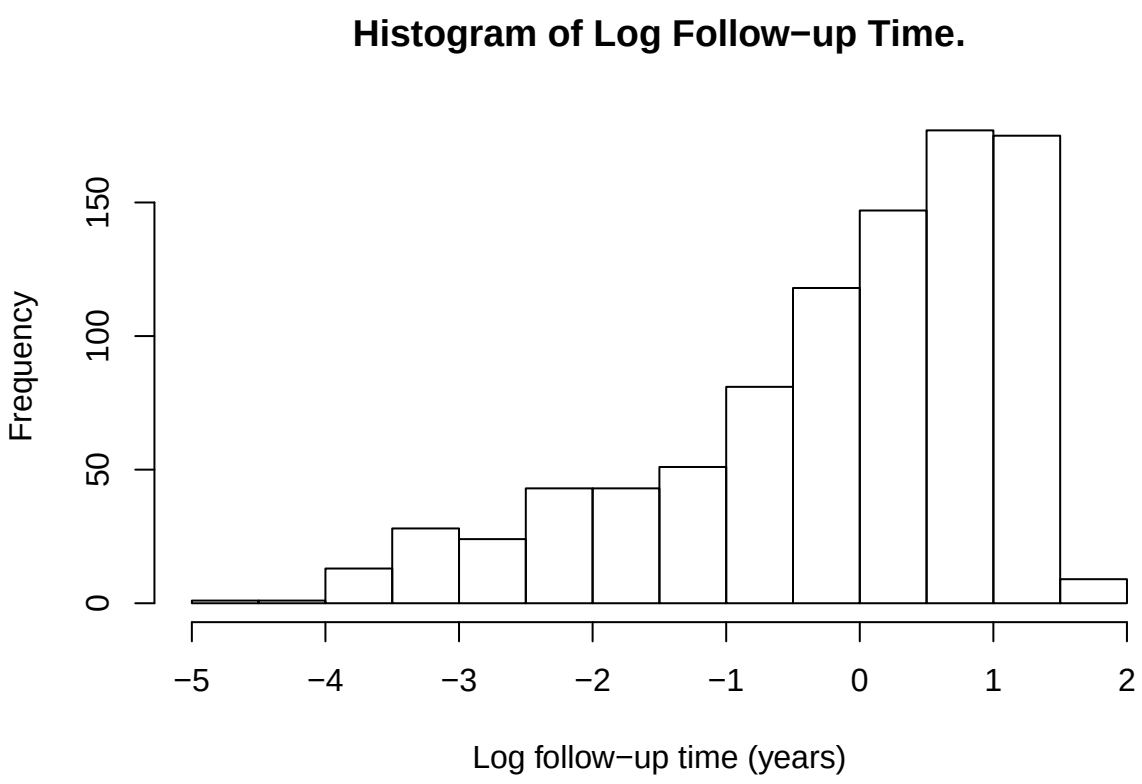
Do we need to create the survival, hazard and cumulative hazard/survival curves for these models.

First, we look at the distribution of the follow-up time of the patients.

```
hist(datta$years, breaks = 20, xlab = "Follow-up time (years)",
       main = "Histogram of Follow-up Time.")
```




```
#?hist
hist(log(datta$years), breaks = 20, xlab = "Log follow-up time (years)",
       main = "Histogram of Log Follow-up Time.")
```

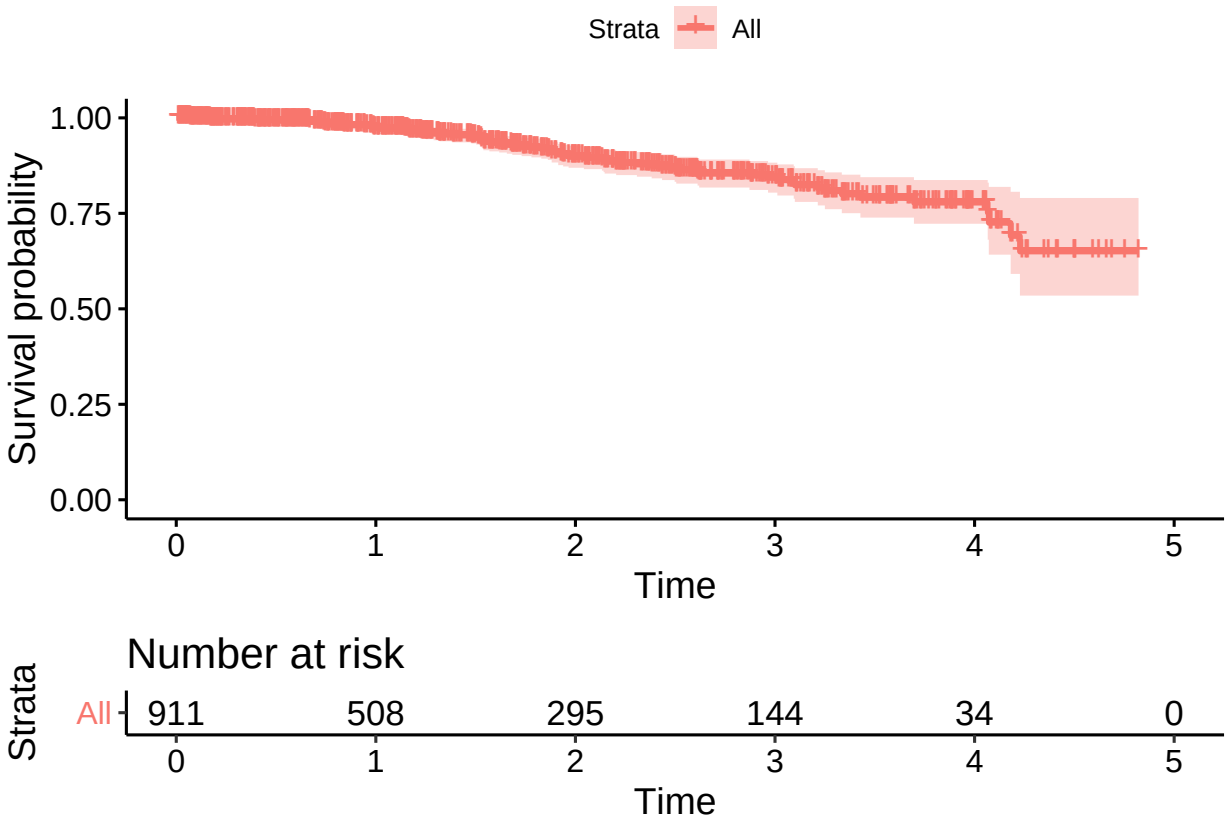



```
#qplot(chol$AGE, geom="histogram")
```

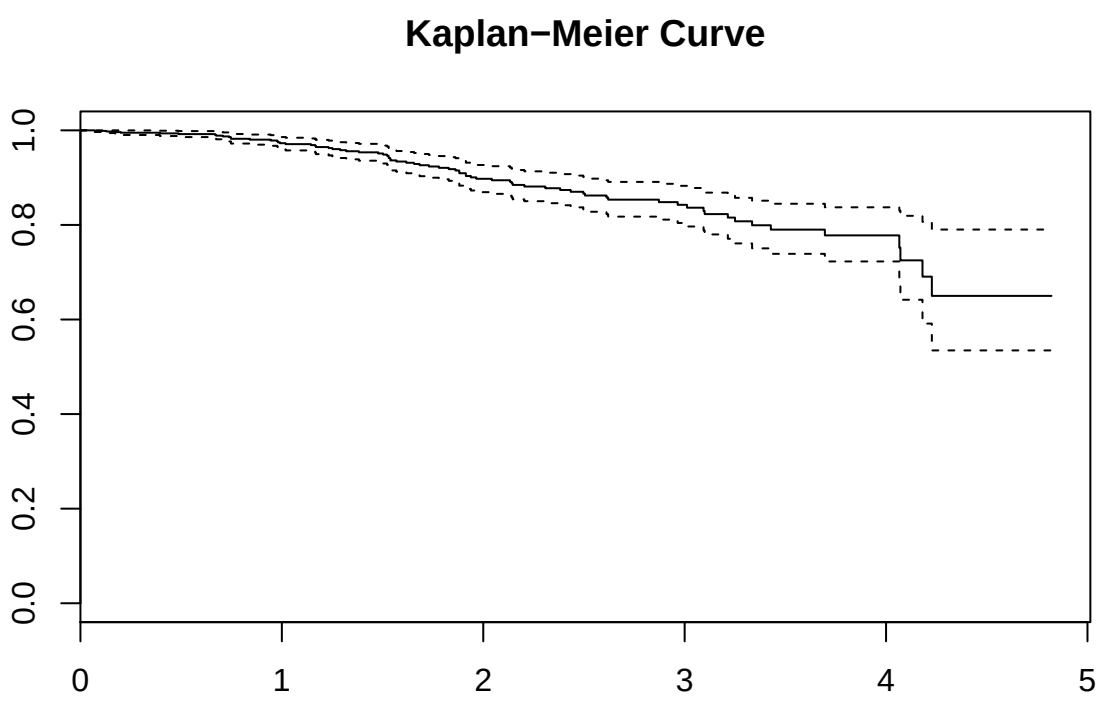
From the above histograms we can see that the follow-up times first have a quick drop, then decrease steadily and then are almost constant and then, again start to steadily decay.

We will now create the Kaplan-Meier curve.

```
heading = "Kaplan-Meier Curve"
fit = survfit(Surv(time = years, event = event) ~ 1, data = datta)
ggsurvplot(fit, risk.table = TRUE, data = datta) #+ ggtitle(heading)
```




```
plot(fit, main = heading)
```

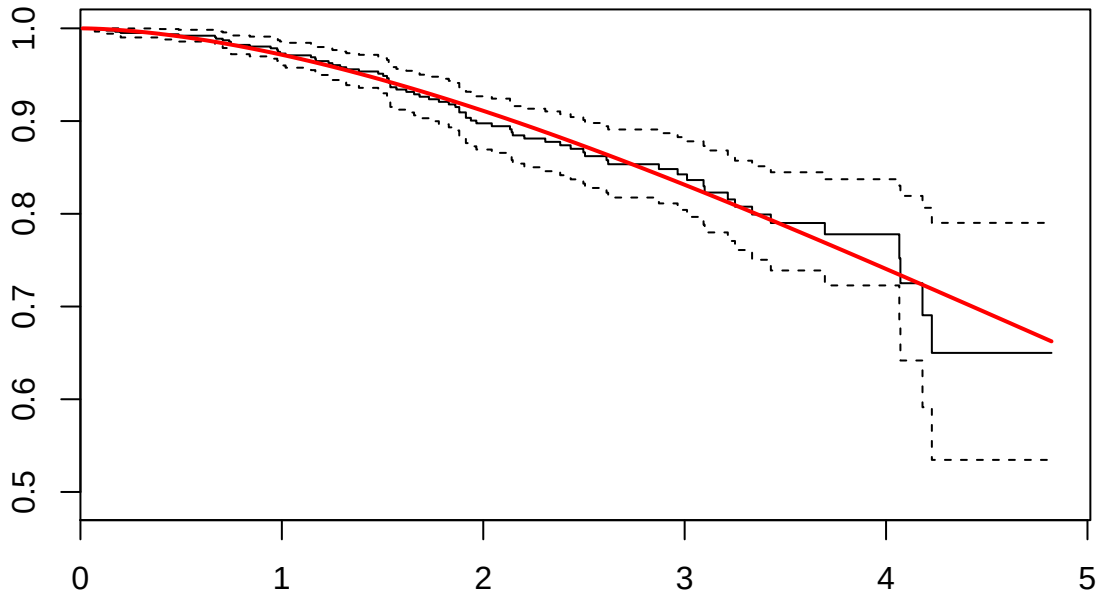


We create parametric survival models.

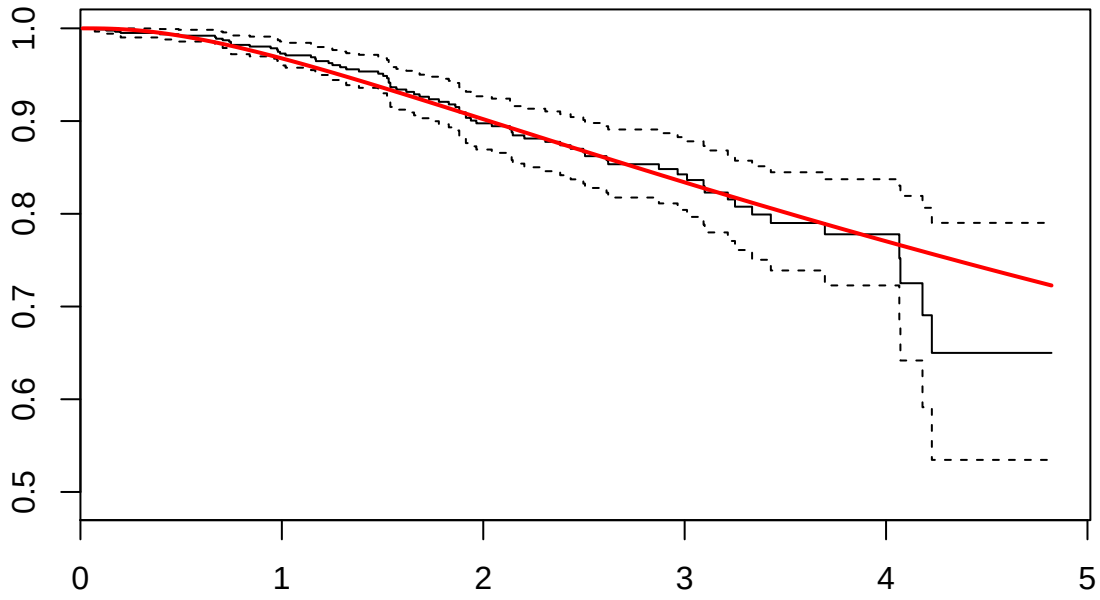
```
# We will fit the parametric survival models using the R package flexsurv.


weilbull = flexsurvreg(formula = Surv(time = years, event = event) ~ 1, data = datta, dist = "weibull")
weilbull
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ 1,
##     data = datta, dist = "weibull")
##
## Estimates:
##        est     L95%    U95%    se
## shape   1.689   1.412   2.020   0.154
## scale   8.152   6.399  10.386   1.007
##
## N = 911,  Events: 73,  Censored: 838
## Total time at risk: 1373.973
## Log-likelihood = -273.9671, df = 2
## AIC = 551.9342
```

```
plot(weilbull, ymin = .49, ci = FALSE)
```



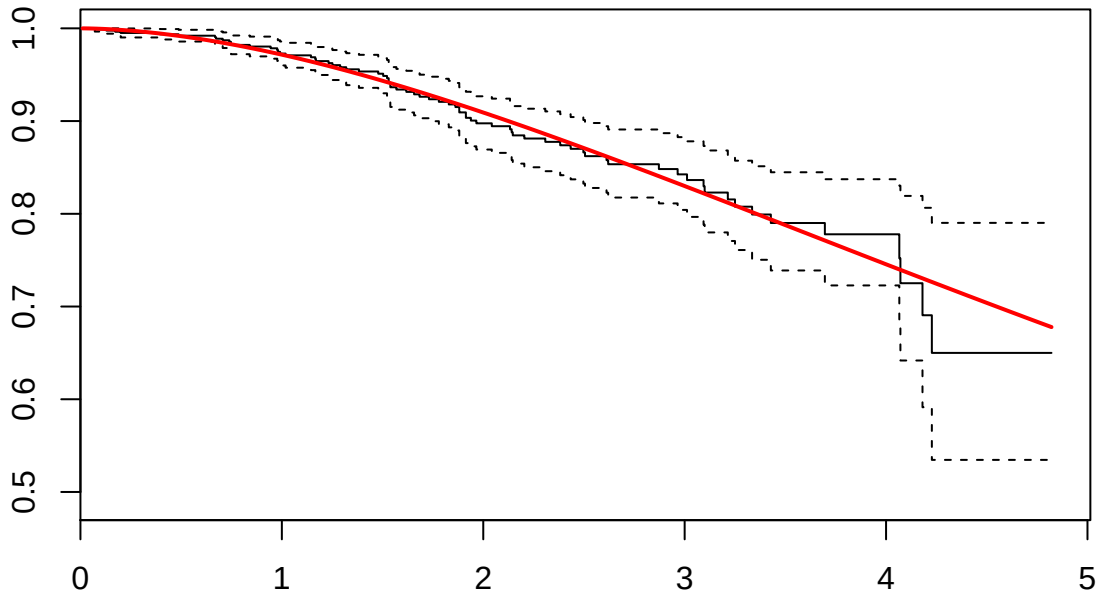

```
#ggsurvplot(weilbull)
# Question, what is default censuring for Surv(), does it need type of censuring as an input.
lnorm = flexsurvreg(formula = Surv(time = years, event = event) ~ 1, data = datta, dist = "lnorm")
plot(lnorm, ymin = .49, ci = FALSE)
```




```
llogis = flexsurvreg(formula = Surv(time = years, event = event) ~ 1, data = datta, dist = "llogis")
llogis
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ 1,
##     data = datta, dist = "llogis")
##
## Estimates:
##        est    L95%   U95%   se
## shape  1.773  1.481  2.123  0.163
## scale  7.334  5.766  9.328  0.900
##
## N = 911,  Events: 73,  Censored: 838
## Total time at risk: 1373.973
## Log-likelihood = -274.0475, df = 2
## AIC = 552.0951
```

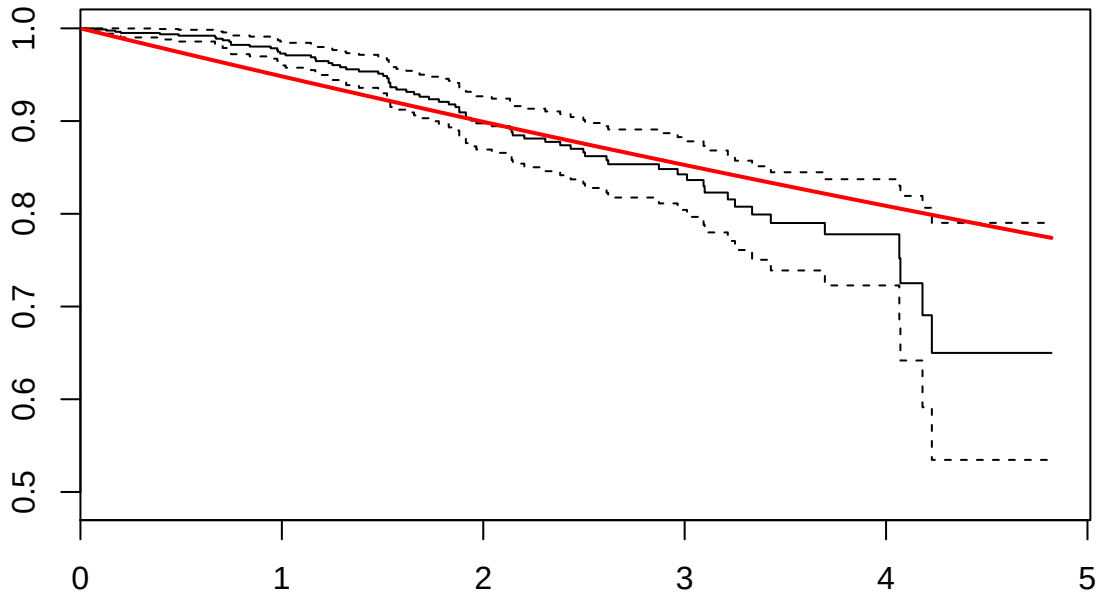
```
plot(llogis, ymin = .49, ci = FALSE)
```




```
exp = flexsurvreg(formula = Surv(time = years, event = event) ~ 1, data = datta, dist = "exp")
exp
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ 1,
##     data = datta, dist = "exp")
##
## Estimates:
##       est      L95%     U95%     se
## rate  0.05313  0.04224  0.06683  0.00622
##
## N = 911,  Events: 73,  Censored: 838
## Total time at risk: 1373.973
## Log-likelihood = -287.2552, df = 1
## AIC = 576.5104
```

```
plot(exp, ymin = .49, ci = FALSE)
```



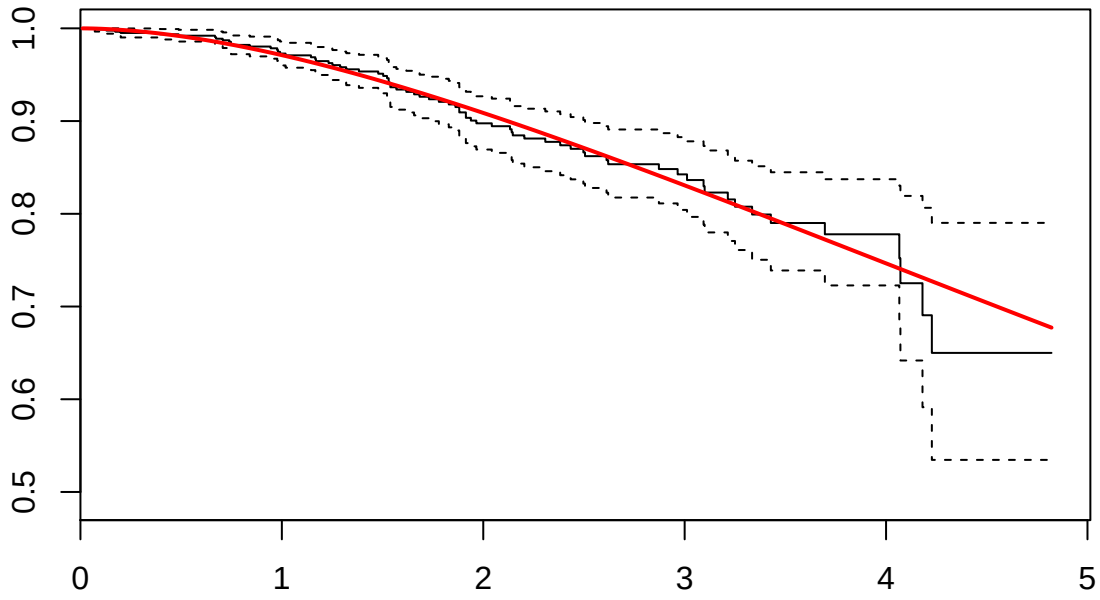

```
gamma = flexsurvreg(formula = Surv(time = years, event = event) ~ 1, data = datta, dist = "gamma")
```

```
## Warning in (function (q, shape, rate = 1, scale = 1/rate, lower.tail =
## TRUE, : NaNs produced
```

```
gamma
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ 1,
##     data = datta, dist = "gamma")
##
## Estimates:
##        est     L95%    U95%    se
## shape  1.8549  1.4828  2.3203  0.2119
## rate   0.2161  0.1379  0.3388  0.0496
##
## N = 911,  Events: 73,  Censored: 838
## Total time at risk: 1373.973
## Log-likelihood = -274.1304, df = 2
## AIC = 552.2607
```

```
plot(gamma, ymin = .49, ci = FALSE)
```




```
list_dist = list(weilbull, lnorm, llogis, exp)
aic = data.frame(AIC = sapply(list_dist, getElement, name = "AIC"))
loglik = data.frame(LogLikelihood = sapply(list_dist, getElement, name = "loglik"))

model_results_table = cbind(aic, loglik)

rownames(model_results_table) = c("Weilbull",
                    "Log-normal", "Log-logistic", "Exponential")

model_results_table = t(model_results_table)
```

?????????????? Do we need to incude any other goodness of fit BIC, AICC and do we need to include the gamma distribution as well??????????

```
kable(model_results_table, caption = "Log-likelihood and AIC of the different parametric models.")
```

Table 3: Log-likelihood and AIC of the different parametric models.

	Weilbull	Log-normal	Log-logistic	Exponential
AIC	551.9342	558.1154	552.0951	576.5104
LogLikelihood	-273.9671	-277.0577	-274.0475	-287.2552

Question: In stats we use multiple imputation and randomforest imputation (nonlinear relationships) to handle multiple imputation. In survival, should either of these be used, not used, or should very specifically created mi and rf methods be used, because of how survival, time to event, is?

Based on the AIC and log-likelihood values alone we would go with the Weillbull distribution because it has the smallest log-likelihood and AIC.

Now look at the Gamma method, suggested in slides 80 and 84 we look at.. and this would agree with ..

We looking at the correlation of the variables we will decide the covariates to be used.

4.2.2 Question 1 d

4.2.3 Question 1 e

We have chosen the weibull distribution model.

R gives us the shape and scale parameter estimates which correspond to the α and μ for the following survival function $S(t) = exp(-(t/\mu)^{\alpha})$.

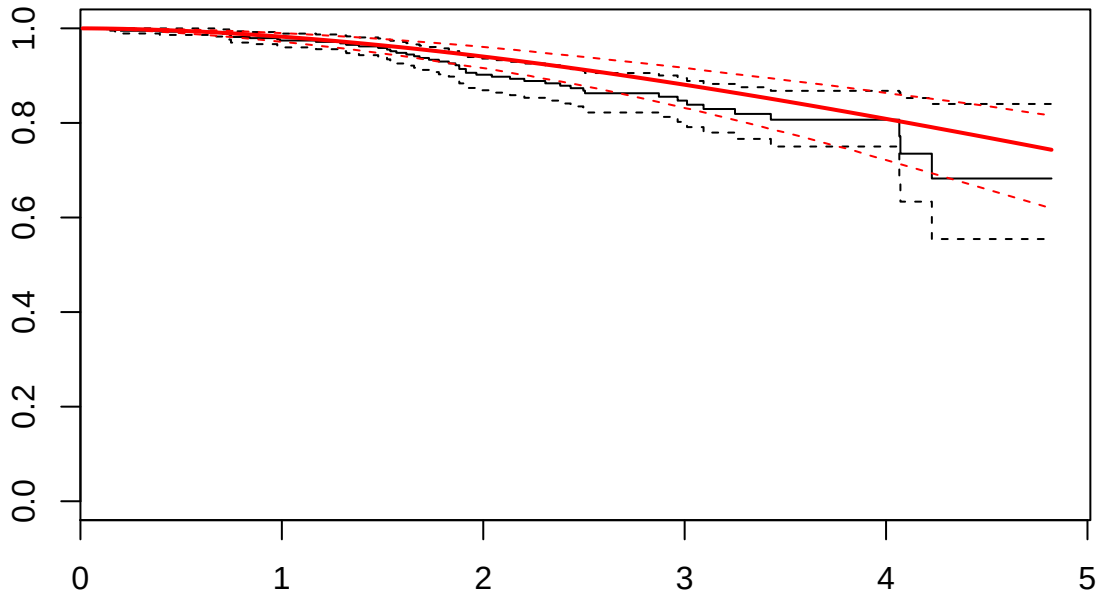
Therefore, we have the following correspondence between our parametrization and the parametrization in the lecture notes: $\lambda = 1/\mu$ and $\gamma = \alpha$ (Jackson 2016, pg 4).

Hence, the

```
#clarklevel +

final_model1 = flexsurvreg(formula = Surv(time = years, event = event) ~ ulceration + thickness, data =

final_model1
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ ulceration +
##     thickness, data = datta, dist = "weibull")
##
## Estimates:
##             data mean  est        L95%       U95%       se
## shape              NA   1.789133   1.441836   2.220085   0.197004
## scale              NA  13.891060   8.851872  21.798953   3.193702
## ulceration   0.332840  -0.785898  -1.181902  -0.389894   0.202047
## thickness    2.060680  -0.057105  -0.113335  -0.000875   0.028689
##             exp(est)   L95%       U95%
## shape              NA         NA         NA
## scale              NA         NA         NA
## ulceration   0.455710   0.306695   0.677129
## thickness    0.944495   0.892851   0.999126
##
## N = 676,  Events: 49,  Censored: 627
## Total time at risk: 1015.545
## Log-likelihood = -171.6334, df = 4
## AIC = 351.2668
```

```
plot(final_model1)
```




```
final_model2 = flexsurvreg(formula = Surv(time = years, event = event) ~ ulceration + thickness + clark

final_model2
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ ulceration +
##     thickness + clarklevel, data = datta, dist = "weibull")
##
## Estimates:
##                 data mean  est      L95%     U95%     se       exp(est)
## shape                NA     1.8175   1.4657   2.2537   0.1995       NA
## scale                NA    11.5402   3.7878  35.1588   6.5594       NA
## ulceration      0.3278    -0.7379  -1.1326  -0.3433   0.2013   0.4781
## thickness       2.0433    -0.0120  -0.0983   0.0742   0.0440   0.9880
## clarklevelII    0.1796     1.0018  -0.5455   2.5491   0.7895   2.7231
## clarklevelIII   0.2814     0.2356  -0.9202   1.3915   0.5897   1.2657
## clarklevelIV    0.4476    -0.1523  -1.2891   0.9845   0.5800   0.8587
## clarklevelV     0.0509    -0.2927  -1.6469   1.0616   0.6909   0.7463
##                 L95%     U95%
## shape                NA       NA
## scale                NA       NA
## ulceration      0.3222   0.7094
## thickness       0.9063   1.0771
## clarklevelII    0.5795  12.7951
## clarklevelIII   0.3984   4.0208
## clarklevelIV    0.2755   2.6765
```

```
## clarklevelV     0.1926   2.8909
##
## N = 668,  Events: 48,  Censored: 620
## Total time at risk: 1000.947
## Log-likelihood = -163.499, df = 8
## AIC = 342.998
```

```
# Why are the shape and scale parameters NA?

# How to determine when the covariates should be transformed. We know when the y and x should be in reg
```

4.2.4 Question 1 f

To assess the goodness of fit, we look at the log-likelihood ratio test. We compare the null model against the fitted model with covariates by subtracting their deviance. We know this difference in deviance follows a ditribution with degrees of freedom equal to the difference in the number of the parameters in the two models, which in our case is...

4.2.5 Question 1 g

4.3 Log-logistic Survival Model

4.3.1 Question 2 a

We will remove the people with missing ulceration data before, we fit a log-logistic model with covariate ulceration (excluding the missing cases).

```
# We check how many missing data there is for the ulceration column.
# sapply(datta, function(x) sum(is.na(x)))

# We will remove all of these.

datta1 = datta[!is.na(datta$ulceration), ]
datta1$ulceration = as.factor(datta1$ulceration)
# Check now how many missing data there is for the ulceration column.
# sapply(datta1, function(x) sum(is.na(x))) # None
```

We will not fit the model.

```
llogmodel = flexsurvreg(formula = Surv(time = years, event = event) ~
                          ulceration, data = datta1, dist = "llogis")

llogmodel
```

```
## Call:
## flexsurvreg(formula = Surv(time = years, event = event) ~ ulceration,
##     data = datta1, dist = "llogis")
##
## Estimates:
##              data mean  est     L95%    U95%    se      exp(est)  L95%
## shape            NA      1.913   1.539   2.379   0.213      NA        NA
## scale            NA     11.350   7.522  17.124   2.382      NA        NA
## ulceration1   0.333     -0.935  -1.309  -0.561   0.191   0.393     0.270
##              U95%
```

```
## shape            NA
## scale            NA
## ulceration1   0.571
##
## N = 676,  Events: 49,  Censored: 627
## Total time at risk: 1015.545
## Log-likelihood = -173.0183, df = 3
## AIC = 352.0367
```

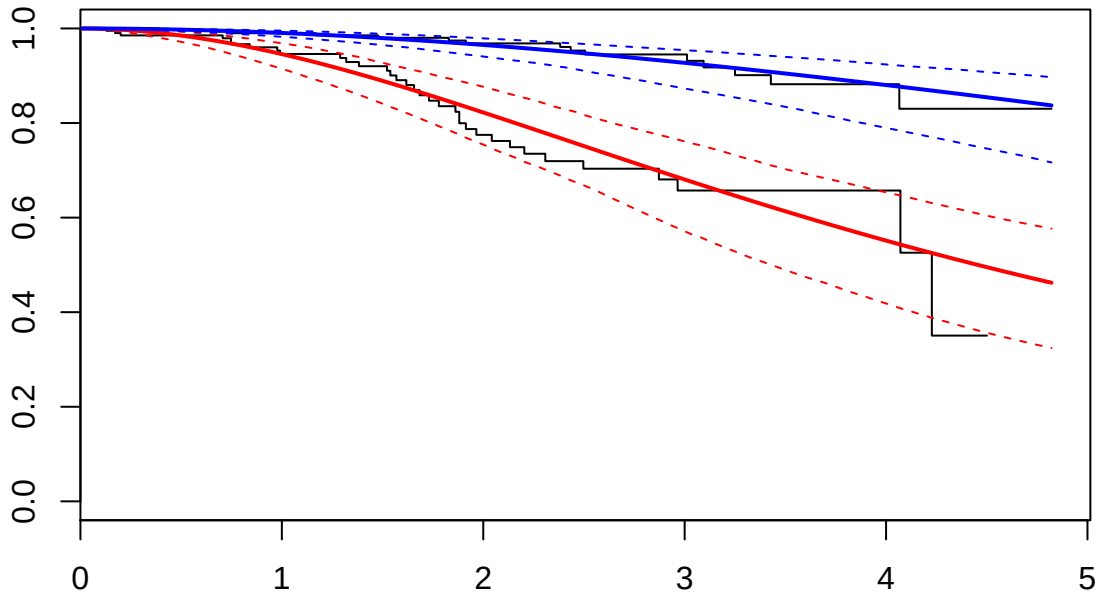
??????????????????????

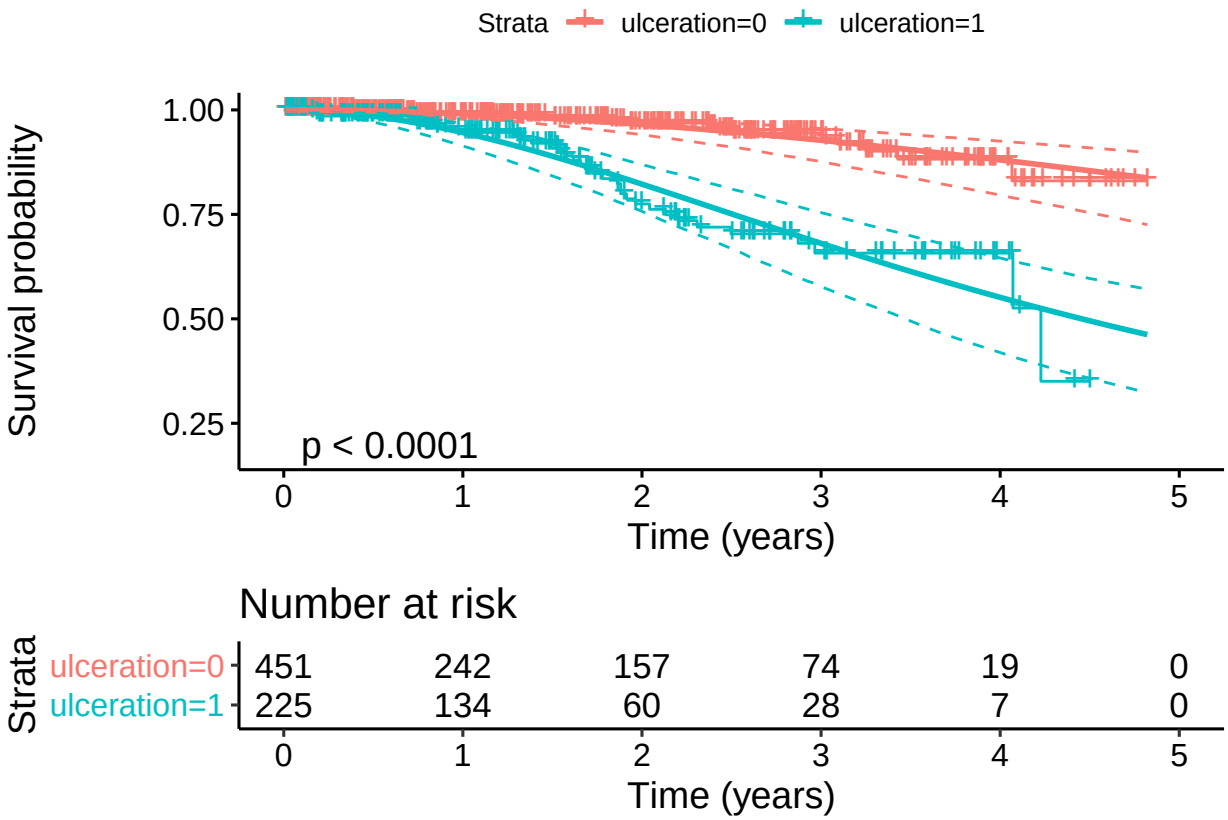
4.3.2 Question 2 b.

Produce two survival plots from this log-logistic model (i.e. one for ulceration=yes and one for ulceration=no).? ???????????? DOES SHE MEAN TWO SURVIVAL CURVES. (NOT TWO SEPERATE PLOTS?)

??????????????????

```
plot(llogmodel, ci = TRUE, col = c(2,4))

ggsurvplot(llogmodel, risk.table = TRUE, pval = TRUE, size = 1,
           #legend.title = "Ulceration",
           #legend.labs = c("No", "Yes"),
           conf.int  = TRUE,
           xlab = "Time (years)",
           risk.table.height = 0.3,
           ylim = c(.18,1))
```





4.3.3 Question 2 c. What is the estimated time ratio and odds ratio for survival and their 95% confidence intervals using the model parameter estimates (for ulceration compared to no ulceration)?

From the documentation of the flexsurv R package's paper (Jackson 2016) we have that the shape and scale parameters are α and μ with the survival function $S(t) = 1/(1 + (t/\mu)^{\alpha})$ which correspond to the parameters from the lecture notes as follows where the parameters from the lectures will be expressed as function of the R parameters: $\alpha = 1/\mu$ and $\gamma = \alpha$, where the later α is the scale parameter of the R output (Jackson 2016, 12).

All the following equations have the lecture notes' paramters.

Odds ratio (2 vs 1) $= \alpha_1/\alpha_2$

?????????????? What is time ratio ??????????

4.3.4 Question 2 d. Demonstrate the time ratio using the estimated median survival for each group. Are the estimated medians observed time points in the data?

The median is equal to $(1/\alpha)^{1/\gamma}$.

4.3.5 Question 2 e. Demonstrate the odds ratio using the estimated proportion surviving 3 years or more.

4.3.6 Question 2 f. How would you describe the time ratio and odds ratio to a member of the study team who does not have a background in statistics?

????????????Also need to plot Residual and diagnostic plots ????????????????????????

Need to present the model summaries as in lecture notes slide 126

5 Discussion

6 References

Jackson, Christopher H. 2016. "Flexsurv: A Platform for Parametric Survival Modeling in R." *Journal of Statistical Software* 70. Europe PMC Funders.

7 Appendix