

Practical Work

Lifetime Data Analysis

Rodrigo Arriaza, Alexander J Ohrt

07 desember, 2021

Introduction

We are given a data set on sexually transmitted diseases (STDs). This is data from a study about gonorrhea and chlamydia in 877 women. The objective with this practical work is to study possible risk factors for a reinfection with gonorrhea or chlamydia in women who have suffered one of both infections previously. The variables of interest are sociodemographic variables or those related to sexual practice. We have a lot of variables at our disposal, as can be seen from the data set below. We have chosen to use the following:

- Age: The age of the woman.
- NumPartners: The number of partners during the last 30 days.
- CondomUse: Use of condoms (1: always, 2: once in a while, 3: never)
- YearsSchool: Years of schooling.
- InitInfect: Initial infection (1: Gonorrhea, 2: Chlamydia, 3: both)
- InvVagAtExam: Involvement vagina at exam (1: yes; 0: no).
- DischargeExam: Discharge at exam (1: yes; 0: no)

The first two were chosen based on results from a study on gonorrhea reinfection in heterosexual STD clinic attendees. The study concluded that increased reinfection risk (of gonorrhea) was associated with younger age and a greater number of recent sex partners, among other risk factors. Moreover, the authors concluded that any type of condom use was a risk factor for reinfection with gonorrhea in women. However by using statistical analysis we found that the Age of the woman and the ethnicity (a variable that was not included in the aforementioned study) are not statistically significant (Necessary to mention Ethnicity here, if it is not a part of the study?). Q: Should perhaps describe what we have done and why it has been done? I am going to ask Klaus about what he thinks about it also I think, just to see what he says.

Another publication reports that, on average, 14% of women with clamydia and 12% of women with gonorrhea get reinfected, with younger women at higher risk. Moreover, they state that many adolescents treated for infection of one of the two STDs are reinfected within three to six months, usually because of resumed sexual contact with an untreated partner. Thus, the marital status might be interesting to analyse. However, this is not added, because, as seen in the exploratory data analysis below, the ages are low, which should mean that the amount in each level of **MaritalStatus** is very skewed towards single. This can be seen in the table below as well. NOT SURE IF WE SHOULD KEEP THIS, DEPENDS ON THE AMOUNT OF ROOM I GUESS. can leave it until the end.

This meta-analysis reports that the relationship between race, socioeconomic status (SES) and chlamydial infection is not clear. It concludes that SES was not associated with chlamydia infection, where they tested for several variables, where level of parent's education was one of them. Either way, we think it might be interesting to see if the years of schooling of the women have any impact on chlamydia reinfection and as is shown below it showed to be statistically significant during the exploratory analysis. Q: I also think that this would be a reason to look at race also (combined with the first study, where they only had afro-americans. Perhaps we can try to mix the variables chosen based on studies and on the statistical method? (that is; if this statistical method is something we can trust))

```
std_data <- read.table("STD_onlydata.txt")
colnames(std_data) <- variable.names <- c("ObsNum", "Ethnicity", "MaritalStatus", "Age", "YearsSchool",

std_data[, variable.names[-c(1, 4, 5, 7,23,24)]] <- lapply(std_data[, variable.names[-c(1,4,5,7,23,24)]]
std_data$cens <- rep(1, length = dim(std_data)[1]) # All uncensored, thus cens = 1 for all rows.

variables.chosen <- c("Age", "NumPartners", "CondomUse", "Ethnicity") # ETC!
continuous.variables <- unlist(lapply(std_data, is.numeric))
continuous.variables <- continuous.variables[-length(continuous.variables)] # Removing the cens-value.
```

Naturally, the categorical variable which states if the woman is reinfected or not (Reinfection) will be used as a dependent variable in the analysis and the time until reinfection since the more time a subject is under study, the greater the risk of the event reoccurring.

Descriptive Analysis

Variable selection

Q: What is the offset(log(TimeUntilReInf))?

```
std_data$Ethnicity <- factor(std_data$Ethnicity)
nb.model <- MASS::glm.nb(Reinfection ~ Ethnicity + MaritalStatus + Age + YearsSchool + InitInfect +
                          NumPartners + OralSex12m + OralSex30d + RectalSex12m + RectalSex30d + AbPain +
                          SignDischarge + SignDysuria+ CondomUse + SignItch + SignLesion + SignRash +
                          SignLymph + InvVagAtExam + DischargeExam + AbnormNodeExam + offset(log(TimeUntilReinf)),
                          data=std_data)
s <- summary(nb.model)
k <- knitr::kable(s$coefficients, caption = 'Variables statistical significance')
kableExtra::row_spec(k, c(6:9,14,18,23,24), color='white', background = 'cyan')
```

Note that age and years of schooling are somewhat correlated. Could do some more EDA probably, and should remove some of this also.

Nonparametric Analysis

The survival curve is estimated by means of Kaplan-Meier and plotted below. The curve below shows the general survival in the data set, i.e. not depending on any specific explanatory variable.

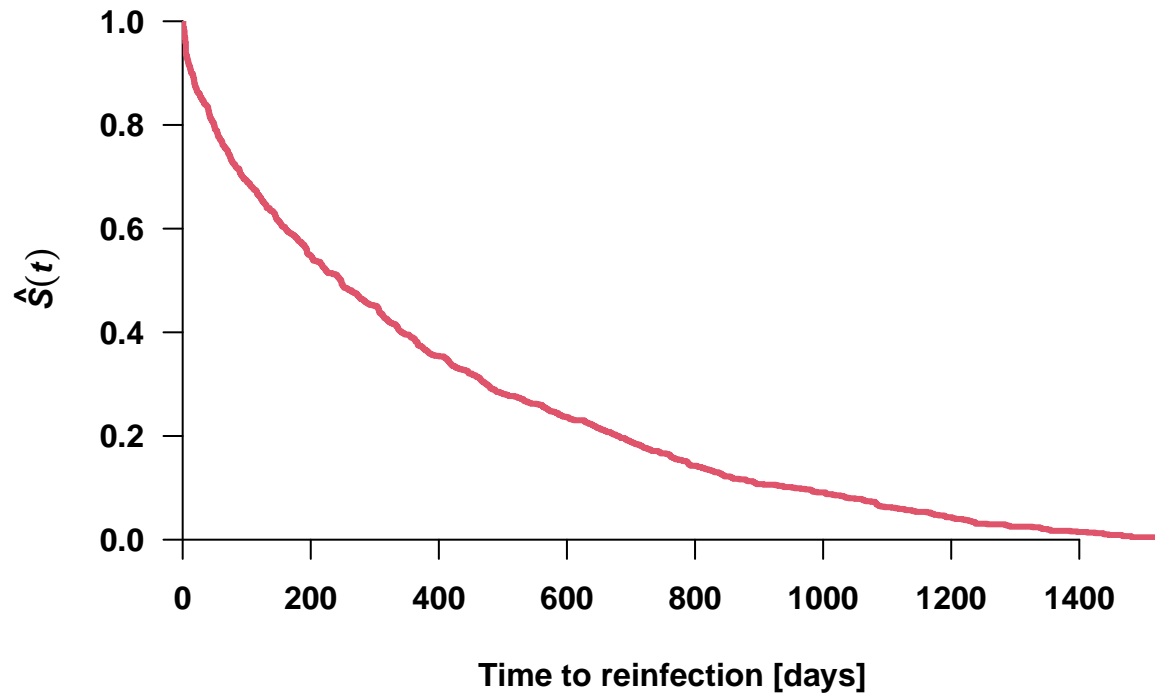
```
s1 <- with(std_data, Surv(TimeUntilReinf, cens) ~ 1)
s1fit <- survfit(s1)
#plot(s1fit)

par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(s1fit, col = 2, xlab = "Time to reinfection [days]",
     ylab = expression(bolditalic(hat(S)(t))),
     lty = 1, lwd = 3, yaxs = "i", xaxs = "i", bty = "n",
     conf.int = F)
title("Survival Function")
```

Table 1: Variables statistical significance

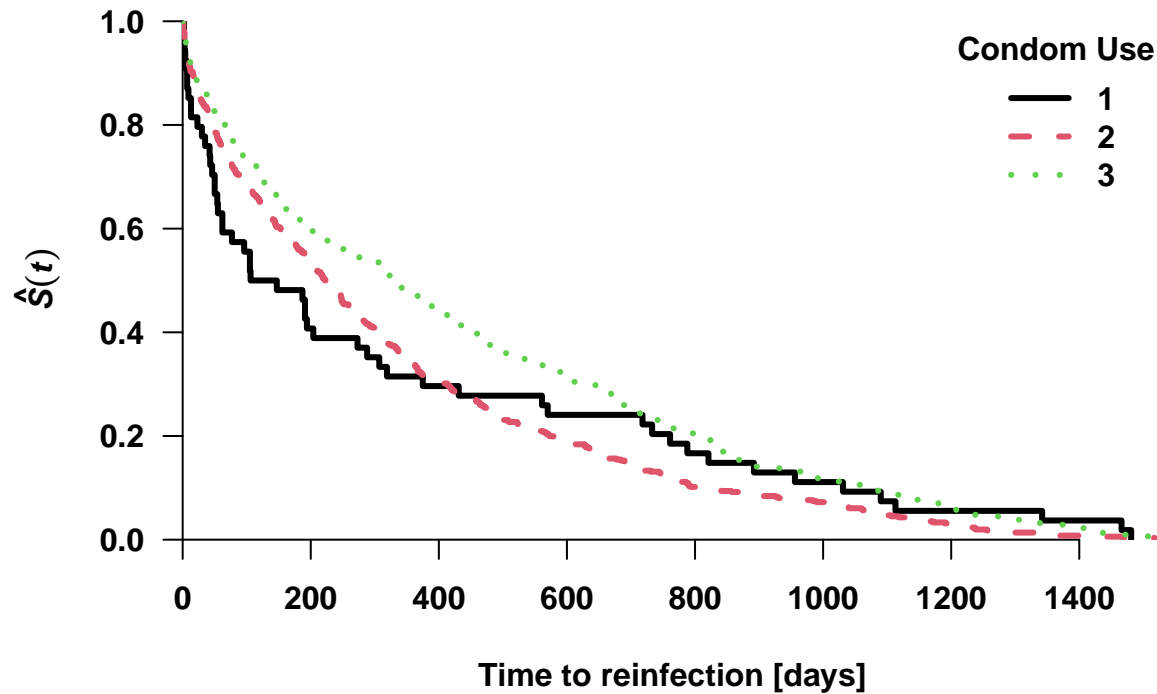
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.4766626	0.6791361	-6.5917022	0.0000000
EthnicityW	-0.0786114	0.1576156	-0.4987540	0.6179527
MaritalStatusM	0.1142920	0.4681139	0.2441542	0.8071114
MaritalStatusS	0.5011754	0.3203698	1.5643657	0.1177317
Age	0.0188481	0.0156143	1.2071011	0.2273932
YearsSchool	-0.1689015	0.0442657	-3.8156308	0.0001358
InitInfect2	-0.3302518	0.1740868	-1.8970524	0.0578210
InitInfect3	-0.3318821	0.1755787	-1.8902183	0.0587288
NumPartners	0.1164568	0.0598373	1.9462257	0.0516276
OralSex12m1	-0.3703474	0.2387666	-1.5510855	0.1208812
OralSex30d1	-0.3246975	0.2643311	-1.2283739	0.2193066
RectalSex12m1	0.0669703	0.4881503	0.1371920	0.8908790
RectalSex30d1	-0.1627456	0.6172379	-0.2636675	0.7920361
AbPain1	0.2969178	0.1771403	1.6761734	0.0937042
SignDischarge1	0.1330009	0.1306664	1.0178660	0.3087416
SignDysuria1	0.1954606	0.1812469	1.0784219	0.2808455
CondomUse2	-0.1553543	0.2725108	-0.5700849	0.5686201
CondomUse3	-0.4582270	0.2819913	-1.6249684	0.1041693
SignItch1	-0.2209724	0.1750560	-1.2622958	0.2068424
SignLesion1	-0.2541307	0.3787052	-0.6710513	0.5021878
SignRash1	-0.0638066	0.4592994	-0.1389215	0.8895122
SignLymph1	0.2368538	0.5922357	0.3999317	0.6892069
InvVagAtExam1	0.5726933	0.2003764	2.8580874	0.0042620
DischargeExam1	-0.5805191	0.2691414	-2.1569301	0.0310111
AbnormNodeExam1	0.0801562	0.5157541	0.1554155	0.8764938

Survival Function



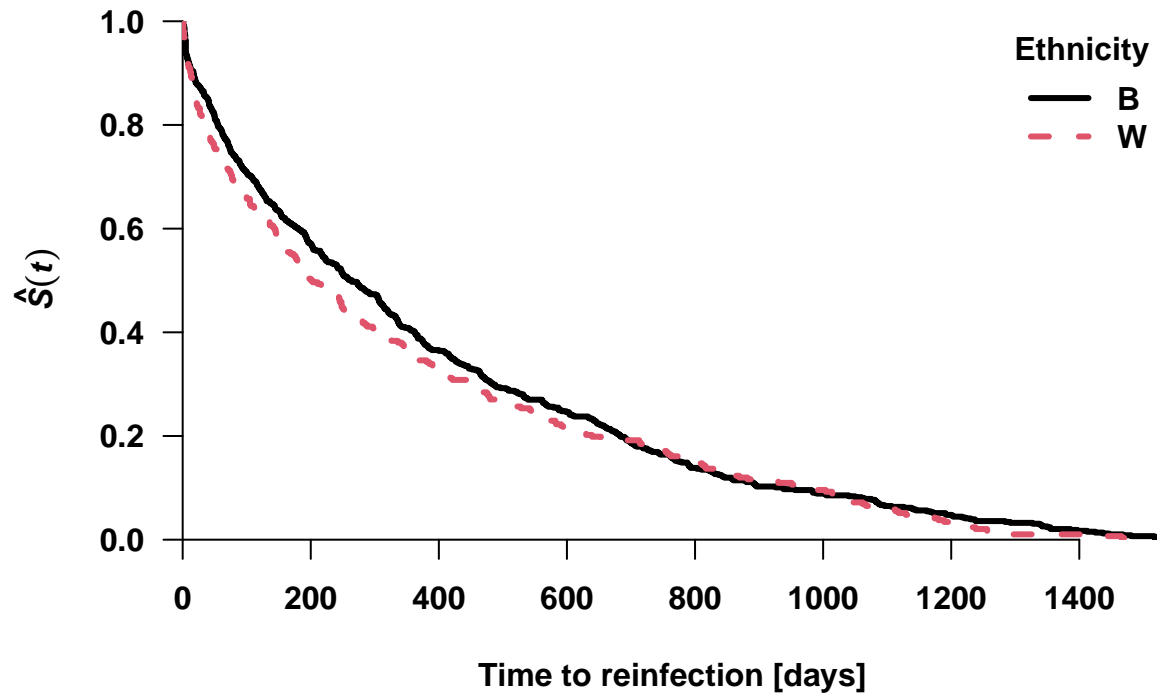
```
s1.CondomUse <- with(std_data, Surv(TimeUntilReinf, cens) ~ CondomUse)
s1fit.CondomUse <- survfit(s1.CondomUse)
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(s1fit.CondomUse, col = c(1,2,3), xlab = "Time to reinfection [days]",
     ylab = expression(bolditalic(hat(S)(t))),
     yaxs = "i", xaxs = "i", bty = "n", lty = c(1,2,3), lwd = rep(3, length.out = 3),
     conf.int = F)
legend("topright", legend = levels(std_data$CondomUse), title = "Condom Use",
     bty = "n", col = c(1,2,3), lty = c(1,2,3), lwd = rep(3, length.out = 3))
title("Survival Function as Function of Condom Use")
```

Survival Function as Function of Condom Use



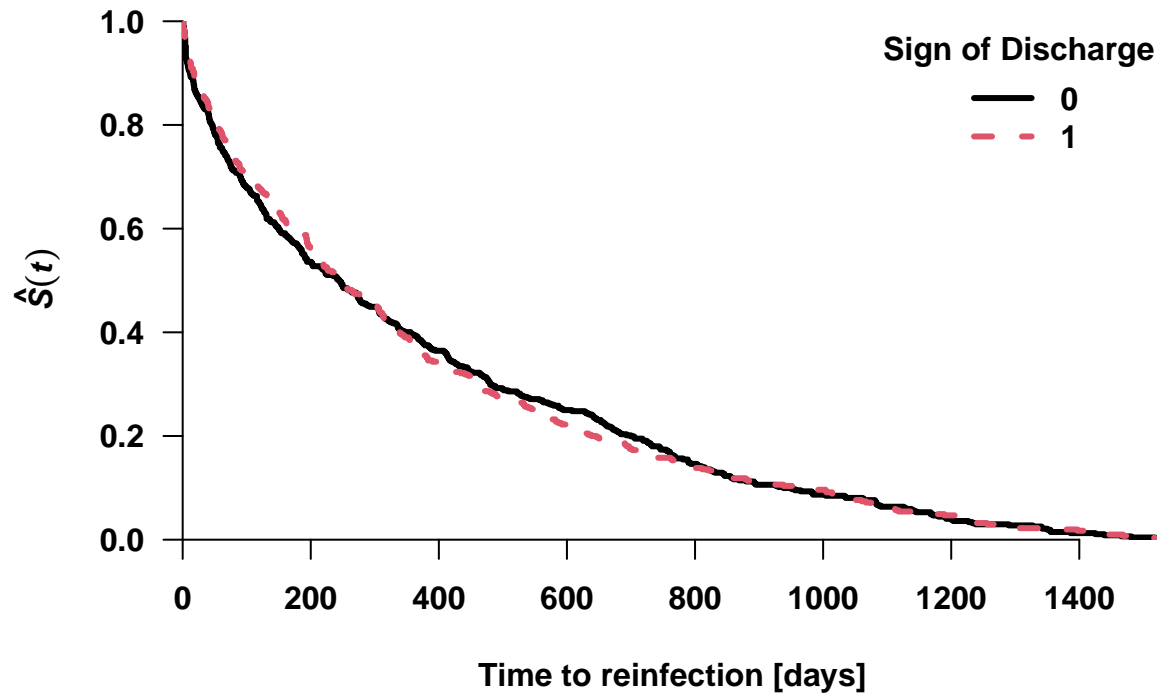
```
s1.Ethnicity <- with(std_data, Surv(TimeUntilReinf, cens) ~ Ethnicity)
s1fit.Ethnicity <- survfit(s1.Ethnicity)
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(s1fit.Ethnicity, col = c(1,2), xlab = "Time to reinfection [days]",
     ylab = expression(bolditalic(hat(S)(t))),
     yaxs = "i", xaxs = "i", bty = "n", lty = c(1,2), lwd = rep(3, length.out = 2),
     conf.int = F)
legend("topright", legend = levels(std_data$Ethnicity), title = "Ethnicity",
     bty = "n", col = c(1,2), lty = c(1,2), lwd = rep(3, length.out = 2))
title("Survival Function as Function of Ethnicity")
```

Survival Function as Function of Ethnicity



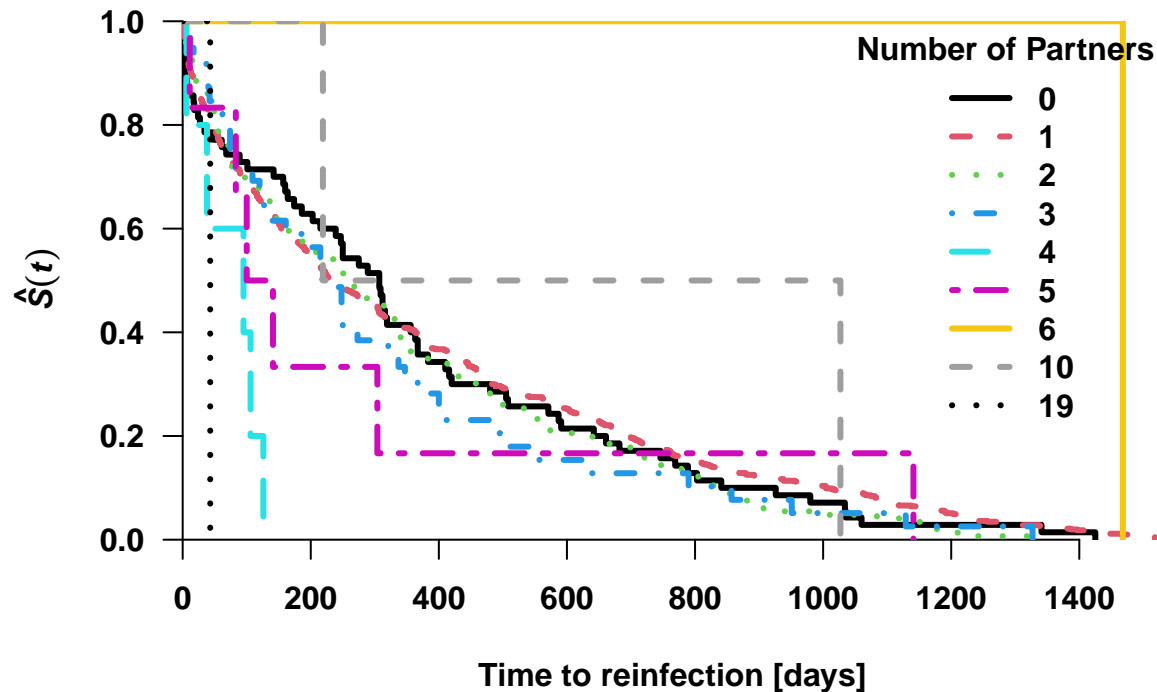
```
s1.SignDischarge <- with(std_data, Surv(TimeUntilReinf, cens) ~ SignDischarge)
s1fit.SignDischarge <- survfit(s1.SignDischarge)
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(s1fit.SignDischarge, col = c(1,2), xlab = "Time to reinfection [days]",
     ylab = expression(bolditalic(hat(S)(t))),
     yaxs = "i", xaxs = "i", bty = "n", lty = c(1,2), lwd = rep(3, length.out = 2),
     conf.int = F)
legend("topright", legend = levels(std_data$SignDischarge), title = "Sign of Discharge",
     bty = "n", col = c(1,2), lty = c(1,2), lwd = rep(3, length.out = 2))
title("Survival Function as Function of Sign of Discharge")
```

Survival Function as Function of Sign of Discharge



```
s1.NumPartners <- with(std_data, Surv(TimeUntilReinf, cens) ~ NumPartners)
s1fit.NumPartners <- survfit(s1.NumPartners)
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(s1fit.NumPartners, col = c(1,2,3,4,5,6,7,8,9), xlab = "Time to reinfection [days]",
     ylab = expression(bolditalic(hat(S)(t))),
     yaxs = "i", xaxs = "i", bty = "n", lty = c(1,2,3,4,5,6,7,8,9), lwd = rep(3, length.out = 9),
     conf.int = F)
legend("topright", legend = levels(as.factor(std_data$NumPartners)), title = "Number of Partners",
     bty = "n", col = c(1,2,3,4,5,6,7,8,9), lty = c(1,2,3,4,5,6,7,8,9), lwd = rep(3, length.out = 9))
title("Survival Function as Function of Number of Partners")
```

Survival Function as Function of Number of Partners



The median survival time is 247 days, as can be seen below.

```
s1fit
```

```
#> Call: survfit(formula = s1)
#>
#>      n events median 0.95LCL 0.95UCL
#> [1,] 877      877    247     216     277
```

Comparison of survival functions by means of nonparametric tests, such as the logrank test. The curve on ethnicity could be interesting to test (the survivals cross though!)

Below, logrank test for all curves plotted above are done.

```
FHtestrcc(s1.CondomUse)
```

```
#>
#> K-sample test for right-censored data
#>
#> Parameters: rho=0, lambda=0
#> Distribution: counting process approach
#>
#> Data: Surv(TimeUntilReinf, cens) by CondomUse
#>
#>      N Observed Expected      O-E (O-E)^2/E (O-E)^2/V
#> CondomUse=1  54      54    51.8   2.16   0.0902   0.0968
#> CondomUse=2 511     511   457.3  53.66   6.2948  13.4467
#> CondomUse=3 312     312   367.8 -55.82   8.4704  14.8506
#>
#> Chisq= 15.1 on 2 degrees of freedom, p-value= 0.000514
#> Alternative hypothesis: survival functions not equal
```



```
FHtestrcc(s1.Ethnicity)
```

```
#>
#> Two-sample test for right-censored data
#>
#> Parameters: rho=0, lambda=0
#> Distribution: counting process approach
#>
#> Data: Surv(TimeUntilReinf, cens) by Ethnicity
#>
#>           N Observed Expected   O-E (O-E)^2/E (O-E)^2/V
#> Ethnicity=B 585      585     603 -18.3   0.555     1.8
#> Ethnicity=W 292      292     274  18.3   1.224     1.8
#>
#> Statistic Z= 1.3, p-value= 0.18
#> Alternative hypothesis: survival functions not equal
```

```
FHtestrcc(s1.SignDischarge)
```

```
#>
#> Two-sample test for right-censored data
#>
#> Parameters: rho=0, lambda=0
#> Distribution: counting process approach
#>
#> Data: Surv(TimeUntilReinf, cens) by SignDischarge
#>
#>           N Observed Expected   O-E (O-E)^2/E (O-E)^2/V
#> SignDischarge=0 472      472     472 -0.461 0.000450 0.00098
#> SignDischarge=1 405      405     405  0.461 0.000525 0.00098
#>
#> Statistic Z= 0, p-value= 0.975
#> Alternative hypothesis: survival functions not equal
```

```
FHtestrcc(s1.NumPartners)
```

```
#>
#> Trend FH test for right-censored data
#>
#> Parameters: rho=0, lambda=0
#> Distribution: counting process approach
#>
#> Data: Surv(TimeUntilReinf, cens) by NumPartners
#>
#>           N Observed Expected   O-E
#> NumPartners=0   70      70   67.796  2.204
#> NumPartners=1  607     607  626.599 -19.599
#> NumPartners=2  146     146  134.200  11.800
#> NumPartners=3   39      39   33.838   5.162
#> NumPartners=4    5       5    1.417   3.583
#> NumPartners=5    6       6    4.944   1.056
#> NumPartners=6    1       1    4.895  -3.895
#> NumPartners=10   2       2    3.104  -1.104
#> NumPartners=19   1       1    0.206   0.794
#>
```

```
#> Statistic Z= 0.7, p-value= 0.488
#> Alternative hypothesis: survival functions not equal
```

Fit of a parametric survival model

Fit a Weibull, log-logistic or lognormal model. Justify the theoretical choice we are taking!!

I am testing all three below, since I am not sure which will be better.

Below we have the null models.

```
s2 <- with(std_data, Surv(TimeUntilReinf, cens))
loglo <- survreg(s2 ~ 1, dist = "loglo")
summary(loglo)
```

```
#>
#> Call:
#> survreg(formula = s2 ~ 1, dist = "loglo")
#>
#> Value Std. Error z p
#> (Intercept) 5.2914 0.0532 99.52 < 2e-16
#> Log(scale) -0.0995 0.0283 -3.51 0.00045
#>
#> Scale= 0.905
#>
#> Log logistic distribution
#> Loglik(model)= -6139.9 Loglik(intercept only)= -6139.9
#> Number of Newton-Raphson Iterations: 6
#> n= 877
```

```
weibull <- survreg(s2 ~ 1)
summary(weibull)
```

```
#>
#> Call:
#> survreg(formula = s2 ~ 1)
#>
#> Value Std. Error z p
#> (Intercept) 5.8289 0.0423 137.94 < 2e-16
#> Log(scale) 0.1762 0.0277 6.37 1.9e-10
#>
#> Scale= 1.19
#>
#> Weibull distribution
#> Loglik(model)= -6040.1 Loglik(intercept only)= -6040.1
#> Number of Newton-Raphson Iterations: 7
#> n= 877
```

```
lognormal <- survreg(s2 ~ 1, dist = "lognormal")
summary(lognormal)
```

```
#>
#> Call:
#> survreg(formula = s2 ~ 1, dist = "lognormal")
#>
#> Value Std. Error z p
#> (Intercept) 5.0988 0.0553 92.2 <2e-16
#> Log(scale) 0.4934 0.0239 20.7 <2e-16
#>
#> Scale= 1.64
```

```

#>
#> Log Normal distribution
#> Loglik(model)= -6148.8   Loglik(intercept only)= -6148.8
#> Number of Newton-Raphson Iterations: 5
#> n= 877

# Below we add some covariates to the models. INCLUDE ALL OF THEM PERHAPS! (OR TEST WITH DIFFERENT ONES)
# First we add Ethnicity e.g.
loglo.Ethnicity <- survreg(s2 ~ Ethnicity, dist = "loglo", data = std_data)
summary(loglo.Ethnicity)

#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity, data = std_data, dist = "loglo")
#>               Value Std. Error      z      p
#> (Intercept)  5.3566      0.0642 83.38 < 2e-16
#> EthnicityW  -0.2016      0.1133 -1.78 0.07520
#> Log(scale)  -0.1013      0.0283 -3.58 0.00035
#>
#> Scale= 0.904
#>
#> Log logistic distribution
#> Loglik(model)= -6138.4   Loglik(intercept only)= -6139.9
#>  Chisq= 3.17 on 1 degrees of freedom, p= 0.075
#> Number of Newton-Raphson Iterations: 4
#> n= 877

weibull.Ethnicity <- survreg(s2 ~ Ethnicity, data = std_data)
summary(weibull.Ethnicity)

#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity, data = std_data)
#>               Value Std. Error      z      p
#> (Intercept)  5.8606      0.0508 115.41 <2e-16
#> EthnicityW  -0.0973      0.0855  -1.14  0.25
#> Log(scale)   0.1759      0.0276   6.36 2e-10
#>
#> Scale= 1.19
#>
#> Weibull distribution
#> Loglik(model)= -6039.4   Loglik(intercept only)= -6040.1
#>  Chisq= 1.29 on 1 degrees of freedom, p= 0.26
#> Number of Newton-Raphson Iterations: 6
#> n= 877

lognormal.Ethnicity <- survreg(s2 ~ Ethnicity, dist = "lognormal", data = std_data)
summary(lognormal.Ethnicity)

#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity, data = std_data, dist = "lognormal")
#>               Value Std. Error      z      p
#> (Intercept)  5.1681      0.0676 76.45 <2e-16
#> EthnicityW  -0.2080      0.1171 -1.78 0.076
#> Log(scale)   0.4916      0.0239 20.59 <2e-16

```

```

#>
#> Scale= 1.63
#>
#> Log Normal distribution
#> Loglik(model)= -6147.2   Loglik(intercept only)= -6148.8
#>  Chisq= 3.15 on 1 degrees of freedom, p= 0.076
#> Number of Newton-Raphson Iterations: 2
#> n= 877

loglo.full <- survreg(s2 ~ Ethnicity + Age + NumPartners + CondomUse + YearsSchool + SignDischarge, data = std_data, dist = "loglo")
summary(loglo.full)

```

```

#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity + Age + NumPartners + CondomUse +
#>   YearsSchool + SignDischarge, data = std_data, dist = "loglo")
#>
#>               Value Std. Error      z      p
#> (Intercept)    4.43176    0.43979 10.08 <2e-16
#> EthnicityW    -0.20875    0.11403  -1.83 0.0672
#> Age           -0.00381    0.01125  -0.34 0.7346
#> NumPartners   -0.01960    0.04942  -0.40 0.6917
#> CondomUse2     0.34934    0.23892   1.46 0.1437
#> CondomUse3     0.67345    0.24619   2.74 0.0062
#> YearsSchool    0.04901    0.03487   1.41 0.1599
#> SignDischarge1 0.05863    0.10562   0.56 0.5788
#> Log(scale)    -0.11108    0.02836  -3.92 9e-05
#>

```

```

#> Scale= 0.895
#>
#> Log logistic distribution
#> Loglik(model)= -6130.8   Loglik(intercept only)= -6139.9
#>  Chisq= 18.39 on 7 degrees of freedom, p= 0.01
#> Number of Newton-Raphson Iterations: 4
#> n= 877

```

```

loglo.pred <- predict(loglo.full, type = "linear")
resids.loglo <- (log(std_data$TimeUntilReinf) - loglo.pred) / loglo.full$scale

weibull.full <- update(loglo.full, dist = "weibull")
summary(weibull.full)

```

```

#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity + Age + NumPartners + CondomUse +
#>   YearsSchool + SignDischarge, data = std_data, dist = "weibull")
#>
#>               Value Std. Error      z      p
#> (Intercept)    5.310720    0.328119 16.19 < 2e-16
#> EthnicityW    -0.110680    0.085688  -1.29   0.20
#> Age           -0.001799    0.008523  -0.21   0.83
#> NumPartners   -0.010662    0.040943  -0.26   0.79
#> CondomUse2     0.009896    0.169834   0.06   0.95
#> CondomUse3     0.285457    0.175474   1.63   0.10
#> YearsSchool    0.043070    0.026987   1.60   0.11
#> SignDischarge1 -0.000572    0.080803  -0.01   0.99
#> Log(scale)     0.166530    0.027666   6.02 1.8e-09

```

```

#>
#> Scale= 1.18
#>
#> Weibull distribution
#> Loglik(model)= -6031.9   Loglik(intercept only)= -6040.1
#>   Chisq= 16.29 on 7 degrees of freedom, p= 0.023
#> Number of Newton-Raphson Iterations: 6
#> n= 877

weibull.pred <- predict(weibull.full, type = "linear")
resids.weibull <- (log(std_data$TimeUntilReinf) - weibull.pred) / weibull.full$scale

lognormal.full <- update(loglo.full, dist = "lognormal")
summary(lognormal.full)

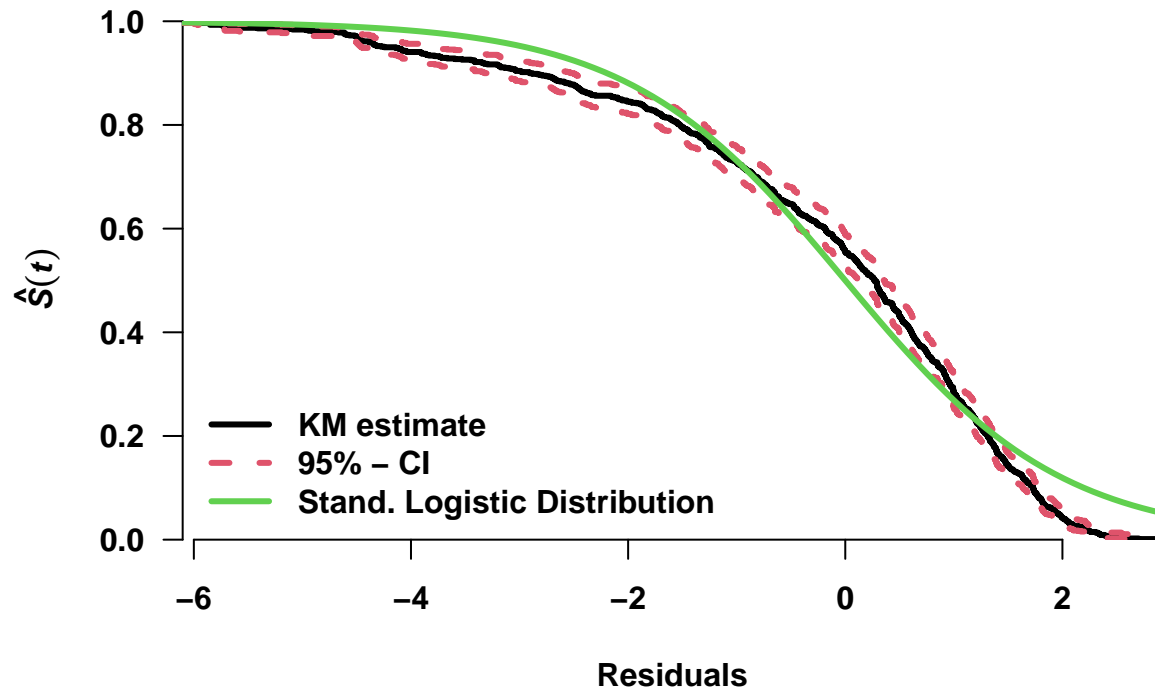
#>
#> Call:
#> survreg(formula = s2 ~ Ethnicity + Age + NumPartners + CondomUse +
#>   YearsSchool + SignDischarge, data = std_data, dist = "lognormal")
#>
#>               Value Std. Error      z      p
#> (Intercept)    4.1690    0.4468  9.33 <2e-16
#> EthnicityW    -0.2198    0.1177 -1.87 0.0620
#> Age           -0.0018    0.0115 -0.16 0.8763
#> NumPartners    0.0172    0.0536  0.32 0.7484
#> CondomUse2     0.3114    0.2331  1.34 0.1815
#> CondomUse3     0.6232    0.2412  2.58 0.0098
#> YearsSchool    0.0510    0.0359  1.42 0.1554
#> SignDischarge1 0.0708    0.1107  0.64 0.5224
#> Log(scale)     0.4841    0.0239 20.27 <2e-16
#>
#> Scale= 1.62
#>
#> Log Normal distribution
#> Loglik(model)= -6140.6   Loglik(intercept only)= -6148.8
#>   Chisq= 16.34 on 7 degrees of freedom, p= 0.022
#> Number of Newton-Raphson Iterations: 3
#> n= 877

lognormal.pred <- predict(lognormal.full, type = "linear")
resids.logno <- (log(std_data$TimeUntilReinf) - lognormal.pred) / lognormal.full$scale

# Model checking (using residuals etc, as done in lab 6).
# Log-logistic
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(survfit(Surv(resids.loglo, std_data$cens) ~ 1), col = c(1,2,2), xlab = "Residuals",
     ylab = expression(bolditalic(hat(S)(t))),
     lty = 1, lwd = 3, yaxs = "i", xaxs = "i", bty = "n")
title("Residuals of the log-logistic Regression Model")
curve(plogis(x, lower.tail = F), from = min(resids.loglo), to = max(resids.loglo), col = 3, lwd = 3,
     add = TRUE)
legend("bottomleft", c("KM estimate", "95% - CI", "Stand. Logistic Distribution"),
     col = c(1, 2, 3), lty = c(1, 2, 1), lwd = 3, bty = "n")

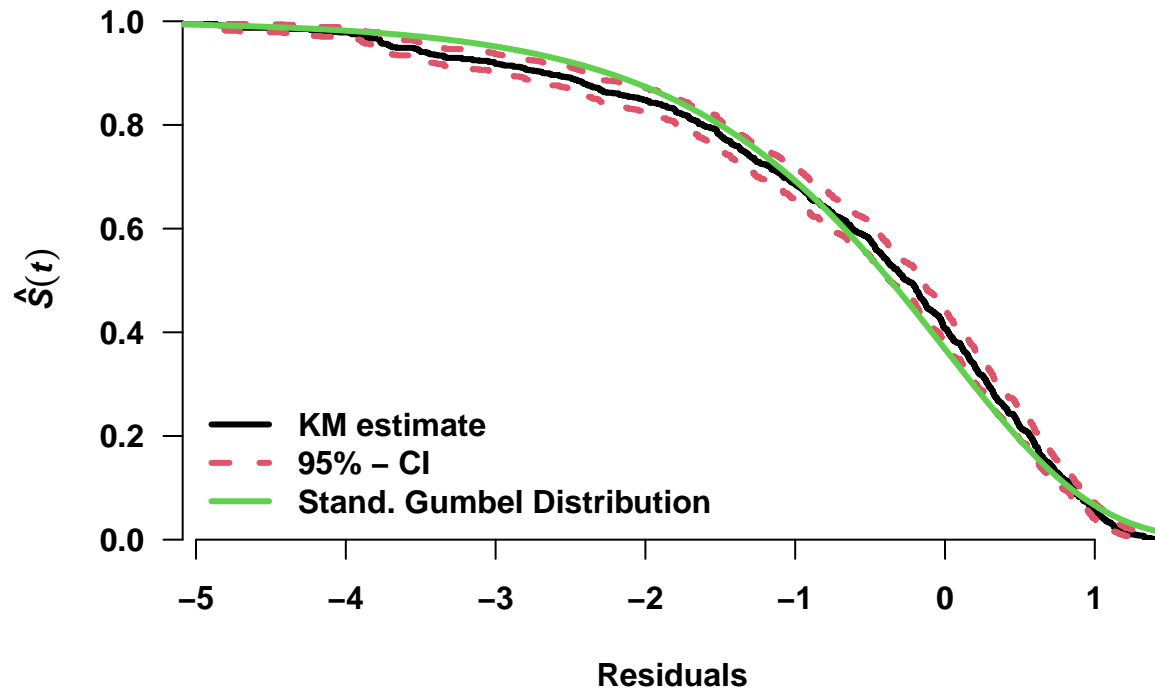
```

Residuals of the log-logistic Regression Model



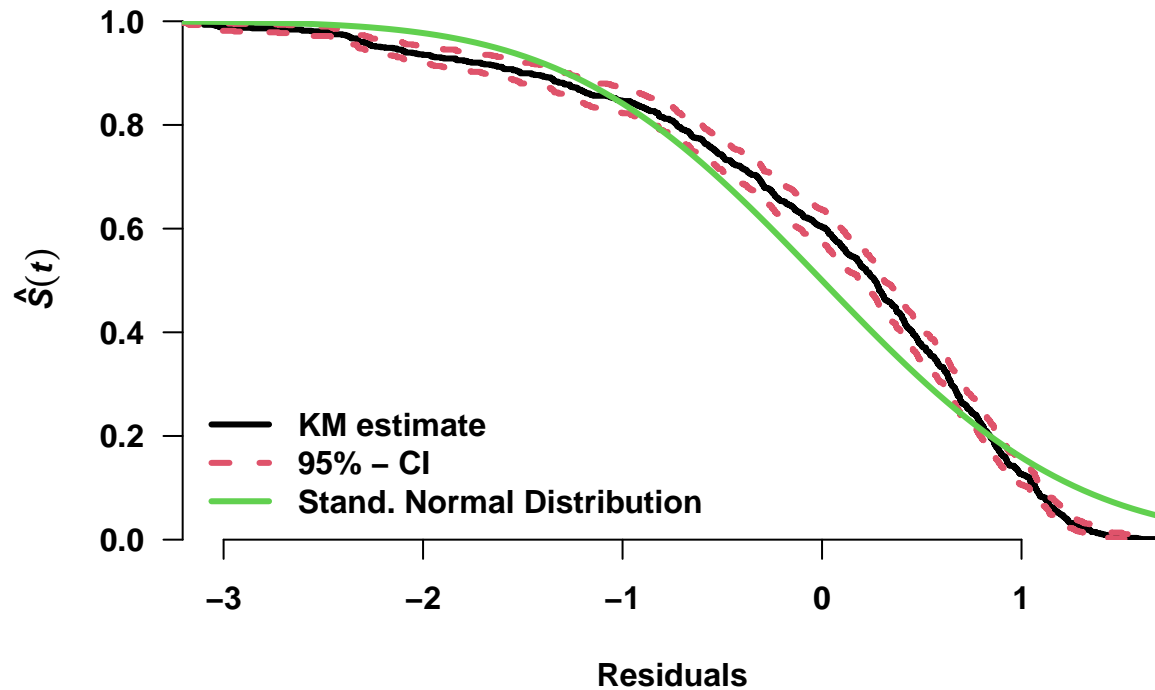
```
# Weibull
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(survfit(Surv(resids.weibull, std_data$cens) ~ 1), col = c(1,2,2), xlab = "Residuals",
     ylab = expression(bolditalic(hat(S)(t))),
     lty = 1, lwd = 3, yaxs = "i", xaxs = "i", bty = "n")
title("Residuals of the Weibull Regression Model")
survgumb <- function(x) {
  return(exp(-exp(x)))
}
curve(survgumb, from = min(resids.weibull), to = max(resids.weibull), col = 3, lwd = 3,
      add = TRUE)
legend("bottomleft", c("KM estimate", "95% - CI", "Stand. Gumbel Distribution"),
      col = c(1, 2, 3), lty = c(1, 2, 1), lwd = 3, bty = "n")
```

Residuals of the Weibull Regression Model



```
# Log-normal
par(font = 2, font.axis = 2, font.lab = 2, las = 1, mar = c(5, 5, 4, 2))
plot(survfit(Surv(resids.logno, std_data$cens) ~ 1), col = c(1,2,2), xlab = "Residuals",
     ylab = expression(bolditalic(hat(S)(t))),
     lty = 1, lwd = 3, yaxs = "i", xaxs = "i", bty = "n")
title("Residuals of the Lognormal Regression Model")
curve(pnorm(x, lower.tail = F), from = min(resids.logno), to = max(resids.logno), col = 3, lwd = 3,
     add = TRUE)
legend("bottomleft", c("KM estimate", "95% - CI", "Stand. Normal Distribution"),
     col = c(1, 2, 3), lty = c(1, 2, 1), lwd = 3, bty = "n")
```

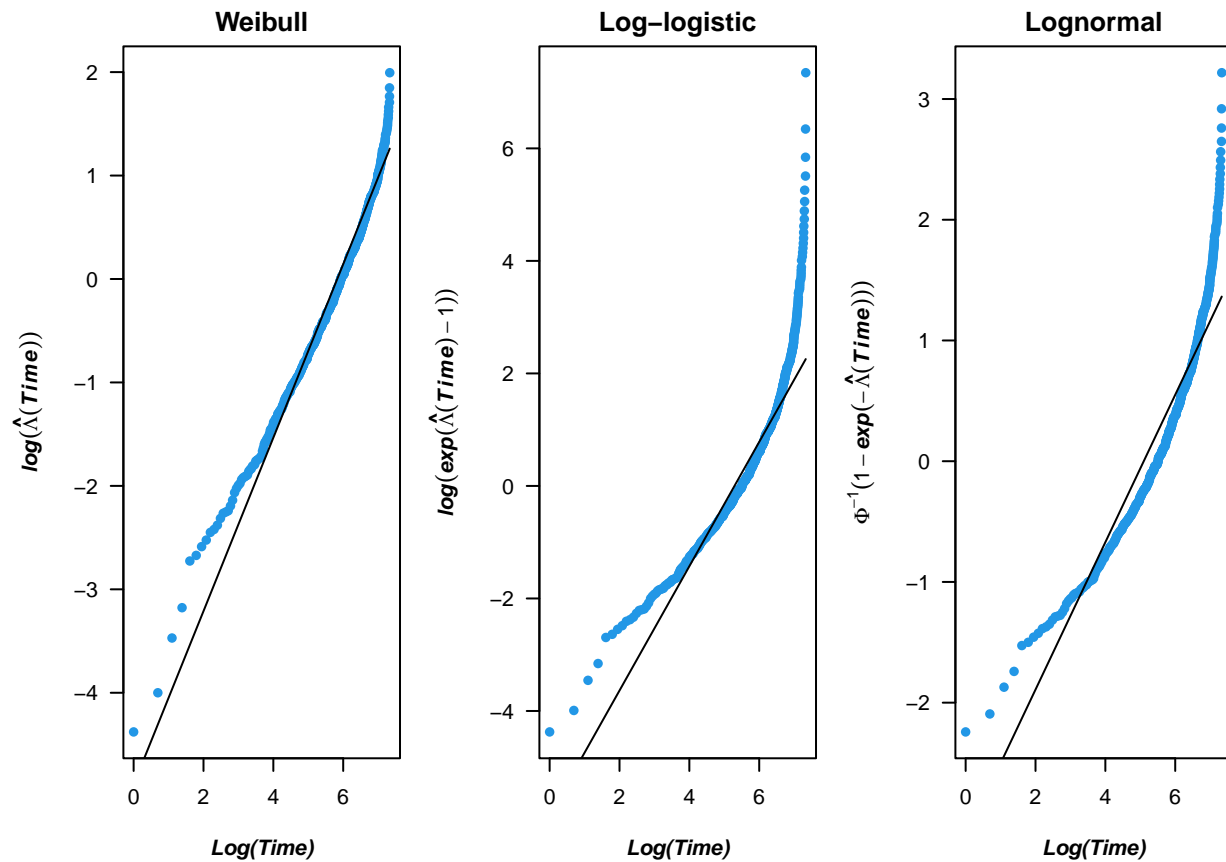
Residuals of the Lognormal Regression Model



I would say that the Weibull looks the best based on the three residual plots that can be seen above. **But perhaps they should be tested in some other ways as well??**

Goodness of fit of the parametric models is done in lab6! These things can be used to choose which of the three distributions should be used!!

Does not really fit into the estimations from above! What have I done wrong?! I think I am misunders
`cumhazPlot(std_data$TimeUntilReinf, std_data$cens,col = 4, distr = c("wei", "loglo", "lognormal"), font`



```
#>
#> Parameter estimates
#> =====
#> $weibull
#>      shape      scale
#> 0.8385125 340.1274020
#>
#> $loglogistic
#>      shape      scale
#> 1.104856 198.683716
#>
#> $lognormal
#> meanlog  sdlog
#> 5.098800 1.637882
```

Effect size measures:

- lognormal: acceleration factor.
- weibull: HR.
- log-logistic: OR.

Fit of a semi-parametric survival model

Conclusions