Practical Work

Lifetime Data Analysis

Rodrigo Arriaza, Alexander J Ohrt

09 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))

InitInfect is interesting to use, because several of the studies above are only done on one of the two diseases, not on both at the same time. Also “almost” statistically significant, even though this is not very precise at all.

```
std_data <- read.table("STD_onlydata.txt")
colnames(std_data) <- variable.names <-c("ObsNum", "Ethnicity", "MaritalStatus",
        "Age", "YearsSchool", "InitInfect", "NumPartners", "OralSex12m",
        "OralSex30d", "RectalSex12m", "RectalSex30d", "AbPain",
        "SignDischarge","SignDysuria","CondomUse","SignItch","SignLesion",
        "SignRash","SignLymph","InvVagAtExam","DischargeExam","AbnormNodeExam",
        "Reinfection", "TimeUntilReinf")
non_factor_indices <- c(1, 4, 5, 7,23, 24) # I THINK Reinfection should be a factor as well!
# This would mean that we should remove 23 from this list.
std_data[, variable.names[-non_factor_indices]] <- lapply(
  std_data[, variable.names[-non_factor_indices]],factor)
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.

Variable selection

Q: What is the offset(log(TimeUntilReInf))? I don’t understand this to be honest ;) Can you explain?

```
#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(6,23,24), color='white', background = 'blue')
```

Descriptive Analysis

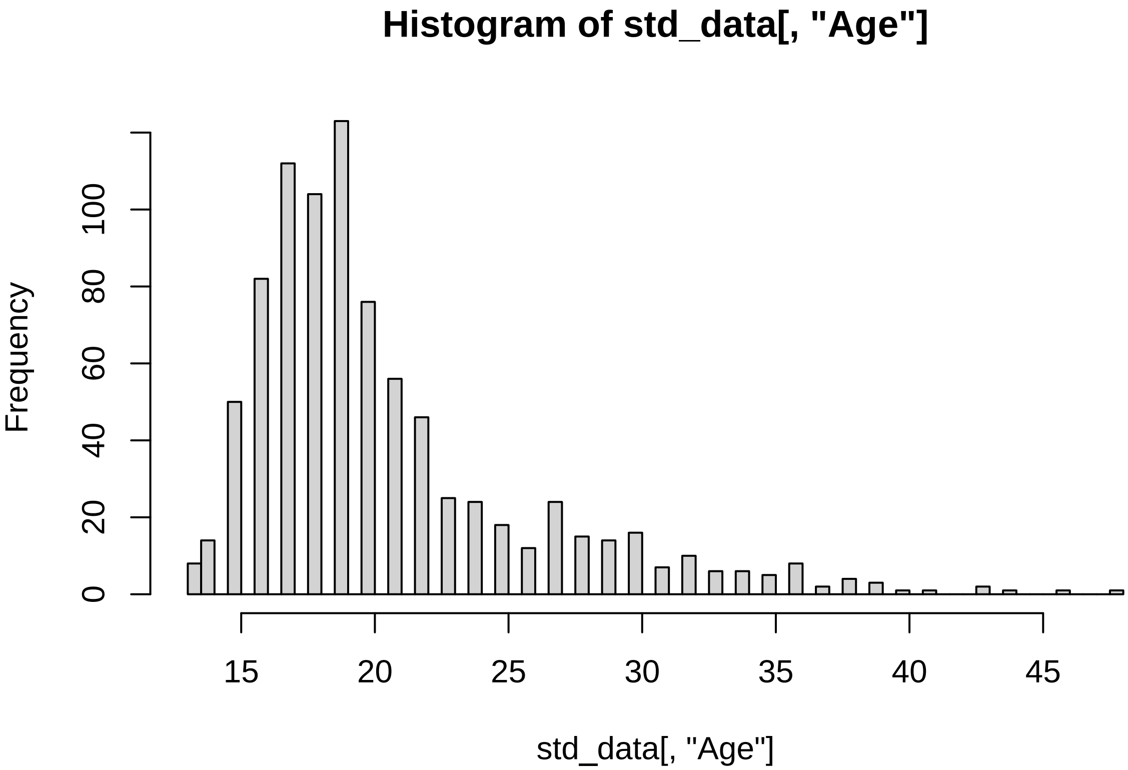
```
hist(std_data[, "Age"], breaks = 100)
```

Table 1: Marital Status Frequency in each Level.

Marital Status	Frequency
D	60
M	28
S	789

Table 2: Sign of Discharge Frequency in each Level.

Sign of Discharge	Frequency
0	472
1	405



```
knitr::kable(table(std_data[, "MaritalStatus"]), col.names = c("Marital Status", "Frequency"), caption =
```

```
knitr::kable(table(std_data[, "SignDischarge"]), col.names = c("Sign of Discharge", "Frequency"), captio
```

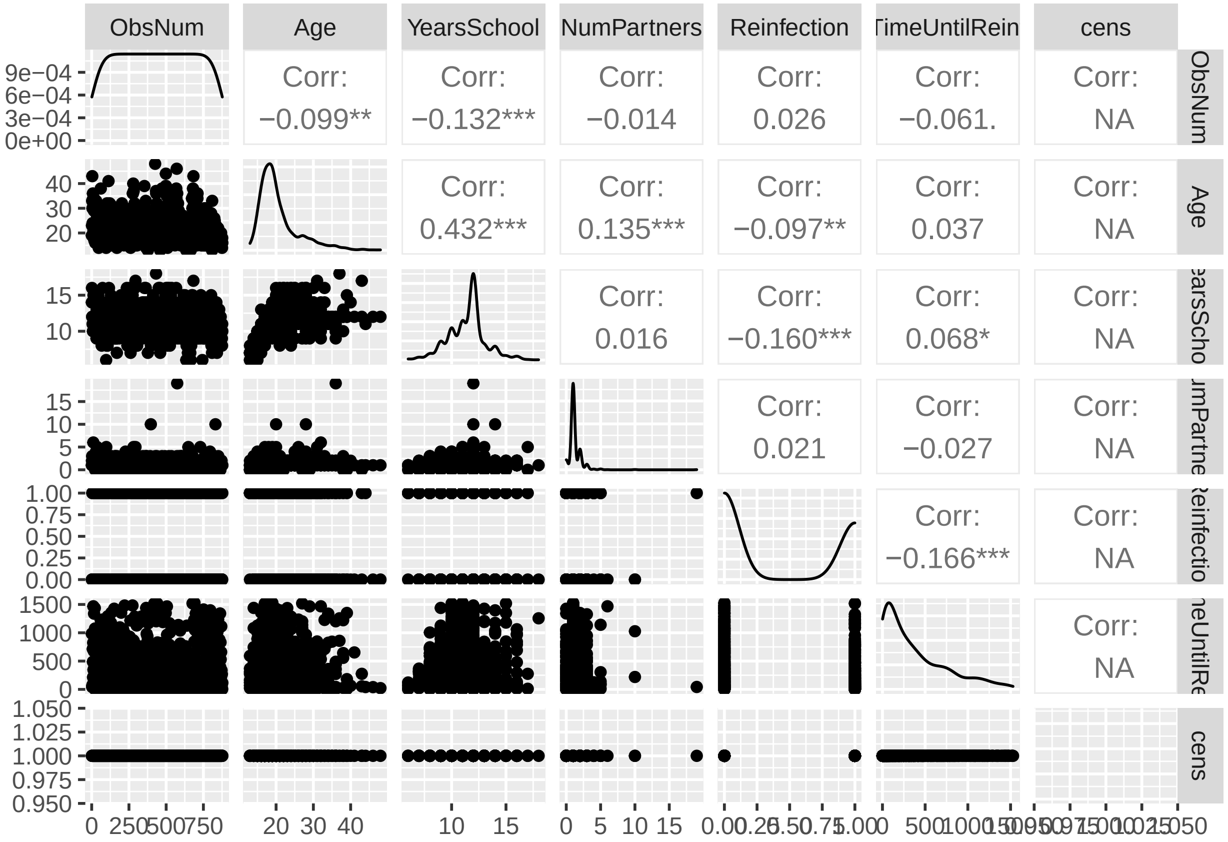
```
knitr::kable(table(std_data[, "DischargeExam"]), col.names = c("Discharge at Exam", "Frequency"), captio
```

```
# Some more tables?
```

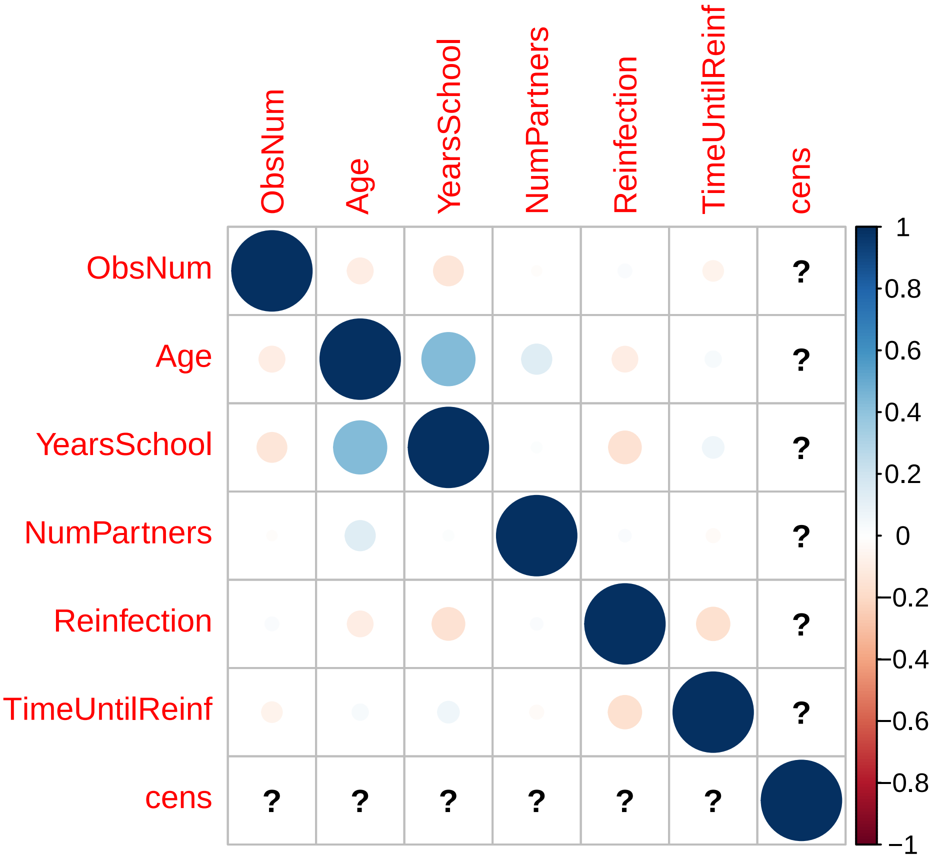
Table 3: Discharge at Exam Frequency in each Level.

Discharge at Exam	Frequency
0	48
1	829

```
ggpairs(std_data[,continuous.variables])
```



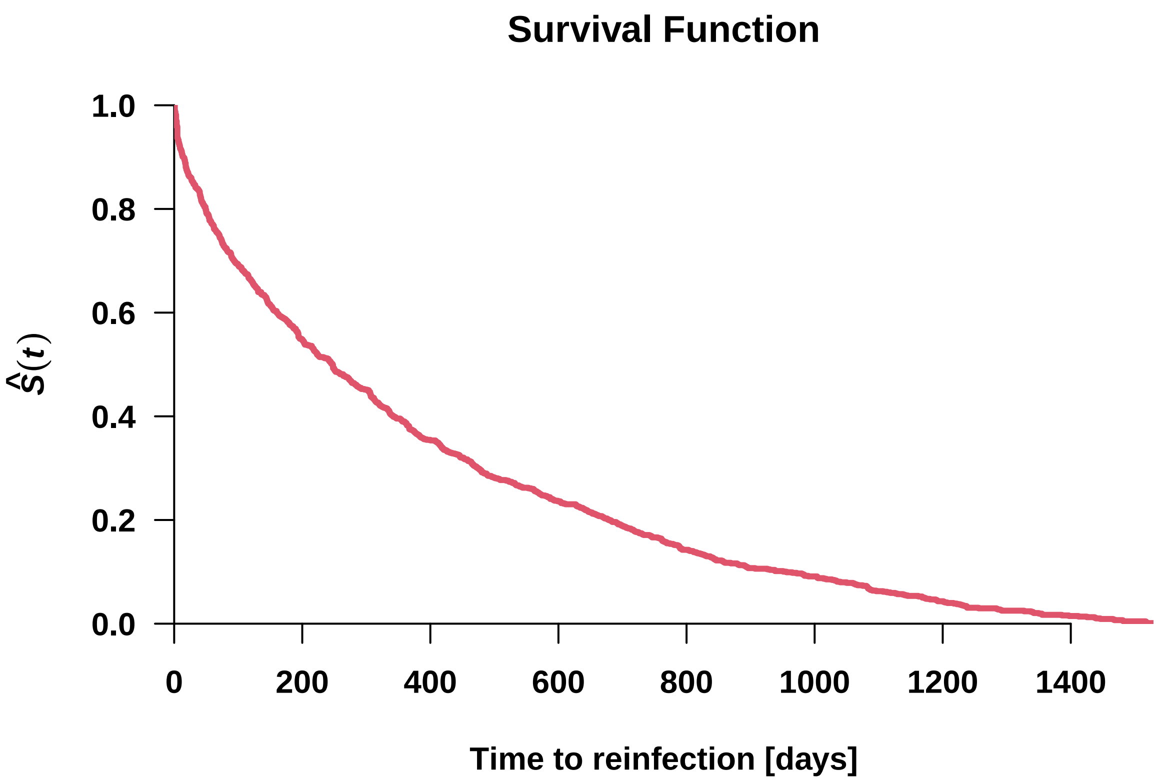
```
corrplot(cor(std_data[, continuous.variables]))
```



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.



The median survival time is 247 days.

Comparison of survival functions by means of nonparametric tests, such as the logrank test.

Below, logrank test for all curves plotted above are done. Other types of tests can also be used (Fleming-Harrington) and test can be done on different variables in the data set.

```
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

After trying to fit Weibull, log-logistic and lognormal log-linear models, we concluded that the Weibull model is best suited to our data.

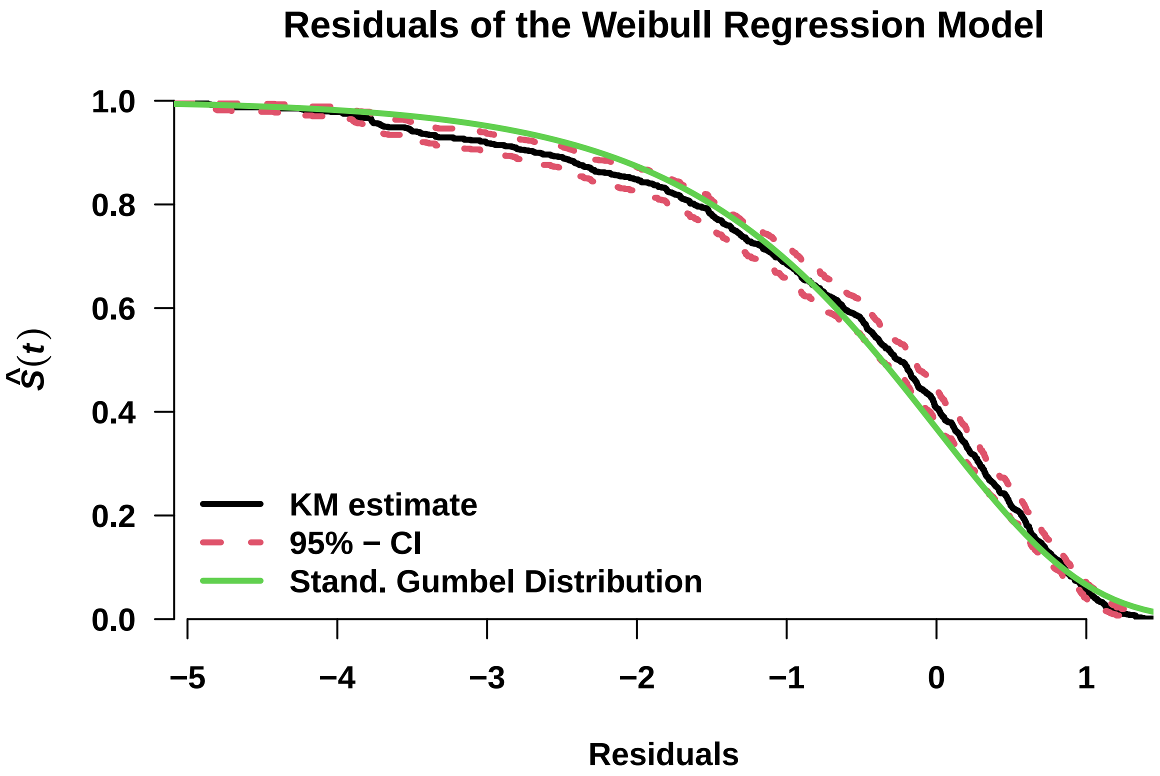
The correct varliables should be inserted here after we decide on them!!

```
# This should be done first with Weibull, to avoid having the log-fit here.
loglo.full <- survreg(s2 ~ Ethnicity + Age + NumPartners + CondomUse + YearsSchool + SignDischarge, dat
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
```

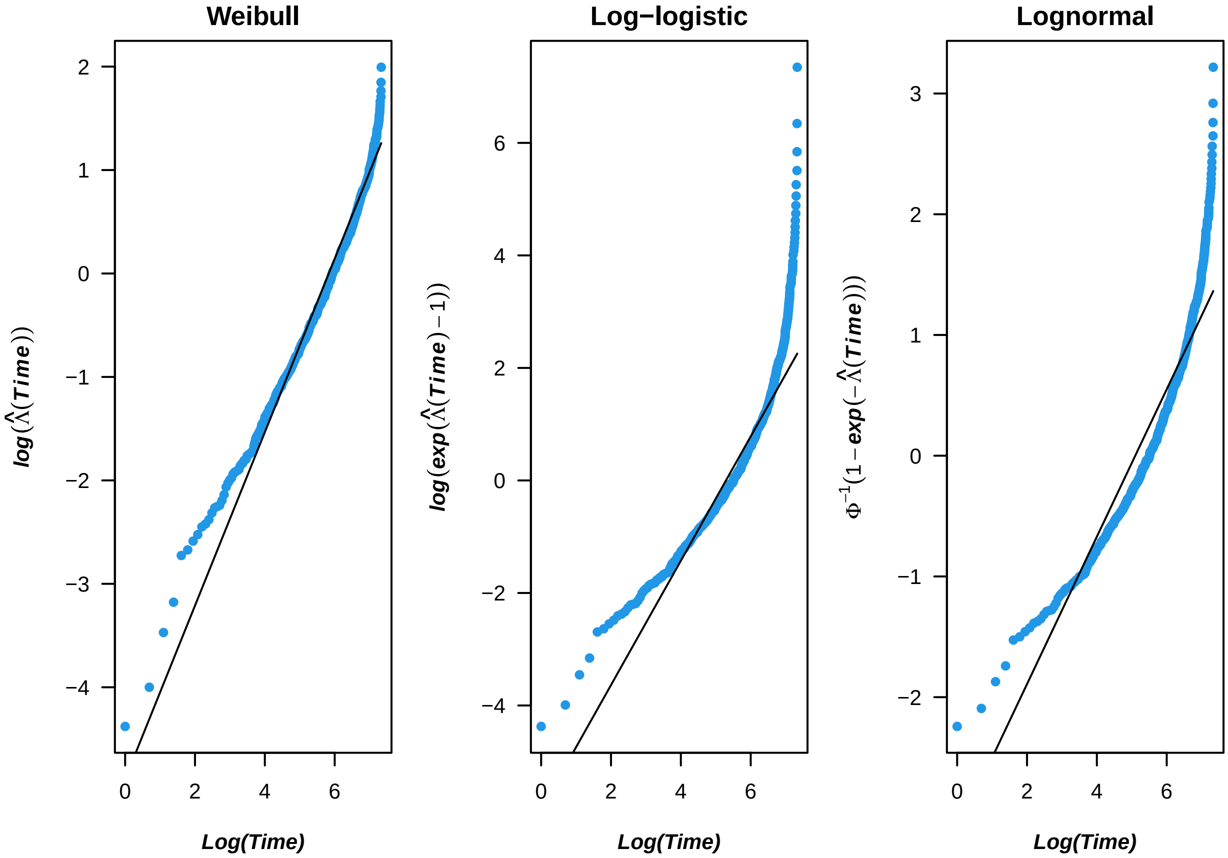
```
weibull.pred <- predict(weibull.full, type = "linear")
resids.weibull <- (log(std_data$TimeUntilReinf) - weibull.pred) / weibull.full$scale
```

```
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")
```



The residuals seem to fit relatively nicely to the Gumbel distribution, which indicates that the Weibull is reasonable to use **Check that the terminology I used here is correct!**

```
cumhazPlot(std_data$TimeUntilReinf, std_data$cens,col = 4, distr = c("wei", "loglo", "lognormal"), font
```



The probability plots above also show that the Weibull is the better parametric model for the data.

But how do we interpret this model fit?

The interpretation in terms of relative hazards (RHs) and the acceleration factor (AF) is . . .

Effect size measures:

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

Fit of a semi-parametric survival model

The proportional hazards model is fit.

Conclusions