Practical Work

Lifetime Data Analysis

Rodrigo Arriaza, Alexander J Ohrt

01 January, 2022

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, but have chosen to use the following, some for statistical reasons and some for medical reasons:

- 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 three 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.

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, the ages in the data set are low, which most likely means that the amount in each level of `MaritalStatus` is very skewed towards "single". This can be seen in the descriptive analysis below.

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 (`YearsSchool`) have any impact on chlamydia reinfection and as is shown below it showed to be statistically significant during the exploratory analysis.

Moreover, we chose to use the initial infection (`InitInfect`) as an explanatory variable, because several of the studies above are only done on one of the two diseases, not on both at the same time. Because of this we wanted to investigate if the initial infection type is a risk factor and, if this is the case, if the risk differs based on which infection was suffered initially.

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.

Statistical Variable Selection

As noted, in addition to medical criteria for selecting variables, we have used a negative binomial model to discover which variables are statistically significant to the event of reinfection. The negative binomial model can be fitted when the occurence of events is a count for each patient, as described in this article. Thereby including all the variables in the model and converting the time into an offset since we are comparing counts for different follow-up times (a person that has been in the study longer would have higher chances of getting reinfected).

REMOVE THIS AFTER: I made a list of things that can be used in the explanation:

- Despite the fact that the age of the woman is found to not be statistically significant in the method above, we have added it because of the mentioned studies (this is thus added based on medical criteria)

```
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)
p.model <- glm(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, family = poisson())
s <- summary(nb.model)
k <- knitr::kable(s$coefficients)
kableExtra::row_spec(k, c(6,23,24), color='white', background = 'blue')
```

	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

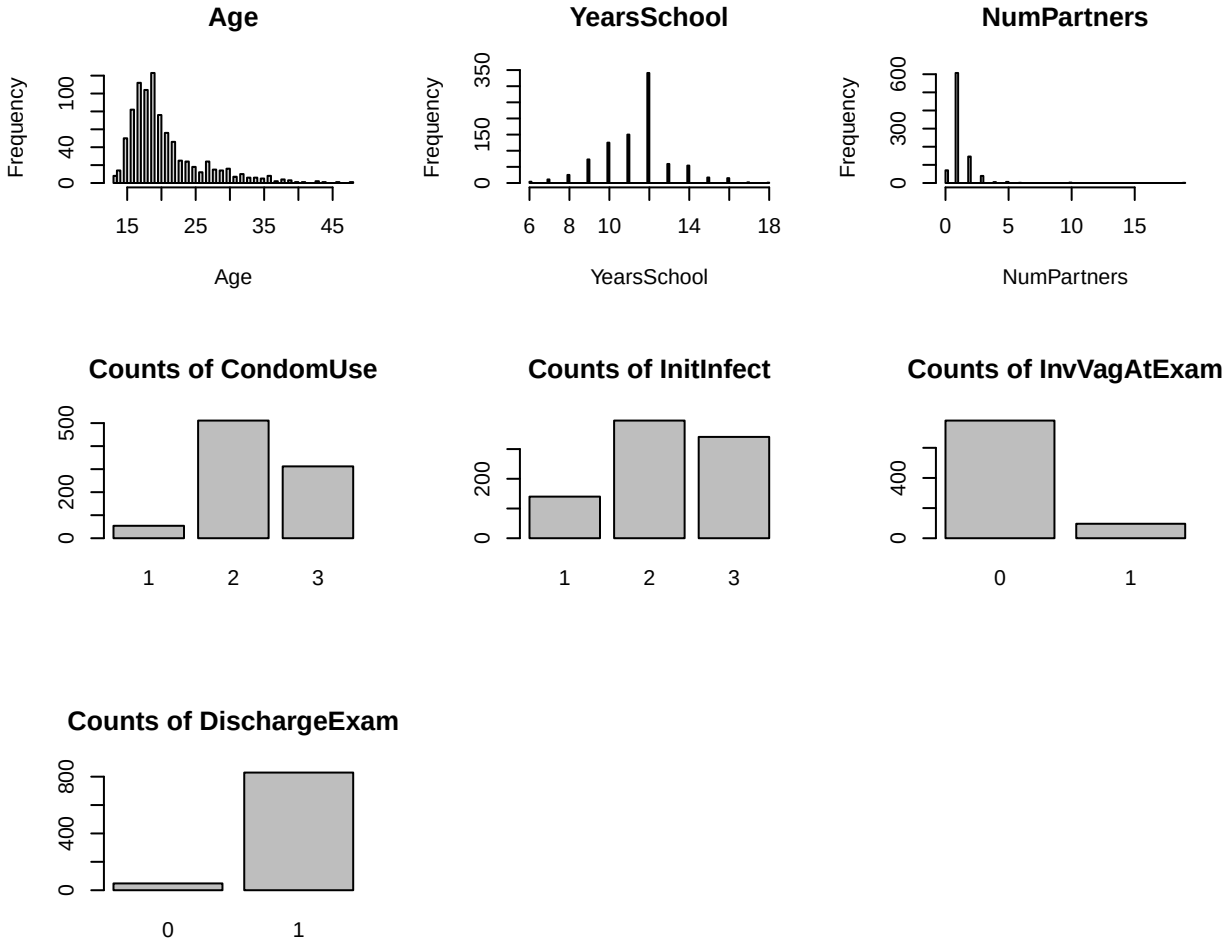
Finally, the vaginal involvement at exam (`InvVagAtExam`) and the discharge at exam (`DischargeExam`) are selected as variables in our analysis, since they are shown as statistically significant in **??**.

Descriptive Analysis

In total, the data set contains 24 variables, but, as noted, we have selected only 7 of them in our analysis. Recall that the data set has 877 women. The percentage of right-censored data in the data set is 60.4, which is a relatively large part of the data set. The women where followed for 1529 days, then the study was stopped.

Table 1: Corr. Between Continuous Variables

	Age	YearsSchool	NumPartners
Age	1.0000000	0.4316163	0.1348591
YearsSchool	0.4316163	1.0000000	0.0155090
NumPartners	0.1348591	0.0155090	1.0000000



The three continuous variables we have chosen to use in the analysis are `Age`, `YearsSchool` and `NumPartners`. The correlations between the variables are shown in table 1. Note that the correlation between `Age` and `YearsSchool` is 0.43, which means that they are somewhat correlated. This could be interesting to have in mind in the following.

Nonparametric Analysis

Survival Curve Estimation

The survival curve is estimated by means of Kaplan-Meier and plotted below. The curve below shows the general survival in the data set.

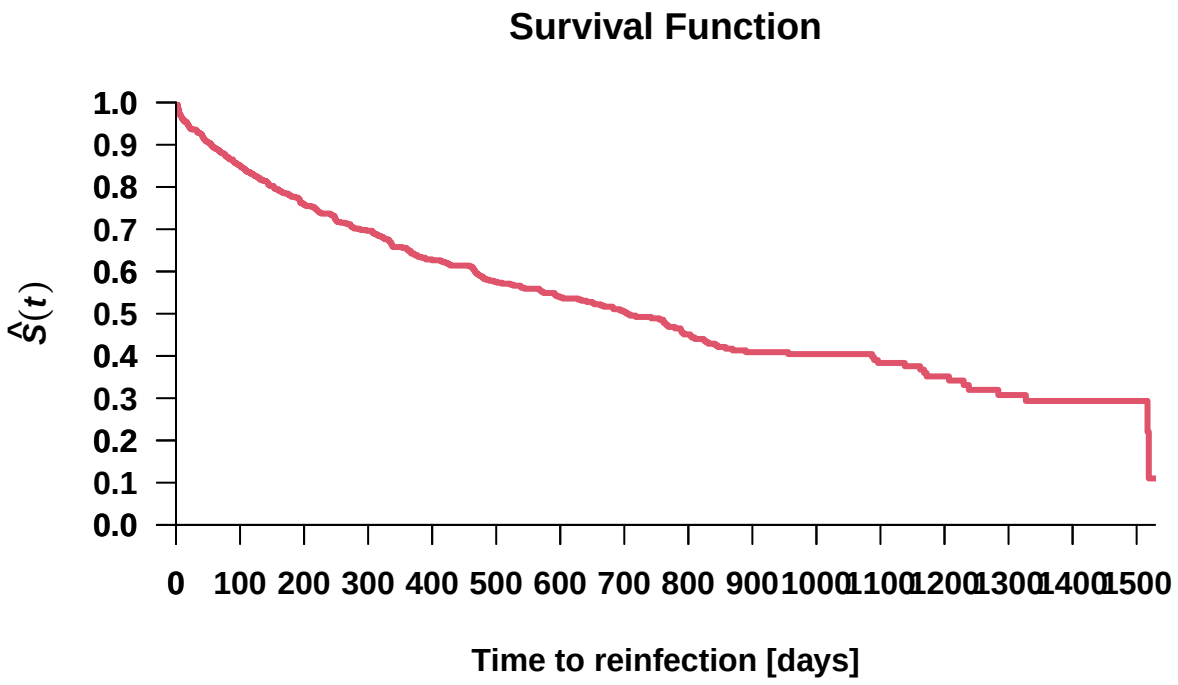


Figure 1: Survival function of time until reinfection

The median survival time is estimated to be 705 days.

Comparison of Survival Curves

Below, survival functions are compared by means of the nonparametric logrank test. Note that other types of tests also can be used (Fleming-Harrington family of tests), but we have only used the logrank test in this case. The general k-sample hypothesis that is tested is

$$H_0 : S_1(t) = \ldots = S_k(t), \forall t \leq \tau \text{ vs. } H_1 : \text{ some } S_i(t) \neq S_l(t), \text{ for some } t \leq \tau,$$

where τ is the chosen limit of the time of examination and k varies depending on the levels of the explanatory variable we are testing. **What about tests on continuous variables, does this make sense as well?** The p-values from each of the tests are given in table 2. For instance, choosing a significance level of $\alpha = 0.05$, we would conclude that reinfection depends on the level of `CondomUse`, `InitInfect` and `InvVagAtExam`, but that there is not enough evidence to conclude that reinfection depends on the level of `DischargeExam`.

Table 2: p-values from logrank tests

	CondomUse	InitInfect	InvVagAtExam	DischargeExam
p-values	0.0132506	0.0145266	0.0068009	0.0558044

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. **Should we have some interaction terms as well? Could be interesting! There are some significant interactions in both parametric and semi-parametric models, so this could be interesting I think! Then the interpretations need to be explained further also (the interaction terms are explained a bit differently).**

```
#>
#> Call:
#> survreg(formula = s2 ~ ., data = final.data, dist = "weibull")
#>                   Value Std. Error     z       p
#> (Intercept)      3.9516     0.6338  6.23 4.5e-10
#> Age              0.0101     0.0161  0.63 0.52912
#> NumPartners     -0.0139     0.0681 -0.20 0.83835
#> CondomUse2       0.0788     0.2995  0.26 0.79244
#> CondomUse3       0.4228     0.3106  1.36 0.17350
#> YearsSchool      0.1704     0.0487  3.50 0.00047
#> InitInfect2      0.5110     0.1907  2.68 0.00738
#> InitInfect3      0.3149     0.1915  1.64 0.10011
#> InvVagAtExam1   -0.5092     0.2212 -2.30 0.02133
#> DischargeExam1   0.4601     0.2894  1.59 0.11184
#> Log(scale)       0.2606     0.0445  5.85 4.9e-09
#>
#> Scale= 1.3
#>
#> Weibull distribution
#> Loglik(model)= -2674.9   Loglik(intercept only)= -2697.1
#>  Chisq= 44.43 on 9 degrees of freedom, p= 1.2e-06
#> Number of Newton-Raphson Iterations: 7
#> n= 877
```

As seen in 2, the standard Gumbel distribution seems to fit relatively nicely to the Kaplan-Meier estimate of the residuals, i.e. it seems like a reasonable choice for the error term W, which indicates that the Weibull is a reasonable model.

The probability plots in 3 also show that the Weibull is the better parametric model for the data, because the log-logistic and lognormal models clearly do not fit the line in the tails.

Interpretation

How do we interpret this model fit? First of all, the model we have fit follows the expression

$$Y = \ln(T) = \mu + \gamma^T \mathbf{Z} + \sigma W,$$

where $W \sim EV(0, 1)$,

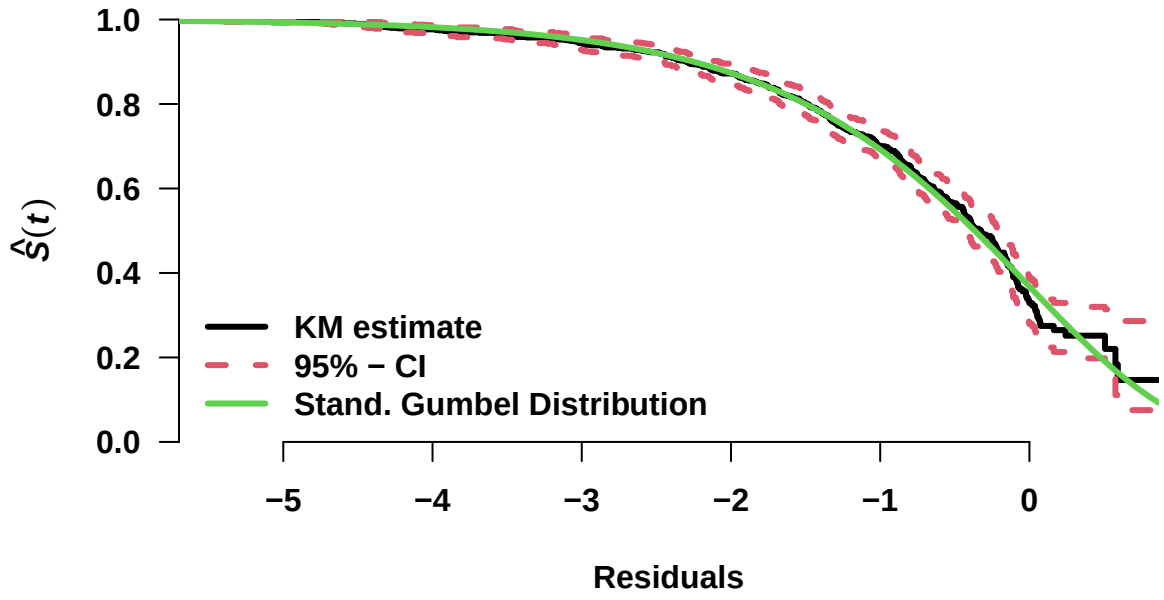


Figure 2: Residuals of the Weibull Regression Model

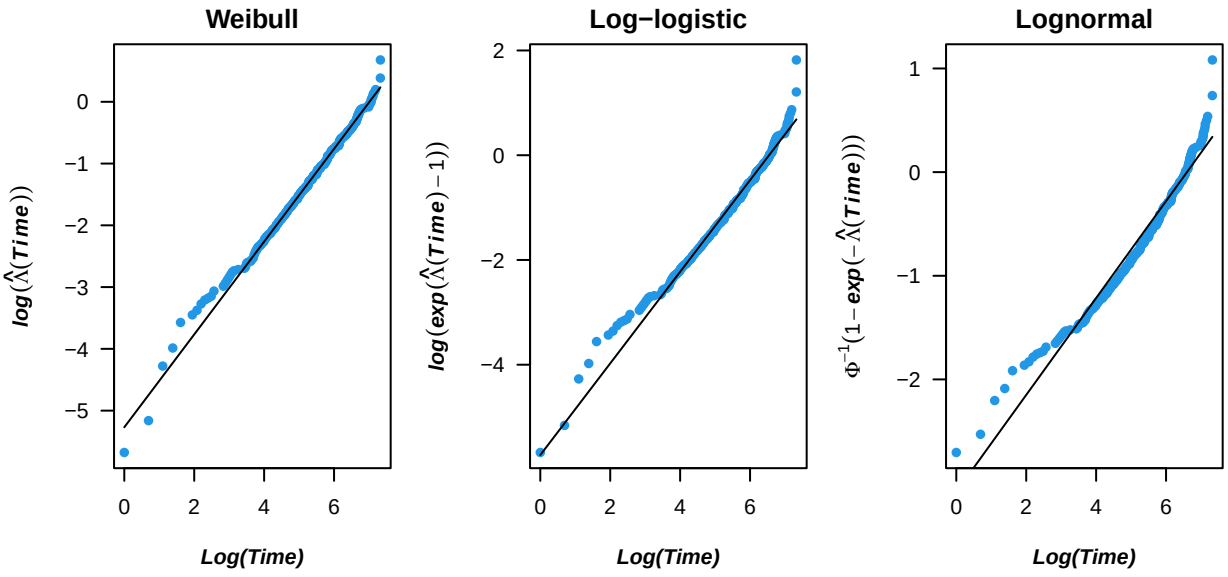


Figure 3: Cumulative hazard plots comparison

Table 3: Parameter Estimates, AF and HR for each Parameter Estimate

	Parameter.Estimate	AF	HR
(Intercept)	3.9516047	0.0192238	0.0475893
Age	0.0101018	0.9899491	0.9922457
NumPartners	-0.0138832	1.0139800	1.0107560
CondomUse2	0.0788112	0.9242144	0.9410747
CondomUse3	0.4227912	0.6552154	0.7219443
YearsSchool	0.1704370	0.8432962	0.8769191
InitInfect2	0.5109972	0.5998971	0.6745026
InitInfect3	0.3149121	0.7298530	0.7845268
InvVagAtExam1	-0.5091714	1.6639119	1.4804896
DischargeExam1	0.4601294	0.6312020	0.7014676

$$\gamma^T = (\gamma_{Age}, \gamma_{NumPartn.}, \gamma_{Cond.}, \gamma_{YSchool}, \gamma_{InitInf.}, \gamma_{InvVagAtExam.}, \gamma_{DischargeAtExam}),$$

are the estimated parameters and

$$\mathbf{Z}^T = (Age, NumPartn., Cond., YSchool, InitInf., InvVagAtExam., DischargeAtExam),$$

is the vector of values. Thus, each of the quantities $\exp(\gamma_i)$ can be interpreted as the unitary change in time until reinfection (when covariate i is continuous), or the change in time until reinfection when changing level (when the covariate i is categorical with different levels), when all the other explanatory variables are kept fixed. This means that a positive parameter estimate $\hat{\gamma}_i$ gives $\exp(\gamma_i) > 0$, which means that the covariate is estimated to being protective by the model, since it increases $\ln(T)$. The opposite holds for $\hat{\gamma}_i < 0$. These interpretations will be done with the acceleration factor and relative hazard next.

In the Weibull model, the acceleration factor (AF) is calculated using the equation

$$AF = \exp(-\hat{\gamma}_i)$$

and the hazard ratio (HR) is calculated using the equation

$$HR = \exp(-\hat{\gamma}_i/\hat{\sigma}).$$

In this case, the model fit gives the scale $\hat{\sigma} \approx 1.298$. These values are calculated for each of the covariates below.

Consider an example using the covariate `CondomUse` when explaining the interpretation of the covariates in terms of the AF. From the table above it is apparent that the AF of `CondomUse3` versus `CondomUse1` is $\approx$ 0.655. This means that the reinfection time for a person that never uses a condom is ≈ 0.655 times the reinfection time for a person that always uses a condom **Not sure that this makes sense!? I think it makes sense with the coefficient value given from the model above, but does not make sense in real life, as this suggests that not using a condom is protective!** The interpretation in terms of the AF is similar when considering the other covariates, except for when considering the `Age` and `NumPartners`, which is not categorical **Perhaps it indeed makes sense to considering the Age in this way also, even though it is weird to treat the Age this way?**

Here I have assumed that the relative hazards is the same as the hazard ratio?? IS THIS CORRECT?! I think so! Looks like the values make sense with the results from the Cox-model below!

Similarly, an example can be used to explain the interpretation of the covariates in terms of the relative hazard. From the table it is apparent that the hazard of `CondomUse3` relative to `CondomUse1` is ≈ 0.722.

This means that the instantaneous risk of reinfection for a person that never uses a condom is ≈ 0.722 times the instantaneous risk of a person that always uses a condom. Similar interpretations can be done with the other covariates.

Fit of a semi-parametric survival model

The proportional hazards model is fit below.

```
#>                       coef exp(coef)   se(coef)          z     Pr(>|z|)
#> Age            -0.00783097 0.9921996 0.01237892 -0.6326050 0.5269915929
#> NumPartners     0.01119536 1.0112583 0.05248331  0.2133127 0.8310830727
#> CondomUse2     -0.05519517 0.9463004 0.23121727 -0.2387156 0.8113261078
#> CondomUse3     -0.33570604 0.7148332 0.23948574 -1.4017788 0.1609813112
#> YearsSchool    -0.13356369 0.8749717 0.03740282 -3.5709522 0.0003556858
#> InitInfect2    -0.38680211 0.6792255 0.14648410 -2.6405740 0.0082765721
#> InitInfect3    -0.23500491 0.7905670 0.14768228 -1.5912871 0.1115449801
#> InvVagAtExam1   0.40191622 1.4946861 0.17048096  2.3575432 0.0183963138
#> DischargeExam1 -0.36743845 0.6925059 0.22324601 -1.6458903 0.0997863366
```

Interpretation

How do we interpret this model fit? First of all, the model we have fit follows the expression

$$\lambda(t|\mathbf{Z}) = \exp(\beta^T \mathbf{Z})\lambda_0(t),$$

where β are the parameters in the model and $\mathbf{Z}$ is the profile of the woman. Additionally, $\lambda_0(t)$ is the hazard at time t for a woman with profile $\mathbf{Z} = 0$, i.e. a woman that always uses a condom, that was initially infected with (only) gonorrhea, that did not experience vaginal involvement at exam and did not experience discharge at exam. The model assumes that the hazard ratio is proportionally equal to $\exp(\beta^T \mathbf{Z})$ at all times. Said in other words, it relates the instantaneous risk for a woman with profile $\mathbf{Z}$ at time t with the instantaneous risk for a woman with the baseline profile at the same time t.

The model parameters β can be interpreted in terms of relative hazards. As is seen from the formula above, the hazard ratio between a woman with profile $\mathbf{Z}$ and a woman with profile $\mathbf{Z} = \mathbf{0}$ is $\exp(\beta^T$, where the values are given in the second column of the table above. The interpretation in terms of relative hazards in this case is the same as the interpretation of the Weibull survival model fit earlier, since the Weibull regression model allows a representation of the proportional hazards model. This is done by setting $\beta = -\gamma/\sigma$. Note that, because of this, the values for `exp(coef)` in the table above and the `HR`-values for the Weibull model calculated earlier are very similar, as they should. They are not exactly the same for numerical reasons when fitting the models.

Analysis of Residuals

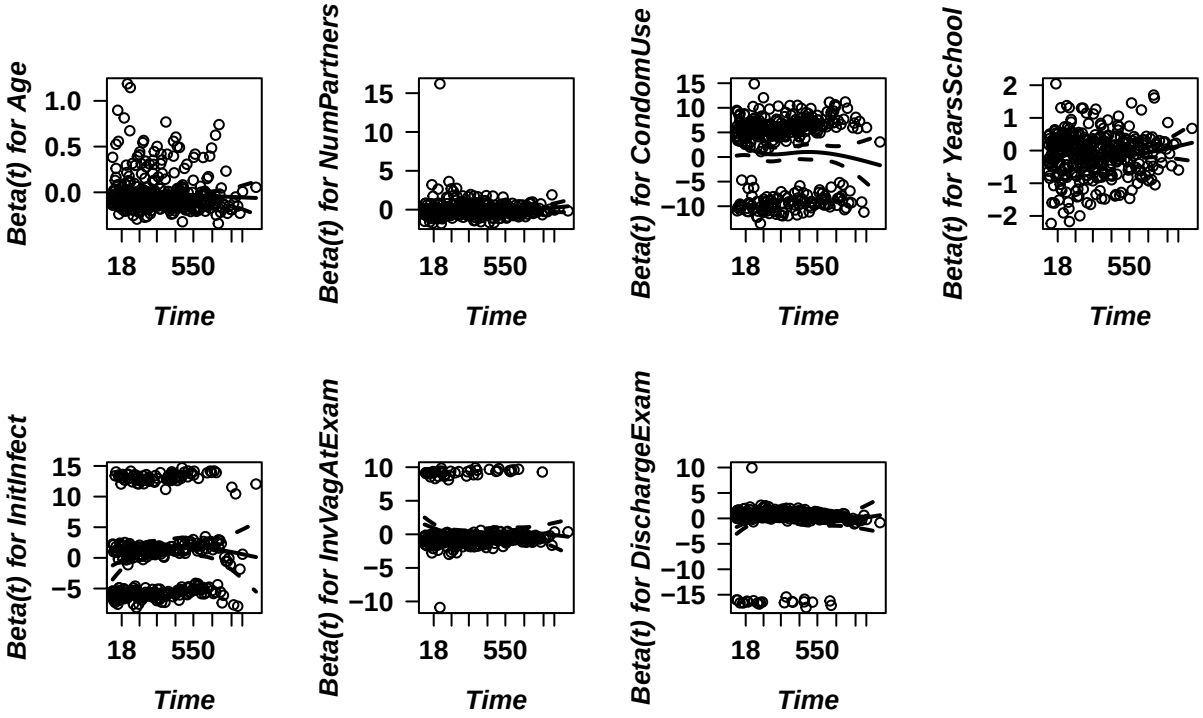
In this section we will check the Cox-model's goodness-of-fit. **Not very satisfied with a lot of the GOF I did below, something else we can do also? Check material.**

Proportional Hazards Assumption

First of all, the Shoenfeld residuals can be used to check the proportional hazards assumption. The residuals are plotted below. The lines in the plots look relatively straight, which indicates that the proportionality assumption might hold.

Table 4: Hypothesis Test for Proportionality Assumption (Shoenfeld)

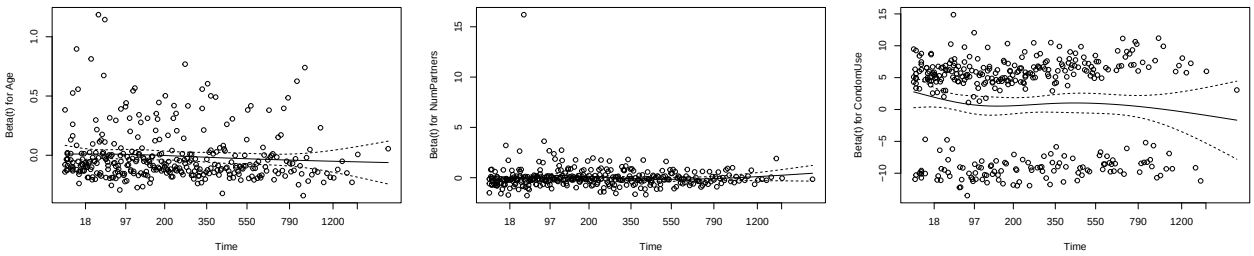
	chisq	df	p
Age	1.5024354	1	0.2202970
NumPartners	0.9963277	1	0.3182007
CondomUse	3.0132522	2	0.2216566
YearsSchool	0.0247726	1	0.8749349
InitInfect	3.5534579	2	0.1691907
InvVagAtExam	1.3535072	1	0.2446659
DischargeExam	0.4134607	1	0.5202182
GLOBAL	11.8922549	9	0.2194519

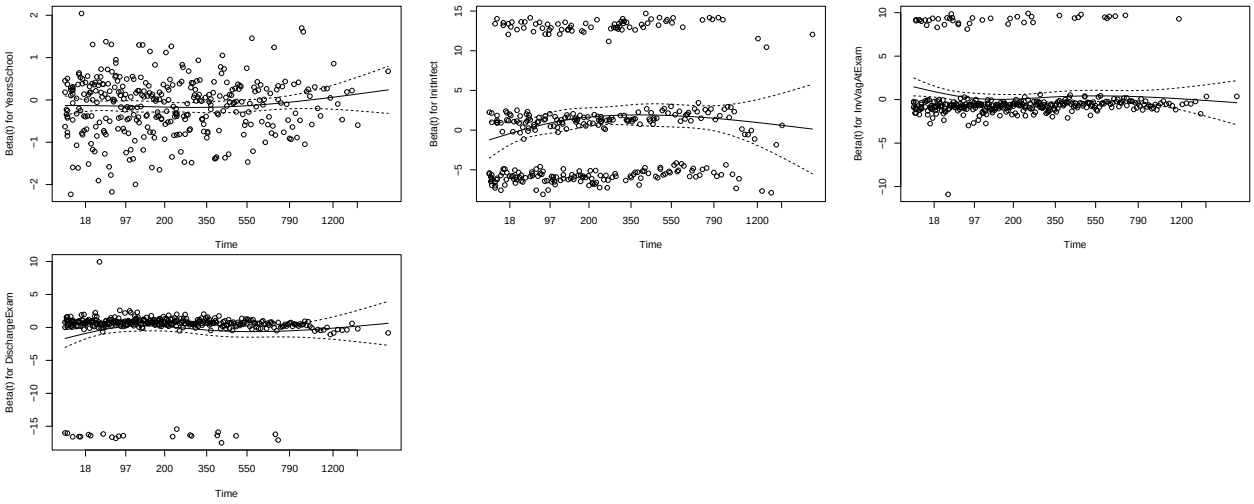


The hypothesis test for the proportionality assumption on each of the covariates is applied above. As the table shows, all p-values are large compared to any reasonable significance level, which means that we do not reject the null hypothesis and we can conclude that the property of proportionality of the covariates is reasonably fulfilled.

The plots or table above might need to be removed later, but we will see in the end.

DO THE ANALYTIC CHECK AND/OR GRAPHICAL CHECK THAT THEY TALK ABOUT IN THE SLIDES ALSO?





In the plots above it is noticeable that they all show a horizontal line which confirms the numerical analysis from 4.

Influential Observations in the Global Fit

Secondly, we can check if there are any influential observations using residuals based on the score residuals. A transformation of the score residuals for each of the four coefficients is plotted below. More precisely, each residual that is plotted is the approximate change in the coefficient vector if the observation in question is dropped, scaled by the standard error of the coefficients.

From the plots above, it becomes apparent that there clearly are some influential observations for each of the covariates.

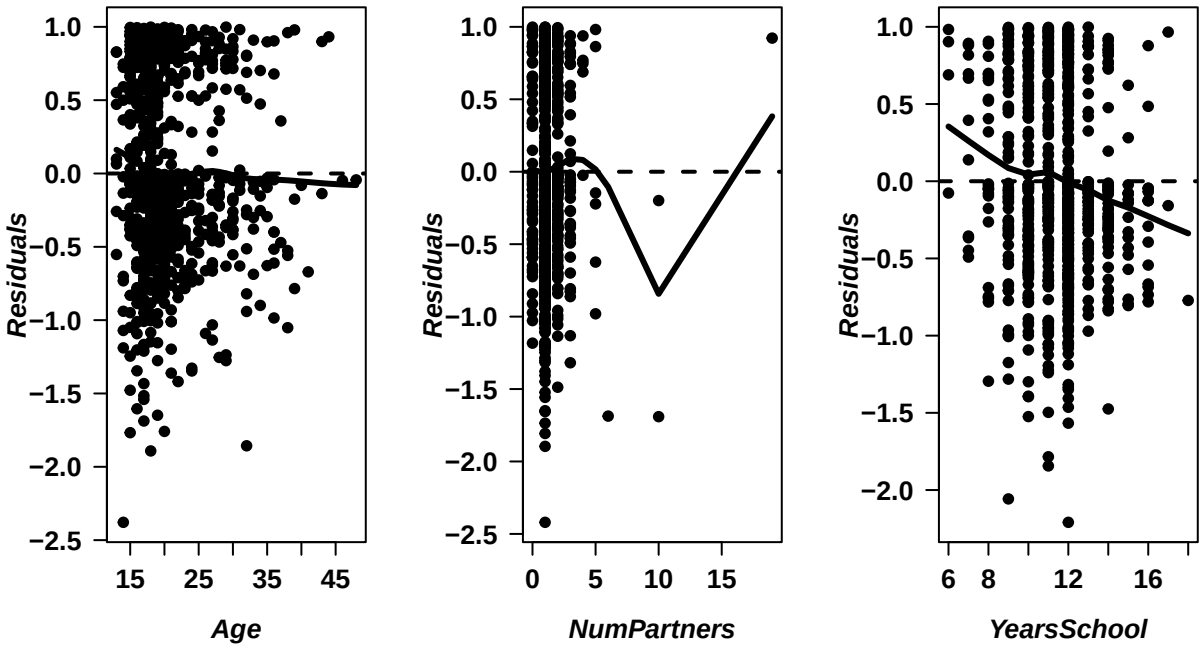
```
#>     ObsNum Ethnicity MaritalStatus Age YearsSchool InitInfect NumPartners
#> 4        4         B             S  43          12          3           1
#> 11      11         B             D  32          12          3           6
#> 154    154         B             D  28          11          2           1
#> 221    221         B             S  18          11          1           1
#> 366    366         W             S  14           9          1           1
#> 498    498         B             M  44          11          3           1
#> 525    525         B             S  15           8          3           1
#> 574    574         B             D  36          12          3          19
#> 831    831         W             S  20          12          2          10
#>     OralSex12m OralSex30d RectalSex12m RectalSex30d AbPain SignDischarge
#> 4            0          0            0            0      0             0
#> 11           1          1            0            0      1             1
#> 154          1          1            0            0      0             0
#> 221          1          1            1            1      0             0
#> 366          0          0            0            0      0             0
#> 498          0          0            0            0      0             0
#> 525          1          1            0            0      0             1
#> 574          1          1            0            0      1             1
#> 831          1          1            1            1      0             1
#>     SignDysuria CondomUse SignItch SignLesion SignRash SignLymph InvVagAtExam
#> 4             0         1        0          0        0         0            0
#> 11            0         2        0          0        0         0            1
#> 154           0         2        0          0        0         0            0
#> 221           0         1        0          0        0         0            0
```

```
#> 366            1         2        0           0        0         0            0
#> 498            0         3        0           0        0         0            0
#> 525            0         3        0           0        0         0            1
#> 574            0         1        1           0        0         0            0
#> 831            0         2        1           0        0         0            1
#>     DischargeExam AbnormNodeExam Reinfection TimeUntilReinf
#> 4               1               0           1            54
#> 11              1               0           0          1468
#> 154             0               0           0           880
#> 221             1               0           0          1481
#> 366             1               0           0          1439
#> 498             1               0           1            42
#> 525             1               0           0          1005
#> 574             1               0           1            43
#> 831             0               0           0          1027
```

By plotting the residuals it was possible to identify as influential observations the individuals with observation numbers: 4, 11, 154, 221, 366, 525, 574, 831. Individuals 4, 498 and 574 showed an unexpectedly short time until reinfection, especially for individuals of their age (older than 40). Whereas the other individuals exhibit a much greater time until reinfection, with almost all of them having more than 1000 days until the event.

Linear Covariates Assumption

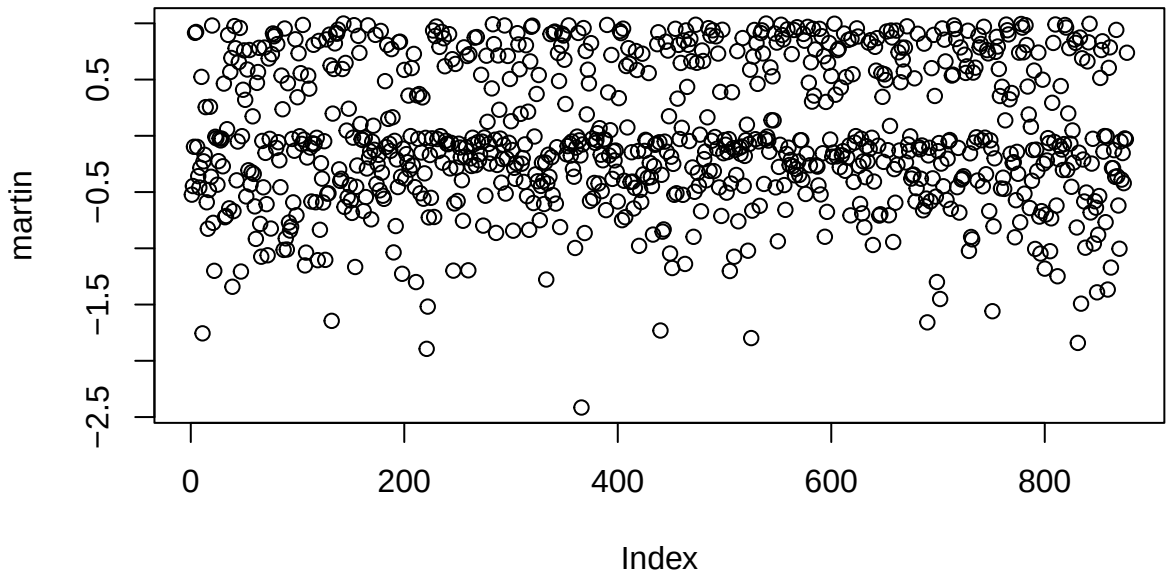
Thirdly, the the linear assumption of the continuous variables `Age`, `NumPartners` and `YearsSchool` is checked.



As can be seen from the plots above, the linear assumption of age seems to hold just fine, while the linear assumption for the two other continuous covariates is more uncertain if holds based on the smoother, which is not flat at all.

Perhaps these plots need to be removed as well? This part of the GOF is perhaps not as important either way. :(Hard to do the GOF-checks without plots perhaps, but I don't think these plots are the most important in the report.

Martingale Residuals (Can be used for global fit? Not sure how)



Conclusions (ADD MORE LATER PROBABLY. Also clean up text.)

Assuming that the parametric and semi-parametric models can be trusted, the data analysis has uncovered some possible protective factors and risk-factors for reinfection with gonorrhea or chlamydia in women who had suffered one of both infections previously.

Protective Factors

The covariates `YearsSchool` and `InitInfect2` (Initial infection of Chlamydia only) are statistically significant to a level $\alpha = 0.05$ in both the parametric and semi-parametric model. Moreover, the logrank tests of the survival curves yield the conclusion that the levels of `InitInfect` yield different survival curve, which supports the conclusion based on the models. Both these covariates are estimated to being protective factors against reinfection. When it comes to the initial infection of Chlamydia, it is estimated to reduce the instantaneous risk compared to a person with Gonorrhea as initial infection by approximately 33% according to the parametric Weibull model and the Cox-model. Moreover, each unitary increase in `YearsSchool` is estimated to reduce the instantaneous risk by approximately 12% according to the parametric Weibull model and the Cox-model. Recall that in all these estimations it is assumed that the rest of the profile of each individual is the same, except for the described change.

Risk-factors

The covariate `InvVagAtExam1` (yes) is statistically significant to a level $\alpha = 0.05$ in both the parametric and semi-parametric model. Moreover, the logrank test of the survival curves yield the conclusion that the levels of the covariate yields different survival curves, which supports the conclusion based on the models. This is estimated to being a risk factor by both models. It is estimated to increasing the instantaneous risk compared a person not experiencing involvement of vagina at exam (with identical profile except for this change) by approximately 48%, when selecting the slightly more optimistic estimate of the Weibull model instead of the more pessimistic estimate from the Cox model 49%.

Perhaps do some predictions of new data (as shown in lab 7 I think), to further quantify the conclusions? We don't have to, but could.