Survival Analysis of STD Reinfection Rates

Claire Hua (9952425), Ana Caklovic (3718673), Moises Sanchez (9866922)

12/01/2018

Abstract

The STD dataset contains 18 categorical variables and 4 numerical variables that describe the patient characteristics and symptoms that affect the rate of STD reinfection. Analysis of the data includes determining the significance of each predictor, modeling an appropriate Cox PH model, diagnostic checking, estimating hazard ratios and survival probabilities of different groups (such as education level), and fitting a parametric model. The direction of our analysis focuses on research questions about the effect of education and the relevance of symptoms in STD reinfection.

Introduction

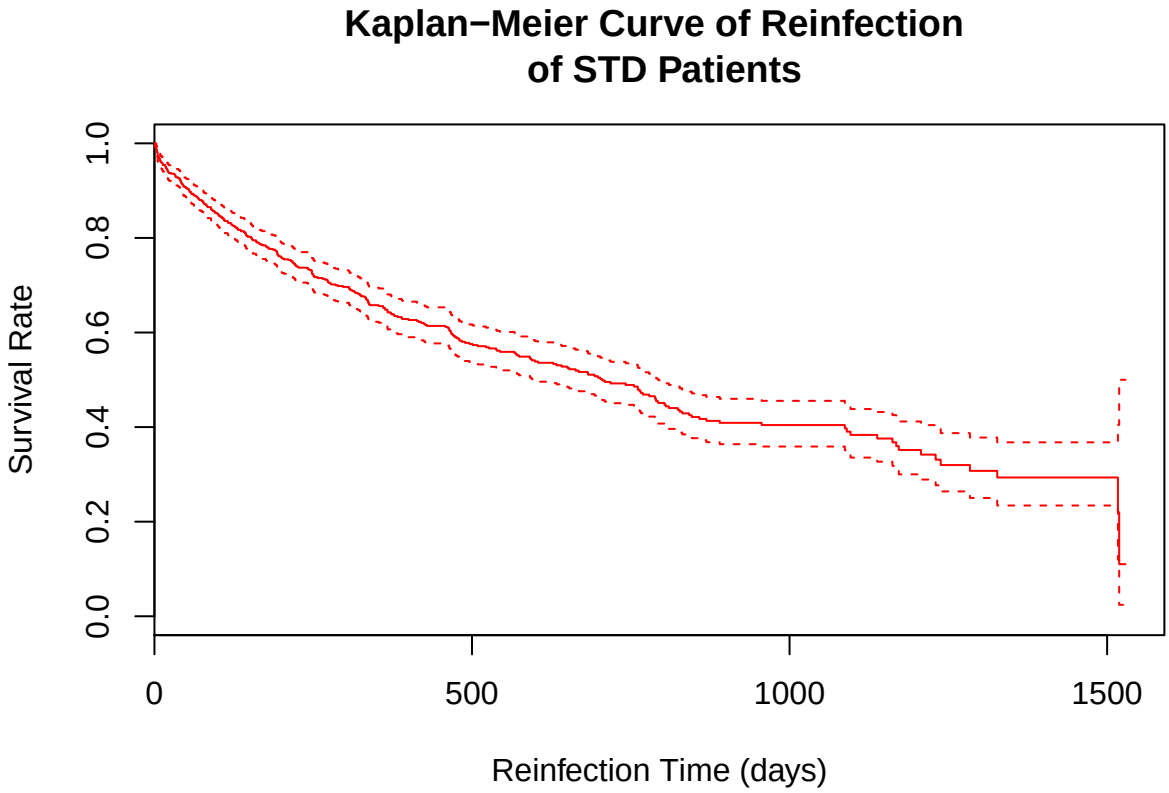
The following report will be analysing a dataset that contains information on the STD reinfection that comes from the textbook Survival Analysis: Techniques for Censored and Truncated Data written by Klein, J. P., & Moeschberger, M. L. (1997). The survey-structured dataset contains 877 observations with time measured in days. The variables examined in the dataset include numerical variables: age, initial infection, years of education, number of partners, and time until reinfection. The binary categorical variables are the symptoms of reinfection (itch, lesion, rash, sign of lymph, and discharge, etc.), time between having oral sex, time between rectal sex, discharge at exam, abnormal node at exam, sign of discharge, sign of dysuria, and reinfection. Our non-binary categorical variables are observation number, marital status, and initial infection. For our analysis we will be trying to fit a model to see the probability of a patient going for a full year without experiencing reinfection. More specifically, we will be doing these tests for those with a maximum of 12 years of schooling and those with less than 12 years of schooling. We will then perform further analysis to determine if a patient's years of schooling is related to the likelihood of STD reinfection, and also to determine which symptom(s) affect STD reinfection the most.

Variables		
Variable	Type	Description
obs	categorical	Observation number
race	categorical	Race (W=white, B=black)
martial	categorical	Marital status (D=divorced / separated, M=married, S=single)
age	numeric	Age
yschool	numeric	Years of schooling
iinfct	categorical	Initial infection (1= gonorrhea, 2=chlamydia, 3=both)
npartner	numeric	Number of partners
os12m	categorical	Oral sex within 12 months (1=yes, 0=no)
os30d	categorical	Oral sex within 30 days (1=yes, 0=no)
rs12m	categorical	Rectal sex within 12 months (1=yes, 0=no)
rs30d	categorical	Rectal sex within 30 days (1=yes, 0=no)
abdpain	categorical	Presence of abdominal pain (1=yes, 0=no)
discharge	categorical	Sign of discharge (1=yes, 0=no)
dysuria	categorical	Sign of dysuria (1=yes, 0=no)
condom	categorical	Condom use (1=always, 2=sometime, 3=never)
itch	categorical	Sign of itch (1=yes, 0=no)
lesion	categorical	Sign of lesion (1=yes, 0=no)
rash	categorical	Sign of rash (1=yes, 0=no)
lymph	categorical	Sign of lymph (1=yes, 0=no)
vagina	categorical	Involvement vagina at exam (1=yes, 0=no)
dchexam	categorical	Discharge at exam (1=yes, 0=no)
abnode	categorical	Abnormal node at exam (1=yes, 0=no)
rinfct	categorical	Reinfection (1=yes, 0=no)
time	numeric.	Time to reinfection

Initial Analysis

We created an initial Kaplan-Meier estimate of survival probabilities based on our dataset. The survival time is measured in days and the status variable is 'rinfct', which indicates whether the observation has been censored or not. The Kaplan Meier curve for the dataset here seems to not be exhibiting any odd behavior so methods will be applied to remedy it. Next we removed the race covariate from our original dataset since we are only trying to conduct a nonracial study on the STD reinfections. The model selection process can now begin after doing the initial preparation.

```
data("std")
std.new <- subset(std, select = -race)
std.surv <- Surv(std.new$time, std.new$rinfct)
std.fit <- survfit(std.surv ~ 1)
plot(std.fit, xlab= "Reinfection Time (days)", ylab= "Survival Rate",
     main = "Kaplan-Meier Curve of Reinfection \n of STD Patients", col=2)
```



Model Creation

We start our model selection by creating a test model with all the covariates in the dataset. Using the summary function on this model, we initially checked all the p-values of the covariates to give us a basic sense of what direction to take our analysis in. We will be performing the likelihood ratio test throughout our analysis by the following method.

To compute our LRT test statistic, we used this equation :

$$2 * [\text{loglikelihood(full model)} - \text{loglikelihood(reduced model)}]$$

The p-value is computed by using our pchisq() function, where degrees of freedom is determined by taking the difference of the number of estimated parameters for each model.

First, we wanted to focus on 'age' and 'yschool' before testing the significance of the symptoms of STD reinfection. We created two initial coxph models (~age + yschool vs. ~yschool) to test for the significance of the covariate 'age' by performing the likelihood ratio test, which tests for significant differences between two models. Our full model is (~age + yschool) and our reduced model is (~age).

```
#testing for significance of age
std.cox <- coxph(std.surv ~ age + yschool, data = std.new) #cox model w/ age and yschool
stdcox1 <- coxph(std.surv ~ yschool, data = std.new) #cox model w/ yschool
lrt = 2*(std.cox$loglik[2]-stdcox1$loglik[2]) #likelihood ratio test statistic
pchisq(lrt, df=1, lower.tail = FALSE)
```

```
## [1] 0.404246
```

```
#new model doesnt contain age
```

After computing the LRT test statistic for our two models with 1 degree of freedom, we got 0.6957, and our p-value was $0.404246 > 0.05$, which means the difference between the model(~age + yschool) and the model(~yschool) was insignificant, therefore we can remove the covariate 'age' from our final model.

```
std.cox.test <- coxph(Surv(std.new$time, std.new$rinfct) ~ ., data = std.new)
stdcox1 <- coxph(Surv(std.new$time, std.new$rinfct) ~ yschool, data = std.new)
```

```
mod.reduc = stdcox1 #reduced model with only yschool
mod.fin = std.cox.test #full model with all covariates
step(mod.reduc, scope = list(mod.reduc, upper = mod.fin))
```

Now, we can run our AIC to find which model should be appropriate to use as our final model. The full model contained all the covariates in the data set and was compared to the reduced model, which only contained 'yschool'. After running AIC, the model with the lowest AIC score of 4101.32 contained the covariates (~yschool + os30d + vagina + condom + abdpain + dchexam). We also applied the cox.zph() to check that the model follows the PH assumption. All the p-values are greater than 0.05, therefore we can accept the null hypothesis, which states that the model follows the PH assumption.

```
#Final model outputted by AIC
new.cox = coxph(std.surv ~ yschool + os30d + vagina + condom + abdpain + dchexam, data = std.new)
summary(new.cox)
```

```
## Call:
## coxph(formula = std.surv ~ yschool + os30d + vagina + condom +
##     abdpain + dchexam, data = std.new)
##
##   n= 877, number of events= 347
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## yschool -0.13450   0.87415  0.03380 -3.979 6.92e-05 ***
## os30d   -0.57990   0.55996  0.15225 -3.809  0.00014 ***
## vagina   0.41809   1.51906  0.16842  2.482  0.01305 *
## condom  -0.22012   0.80242  0.09316 -2.363  0.01814 *
## abdpain  0.28360   1.32790  0.14783  1.918  0.05506 .
## dchexam -0.41337   0.66142  0.22235 -1.859  0.06302 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## yschool    0.8742     1.1440    0.8181    0.9340
## os30d      0.5600     1.7859    0.4155    0.7547
## vagina     1.5191     0.6583    1.0920    2.1132
## condom     0.8024     1.2462    0.6685    0.9632
## abdpain    1.3279     0.7531    0.9939    1.7742
## dchexam    0.6614     1.5119    0.4278    1.0227
##
## Concordance= 0.624  (se = 0.017 )
## Rsquare= 0.063   (max possible= 0.991 )
## Likelihood ratio test= 57.21  on 6 df,   p=1.659e-10
## Wald test            = 54.75  on 6 df,   p=5.214e-10
## Score (logrank) test = 55.45  on 6 df,   p=3.76e-10
```

```
cox.zph(new.cox)
```

```
##              rho  chisq     p
## yschool  0.01313 0.0595 0.807
## os30d   -0.05193 0.9358 0.333
## vagina  -0.04637 0.7392 0.390
## condom   0.06788 1.4836 0.223
## abdpain -0.00911 0.0284 0.866
## dchexam  0.01772 0.1093 0.741
## GLOBAL        NA 3.4916 0.745
```

Although this model had the lowest AIC score, our summary of the model shows that 'abdpain' and 'dchexam' had p-values of 0.05506 and 0.06302, which means they are insignificant in the model.

To determine whether we can remove 'abdpain' and dchexam', we perform additional likelihood ratio tests:

Testing for significance of 'abdpain' by performing LRT on these two models:
(~yschool + os30d + vagina + condom + abdpain + dchexam) vs. (~yschool + os30d + vagina + condom + dchexam)
After calculating the appropriate LRT test statistic (3.460087, dof=1), the resulting p-value of $0.0629 > 0.05$ tells us that removing 'abdpain' did not make a significant difference to the model, therefore we can remove 'abdpain' from the final model.

Testing for significance of 'dchexam' by performing LRT on these two models:
(~yschool + os30d + vagina + condom + abdpain + dchexam) vs. (~yschool + os30d + vagina + condom + abdpain)

```
#new.cox1 has removed abdpain
new.cox1 <- coxph (std.surv ~ yschool + os30d + vagina + condom + dchexam, data = std.new)
#newcox2 has removed dchexam
new.cox2 <- coxph (std.surv ~ yschool + os30d + vagina + condom + abdpain, data = std.new)
lrt1 = 2*(new.cox$loglik[2]-new.cox1$loglik[2]) #test statistic of new.cox vs. new.cox1
lrt2 = 2*(new.cox$loglik[2]-new.cox2$loglik[2]) #test statistic of new.cox vs. new.cox2
pchisq(lrt1, df=1, lower.tail = FALSE)
```

```
## [1] 0.06286701
```

```
pchisq(lrt2, df=1, lower.tail = FALSE)
```

```
## [1] 0.07912472
```

After calculating the appropriate LRT test statistic (3.082789, dof=1), the resulting p-value of $0.0791 > 0.05$ tells us that removing 'dchexam' did not make a significant difference to the model, therefore we can also remove 'dchexam' from the final model.

```
#new model has removed abdpain and dchexam, this is our final model
new.mod <- coxph(std.surv ~ yschool + os30d + vagina + condom, data=std.new)
cox.zph(new.mod)
```

```
##             rho  chisq     p
## yschool  0.0111 0.0418 0.838
## os30d   -0.0508 0.8907 0.345
## vagina  -0.0464 0.7417 0.389
## condom   0.0764 1.9187 0.166
```

```
## GLOBAL       NA 3.6202 0.460
```

In the end, our final model contains the covariates (~yschool + os30d + vagina + condom) and we use the cox.zph function to check that our model upholds the proportional hazards assumption and whether stratification is neccesary for certain covariates. The resulting p-values of the covariates are all greater than 0.05, therefore we can accept our null hypothesis which states that the model follows the PH assumption.

Here is the hazard rate of our final Cox model:

$$h(t) = h_0(t) * exp[-0.13564X_1 - 0.55498X_2 - 0.44055X_3 - 0.22594X_4]$$

Where X_1 = 'yschool', X_2 = 'os30d', X_3 = 'vagina', and X_4 = 'condom'.

Interpretation of the Model

```
summary(new.mod)
```

```
## Call:
## coxph(formula = std.surv ~ yschool + os30d + vagina + condom,
##     data = std.new)
##
##   n= 877, number of events= 347
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## yschool -0.13574   0.87307  0.03352 -4.050 5.12e-05 ***
## os30d   -0.55498   0.57408  0.15177 -3.657 0.000255 ***
## vagina   0.44055   1.55357  0.16689  2.640 0.008295 **
## condom  -0.22594   0.79776  0.09368 -2.412 0.015866 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## yschool    0.8731     1.1454    0.8176    0.9323
## os30d      0.5741     1.7419    0.4264    0.7730
## vagina     1.5536     0.6437    1.1201    2.1547
## condom     0.7978     1.2535    0.6640    0.9585
##
## Concordance= 0.615  (se = 0.017 )
## Rsquare= 0.056   (max possible= 0.991 )
## Likelihood ratio test= 50.5  on 4 df,   p=2.838e-10
## Wald test            = 48.12  on 4 df,   p=8.925e-10
## Score (logrank) test = 48.7  on 4 df,   p=6.758e-10
```

Notice since the LRT, Wald Test, and score all have small p-value smaller than 0.05 we reject the null hypothesis that the model is not significant. Since, we the final model is significant then we are start interpreting the other values outputted from the summary function. Notice that

$$p_{yschool} = 5.12e^{-}05 < \alpha$$
$$p_{os30d} = 0.000255 < \alpha$$
$$p_{vagina} = 0.008295 < \alpha$$
$$p_{condom} = 0.015866 < \alpha$$

for significance level of 0.05. These p-values agree with the model selection process, and that the covariates chosen for the model are significant.

The yschool coefficient is negative. The hazard ratio for this covariate is 0.8731 and using its respective p-value one can see their is a strong relationship between not being reinfected and the amount of years one goes to school. Specifically for every year of school the hazard ratio decreases by factor of 0.8731.

The os30d coefficient is negative. The hazard ratio for this covariate is 0.57408 and using its respective p-value one can see their is a strong relationship between not being reinfected and if the patient has had oral sex in 30 days before testing. Not having oral sex 30 days before testing increasing chances for NOT testing positive. Specifically for every year of school the hazard ratio decreases by factor of 0.5741.

The condom coefficient is negative. The hazard ratio for this covariate is 0.7978 and using its respective p-value one can see is a strong relationship between not being reinfected and whether that individual used a condom. Specifically for every year of school the hazard ratio decreases by factor of 0.7978.

The Vagina coefficient is positive. The hazard ratio for this covariate is 1.5536 and using its respective p-value one can see is a strong relationship between being reinfected and whether had involvement vagina at exam. Specifically for every year of school the hazard ratio increases by factor of 1.5536. The confidence intervals for each covariate hazard ratio are as shown:

$$C.I_{yschool} = [2.24, 2.55]$$

$$C.I_{os30d} = [1.31, 2.39]$$

$$C.I_{Vagina} = [3.39, 6.53]$$

$$C.I_{condom} = [0.66, 0.95]$$

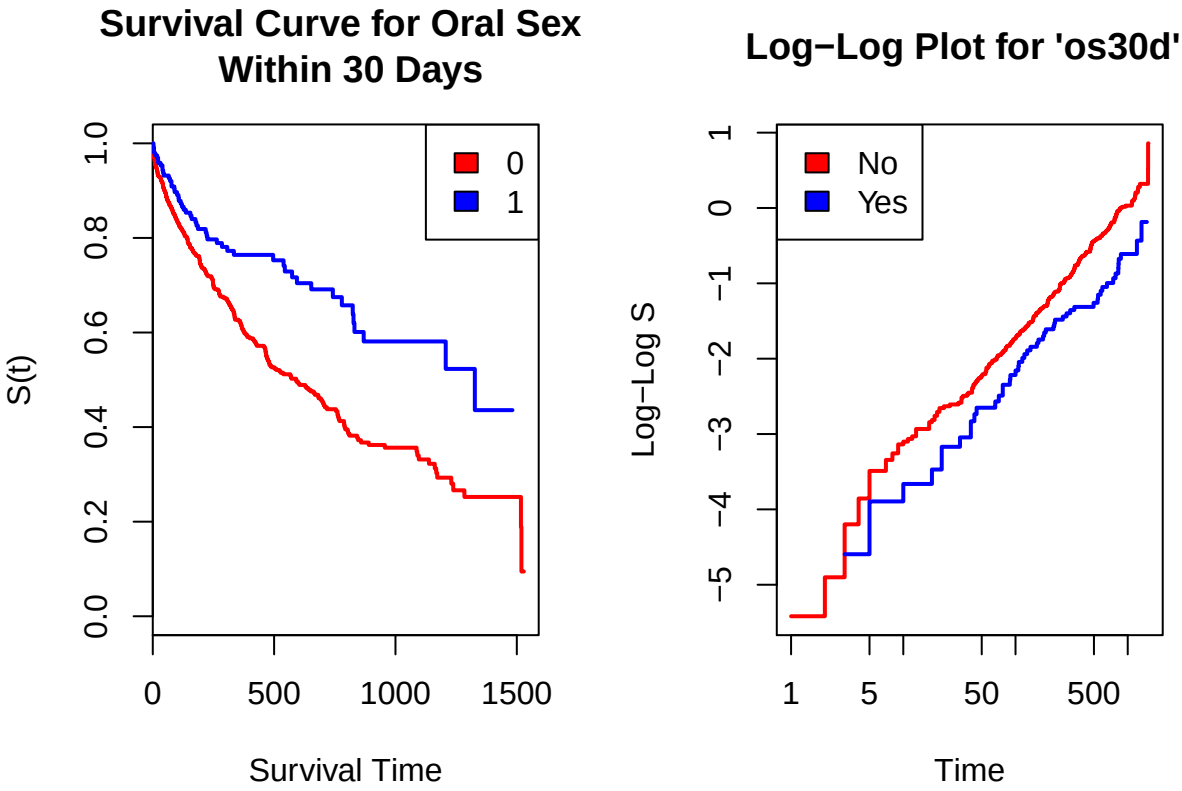
, thus confirming our claims about the relationships with reinfection.

Analysis of Covariates in the Model

The analysis of the covariates includes a Cox PH and clog-log plot of each symptom variable that was determined to be significant and included in the model (os30d, vagina and condom) in order to model the survival curves and test the PH assumption.

```
#converting into factor variables for cloglog plots
os30d.f <- factor(std.new$os30d, levels=c(0,1), labels=c("no","yes"))
vagina.f <- factor(std.new$vagina, levels=c(0,1), labels=c("no", "yes"))
condom.f <- factor(std.new$condom, levels=c(1,2,3), labels=c("always","sometimes","never"))
```

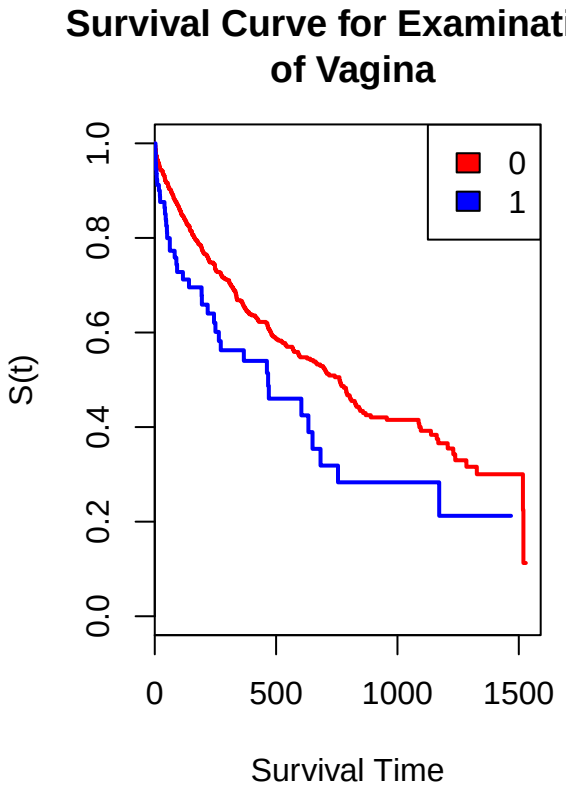
```
par(mfrow=c(1,2))
# Cox PH Plot of os30d:
std.surv <- Surv(std.new$time, std.new$rinfct)
plot(survfit(std.surv ~ os30d.f, data=std.new), xlab="Survival Time", ylab="S(t)",
     main = "Survival Curve for Oral Sex \n Within 30 Days", lwd=2,col=c(2,4))
legend("topright", c("0", "1"), fill=c(2,4))
#log-log plot of os30d:
plot(survfit(std.surv~os30d.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog",
     xlab="Time", ylab="Log-Log S", main= "Log-Log Plot for 'os30d' ")
legend("topleft",c("No", "Yes"),fill=c(2,4))
```

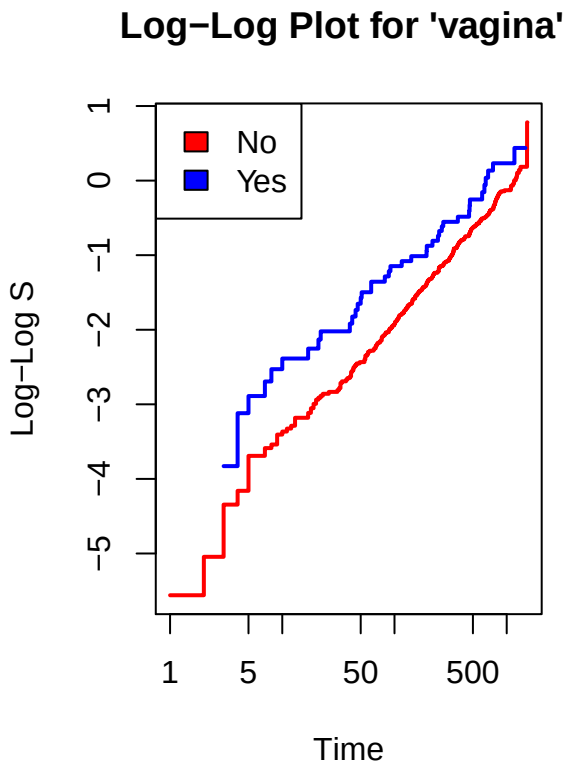


The plot of the Cox PH model for the covariate "os30d" (oral sex within 30 days) indicates that for the variable's levels (1=yes, 0=no), the survival curve for the level "1" is consistently greater than the survival curve for level "0". This means that the data indicates that having oral sex within 30 days has a higher survival probability (lower rate of STD reinfection) than not having oral sex within 30 days.

The clog-log plot for "os30d" shows no intersections within the plot (the curves are mostly parallel), which means that we can visually confirm that the PH assumption is not violated. This means that we can state that the proportional hazards assumption is appropriate for this covariate.

```
par(mfrow=c(1,2))
# Cox PH Plot of vagina:
fit.vagina <- coxph(std.surv ~ vagina.f, data = std.new)
plot(survfit(std.surv ~ vagina.f, data=std.new), xlab="Survival Time", ylab="S(t)",
     main = "Survival Curve for Examination \n of Vagina", lwd=2,col=c(2,4))
legend("topright", c("0", "1"), fill=c(2,4))
#log-log plot of vagina.f:
plot(survfit(std.surv~vagina.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog",
     xlab="Time", ylab="Log-Log S", main= "Log-Log Plot for 'vagina' ")
legend("topleft",c("No", "Yes"),fill=c(2,4))
```

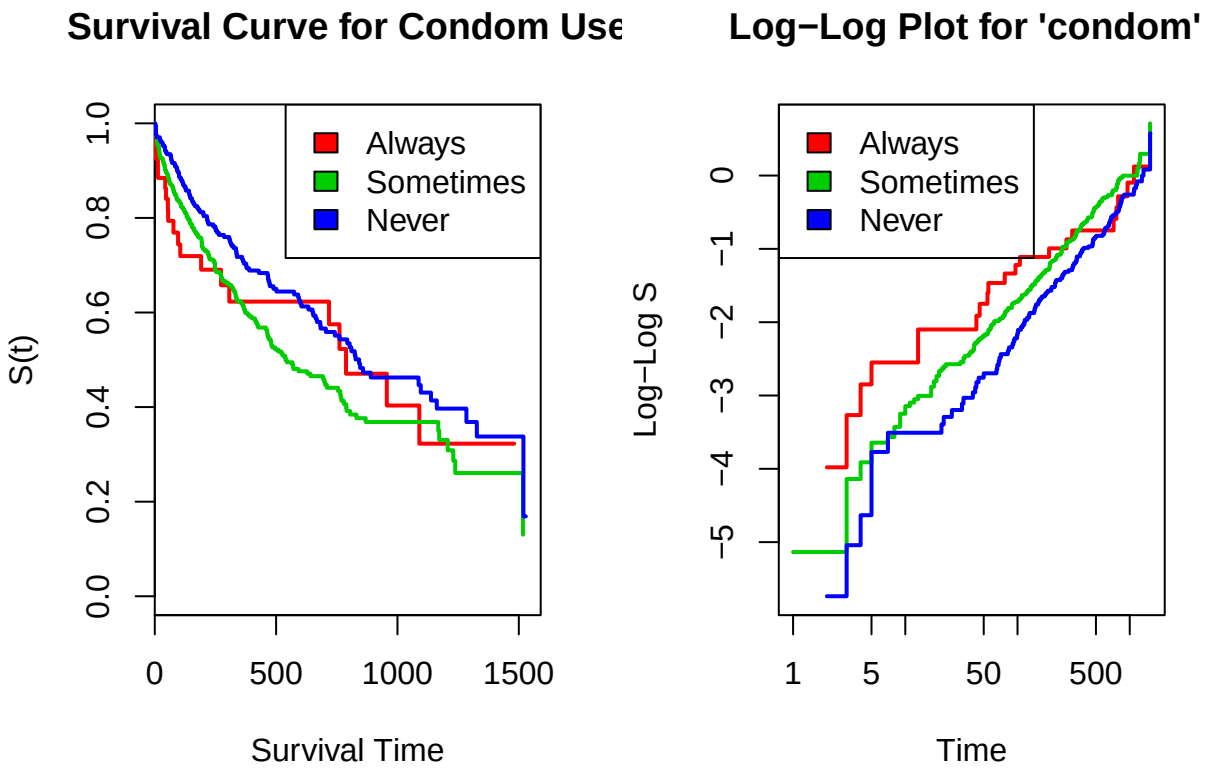




The Cox PH plot for the covariate "vagina"" (examination of vagina at doctor appointment (1=yes, 0=no)) below shows that the rate of reinfection for patients that had a vagina examination is consistenlty lower than that of patients who did not have a vagina examination, since the survival curve for "yes" is consistently lower than the survival curve for "no".

The clog-log plot for "vagina" illustrates that there are no intersections between the two plotted curves (for the levels "yes" and "no"). This means the PH assumption is being upheld.

```
par(mfrow=c(1,2))
# Cox PH Plot of condom:
plot(survfit(std.surv ~ condom.f, data=std.new), xlab="Survival Time", ylab="S(t)",
     main = "Survival Curve for Condom Use", lwd=2,col=c(2,3,4))
legend("topright", c("Always","Sometimes","Never"), fill=c(2,3,4))
#log-log plot of condom.f:
plot(survfit(std.surv~condom.f, data = std.new), lwd=2, col=c(2,3,4), fun="cloglog",
     xlab="Time", ylab="Log-Log S", main= "Log-Log Plot for 'condom' ")
legend("topleft",c("Always", "Sometimes","Never"),fill=c(2,3,4))
```



The Cox PH plot below for the covariate "condom" (condom use (1=always, 2=sometime, 3=never)) illustrates an interesting intersection between the three levels: always wearing a condom is only linked to a lower rate of reinfection (in comparison to sometimes and never wearing a condom) at one point in the middle of the plot (between time 500 and 1000). In fact, never wearing a condom is the level most consistently linked to a lower rate of reinfection of time (while sometimes wearing a condom is most consistently linked to a higher rate of reinfection).

The clog-log plot for "condom" indicates that there are no intersections that we should be concerned about (although there is an intersection at the beginning, and some intersection at the end of the plot, we are only concerned with the middle of the plot). Thus, we can conclude that the PH assumption is not being violated.

All the clog-log plots above show that the PH assumption is never violated, therefore none of these covariates need to be stratified and using a Cox PH model is appropriate.

```
pchisq(inter.lrt, df=2, lower.tail = FALSE)
```

After determining which covariates were significant, we wanted to test for significant interactions between them. We first created new cox objects and then performed LRT to compare models with and without the interaction term. After cycling through all the possible interactions, we found that the only significant interaction was between the covariates 'os30d' and 'condom'. This implies that the effect of having oral sex within 30 days on condom use significantly affects the rate of STD reinfection.

Research Questions

1. What is the probability of a patient going for a full year without experiencing reinfection?

```
summary(std.fit, std.fit$time[std.fit$time == est]) #to find 95% CI
```

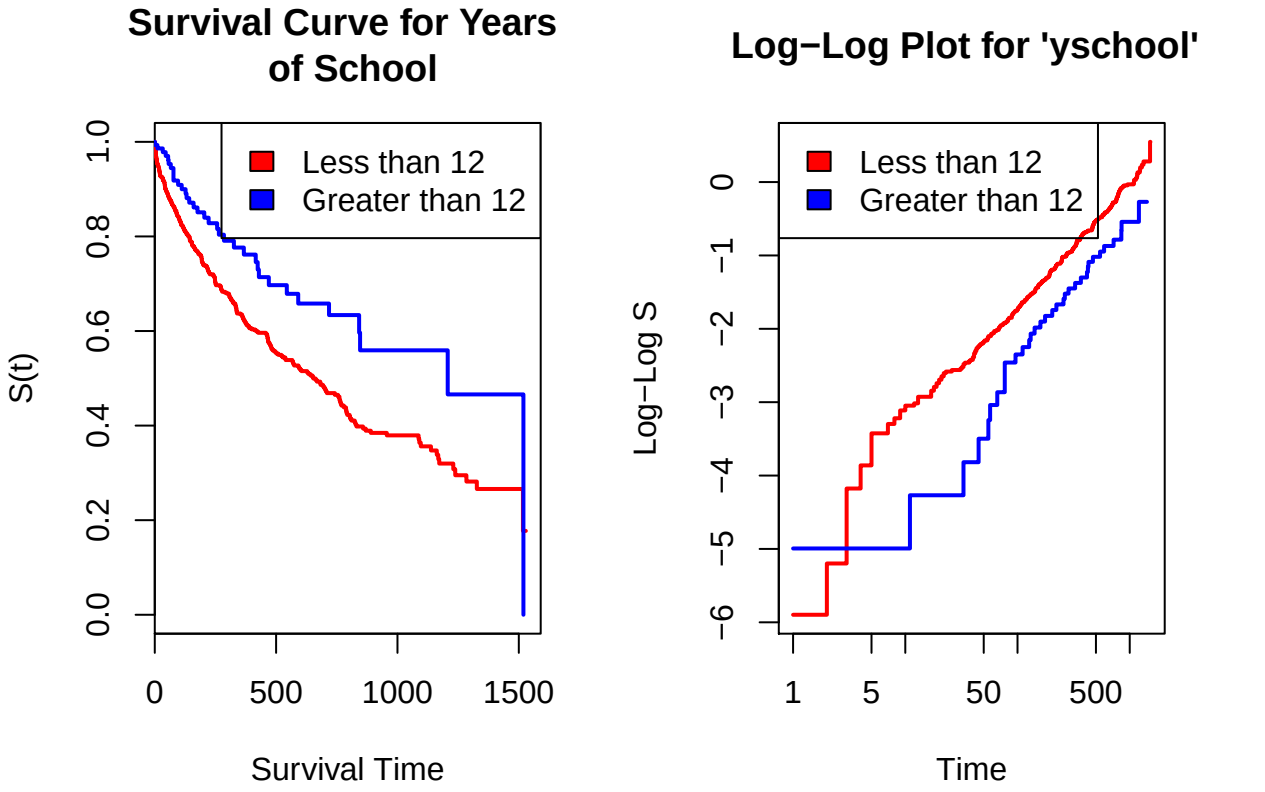
The probablity of a a patient going for a full year without experiencing reinfection is 0.648. We can also say with 95% confidence that the probability of a patient going for a full year without reinfection is between the

values of [0.613,0.686].

2. Is there a difference in the probability of STD reinfection between patients with low years of schooling and high years of schooling?

```
less12 <- std.new$age[std.new$yschool <= 12]
great12 <- std.new$age[std.new$yschool > 12]
yschool.f <- ifelse(std.new$yschool <= 12, 1,2)
yschool.f <- factor(yschool.f, levels=c(1,2), labels=c("less than 12", "greater than 12"))
school.cox <- coxph(std.surv ~ yschool.f, data=std.new)
summary(school.cox)
```

```
par(mfrow=c(1,2))
#Cox PH Plot for yschool.f
plot(survfit(std.surv ~ yschool.f, data=std.new), xlab="Survival Time", ylab="S(t)",
     main = "Survival Curve for Years \n of School", lwd=2,col=c(2,4))
legend("topright", c("Less than 12", "Greater than 12"), fill=c(2,4))
#test to see if model follows coxph assumption
cox.zph(school.cox) #follows coxph model
plot(survfit(std.surv~yschool.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog",
     xlab="Time", ylab="Log-Log S", main= "Log-Log Plot for 'yschool' ")
legend("topleft",c("Less than 12", "Greater than 12"),fill=c(2,4))
```



Here we get that for the difference of years of schooling STD reinfection rate has an LRT p-value of 0.0007347. Since this p-value is less than 0.05 ($0.0007347 < 0.05$) then we can conclude that there is a difference between the years of school. Since the hazard ratio is a negative value and our p-value is small we can conclude that there is a strong relationship between more than 12 years of school and having less. We are able to calculate a hazard ratio

$$[C_{yschool.fgreaterthan12} = 0.40, 0.81]$$

, which confirms our last claim.

We can see from the Cox PH above that the group of patients with more than 12 years of schooling, consistently has a higher survival probability than patients with less than 12 years of schooling. In addition, we tested to see if the Cox PH assumption was upheld by performing a cox.zph test and plotting a clog-log plot. The cox.zph test gave us a p=0.467, which is greater than 0.05 so we fail to reject the H0: Proportional hazards (PH) assumption is not violated, and thus we can conclude here that the Cox PH assumption is upheld. The clog-log plot visually supports this claim since there are no intersections between the two curves (except for one in the beginning, but we are only concerned with the beginning of the plot).

3. What is the probability of a low years of schooling patient going a full year without a reinfection?

```
std.fit2 <- survfit(std.surv ~ yschool.f)
summary(std.fit2, std.fit2$time[std.fit2$time == est]) #to find 95% CI
```

```
## Call: survfit(formula = std.surv ~ yschool.f)
##
##                 yschool.f=less than 12
##         time       n.risk      n.event     survival      std.err
##     364.0000     283.0000     221.0000       0.6261       0.0204
## lower 95% CI upper 95% CI
##       0.5873       0.6675
##
##                 yschool.f=greater than 12
##         time       n.risk      n.event     survival      std.err
##     364.0000      53.0000      23.0000       0.7765       0.0424
## lower 95% CI upper 95% CI
##       0.6977       0.8642
```

The probablity of a low years of schooling patient going a full year without a reinfection is 0.6261. We can also say with 95% confidence that the probability of a patient going for a full year without reinfection is between the values of 0.5873 and 0.6675. This is a smaller probability in comparison to a high years of schooling patient, who has a 0.7765 survival probability.

4. Given low and high years of schooling, how does condom use affect survival probability?

```
std.fit3 <- survfit(std.surv ~ yschool.f + condom.f, data=std.new)
summary(std.fit3, std.fit3$time[std.fit3$time == est]) #to find 95% CI
```

```
## Call: survfit(formula = std.surv ~ yschool.f + condom.f, data = std.new)
##
##                 yschool.f=less than 12, condom.f=always
##         time       n.risk      n.event     survival      std.err
##     364.0000      14.0000      14.0000       0.5794       0.0891
## lower 95% CI upper 95% CI
##       0.4287       0.7831
##
##                 yschool.f=less than 12, condom.f=sometimes
##         time       n.risk      n.event     survival      std.err
##     364.0000     141.0000     143.0000       0.5795       0.0277
## lower 95% CI upper 95% CI
##       0.5277       0.6364
##
##                 yschool.f=less than 12, condom.f=never
##         time       n.risk      n.event     survival      std.err
##     364.0000     128.0000      64.0000       0.7053       0.0315
## lower 95% CI upper 95% CI
##       0.6462       0.7697
##
```

```
##                 yschool.f=greater than 12, condom.f=always
##         time       n.risk      n.event     survival      std.err
##      364.000        3.000        2.000        0.800        0.126
## lower 95% CI upper 95% CI
##        0.587        1.000
##
##                 yschool.f=greater than 12, condom.f=sometimes
##         time       n.risk      n.event     survival      std.err
##     364.0000      30.0000      12.0000       0.8071       0.0524
## lower 95% CI upper 95% CI
##       0.7107       0.9166
##
##                 yschool.f=greater than 12, condom.f=never
##         time       n.risk      n.event     survival      std.err
##      364.000       20.000        9.000        0.726        0.079
## lower 95% CI upper 95% CI
##        0.586        0.898
```

For the group of subjects with less than 12 years of education, the survival probabilities of those that used a condom always, sometimes, and never are 0.5794, 0.5795, and 0.7053. For the group of subjects with more than 12 years of education, the surival probabilites of those that used a condom always, sometimes, and never are 0.800, 0.8071, and 0.726. Interestingly, within the group with less than 12 years of education, those who did not use a condom had a higher survival rate than those who did use a condom, which may be a result of other hidden factors that were not included in the data.

5. Which symptom contributes most to reinfection rate?

```
summary(new.mod)
```

```
## Call:
## coxph(formula = std.surv ~ yschool + os30d + vagina + condom,
##     data = std.new)
##
##   n= 877, number of events= 347
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## yschool -0.13574   0.87307  0.03352 -4.050 5.12e-05 ***
## os30d   -0.55498   0.57408  0.15177 -3.657 0.000255 ***
## vagina   0.44055   1.55357  0.16689  2.640 0.008295 **
## condom  -0.22594   0.79776  0.09368 -2.412 0.015866 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## yschool    0.8731     1.1454    0.8176    0.9323
## os30d      0.5741     1.7419    0.4264    0.7730
## vagina     1.5536     0.6437    1.1201    2.1547
## condom     0.7978     1.2535    0.6640    0.9585
##
## Concordance= 0.615  (se = 0.017 )
## Rsquare= 0.056   (max possible= 0.991 )
## Likelihood ratio test= 50.5  on 4 df,   p=2.838e-10
## Wald test            = 48.12  on 4 df,   p=8.925e-10
## Score (logrank) test = 48.7  on 4 df,   p=6.758e-10
```

'os30d' contributed the most to reinfection rate because it had the most significant p-value of 0.000255,

compared to the p-values of 'vagina'(p-value=0.008295) and 'condom'(p-value=0.015866. This means that having oral sex within 30 days had the most significant effect on STD reinfection compared to condom use and involvement of the vagina at a doctor's appointment.

Further Analysis: Fitting a Parametric Model

To perfom additional analysis, we decided to fit parametric models to our data. We fitted three models with a Weibull, Exponential, and Log-Normal distribution and used the LRT from the anova() function to determine which model had the best fit. When comparing the exponential to the Weibull model, the Weibull model was significant and when comparing the exponential to the log-normal model, the log-normal model was significant. We could not use the LRT from anova() function to compare the Weibull and log-normal model, due to the residuals having the same degrees of freedom.

```
#creating parametric models: weibull, exponential, log-normal
w.mod <- survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f + condom.f,
                 data = std.new, dist = "weibull")
e.mod <- survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f + condom.f,
                 data = std.new, dist = "exponential")
log.mod <- survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f + condom.f,
                   data = std.new, dist = "lognormal")
anova(e.mod, w.mod) #p-value 6.223076e-10 means weibull is better
```

```
##                                       Terms Resid. Df    -2*LL Test Df
## 1 yschool.f + os30d.f + vagina.f + condom.f       871 5392.393      NA
## 2 yschool.f + os30d.f + vagina.f + condom.f       870 5351.865    =  1
##   Deviance     Pr(>Chi)
## 1       NA           NA
## 2 40.52777 1.938416e-10
```

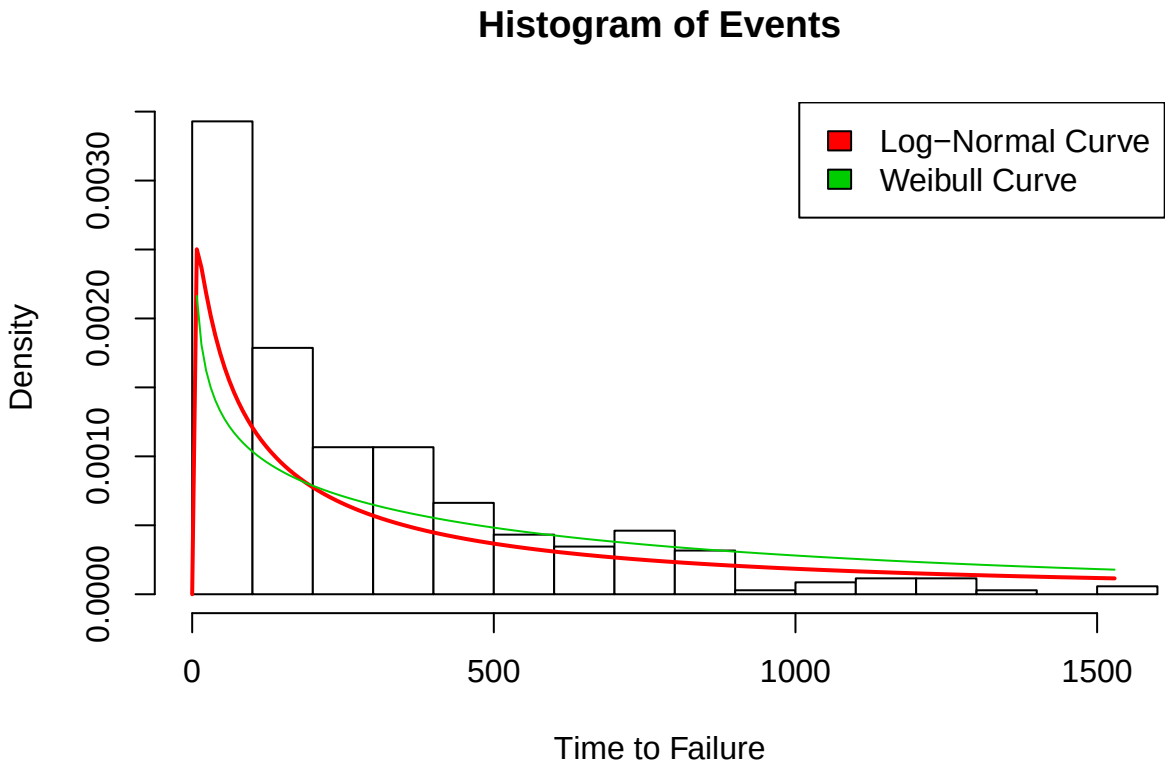
```
anova(e.mod, log.mod) #p-value 5.401376e-08 means log.mod is better
```

```
##                                       Terms Resid. Df    -2*LL Test Df
## 1 yschool.f + os30d.f + vagina.f + condom.f       871 5392.393      NA
## 2 yschool.f + os30d.f + vagina.f + condom.f       870 5378.140    =  1
##   Deviance     Pr(>Chi)
## 1       NA           NA
## 2  14.2534 0.0001597722
```

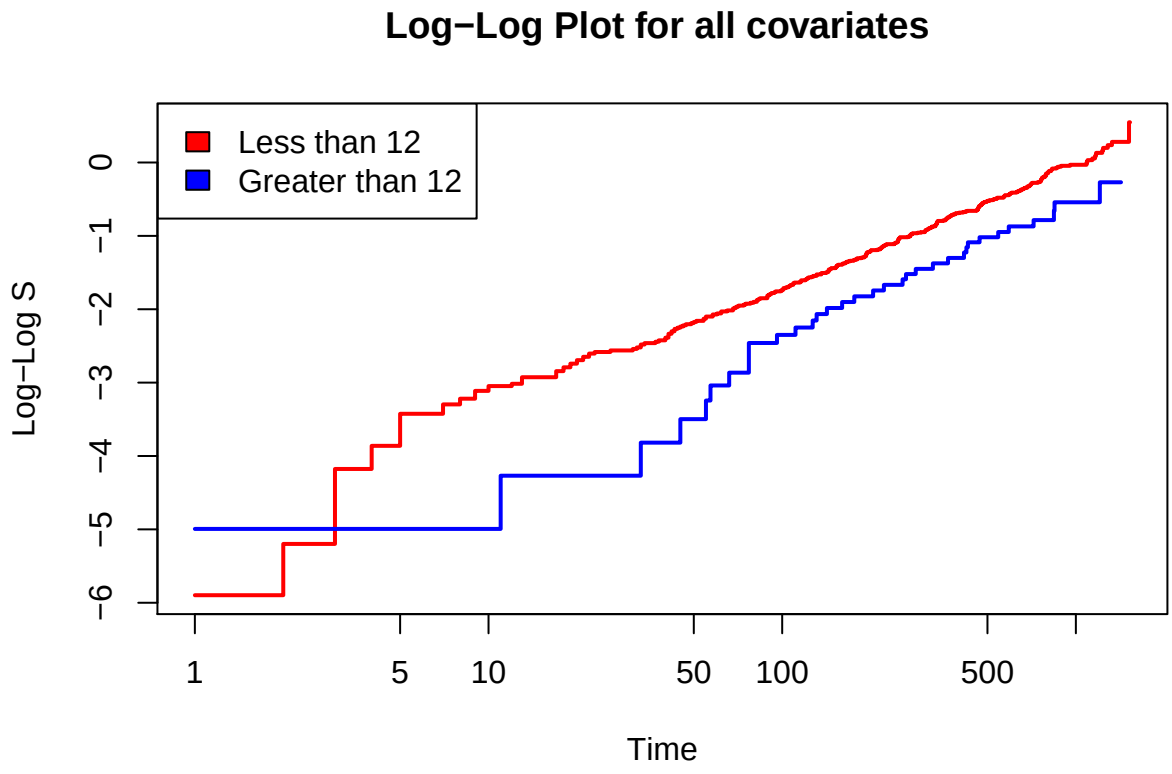
```
anova(w.mod, log.mod)  #no p-value bc the dof were the same
```

```
##                                       Terms Resid. Df    -2*LL Test Df
## 1 yschool.f + os30d.f + vagina.f + condom.f       870 5351.865      NA
## 2 yschool.f + os30d.f + vagina.f + condom.f       870 5378.140    =  0
##    Deviance Pr(>Chi)
## 1        NA       NA
## 2 -26.27437       NA
```

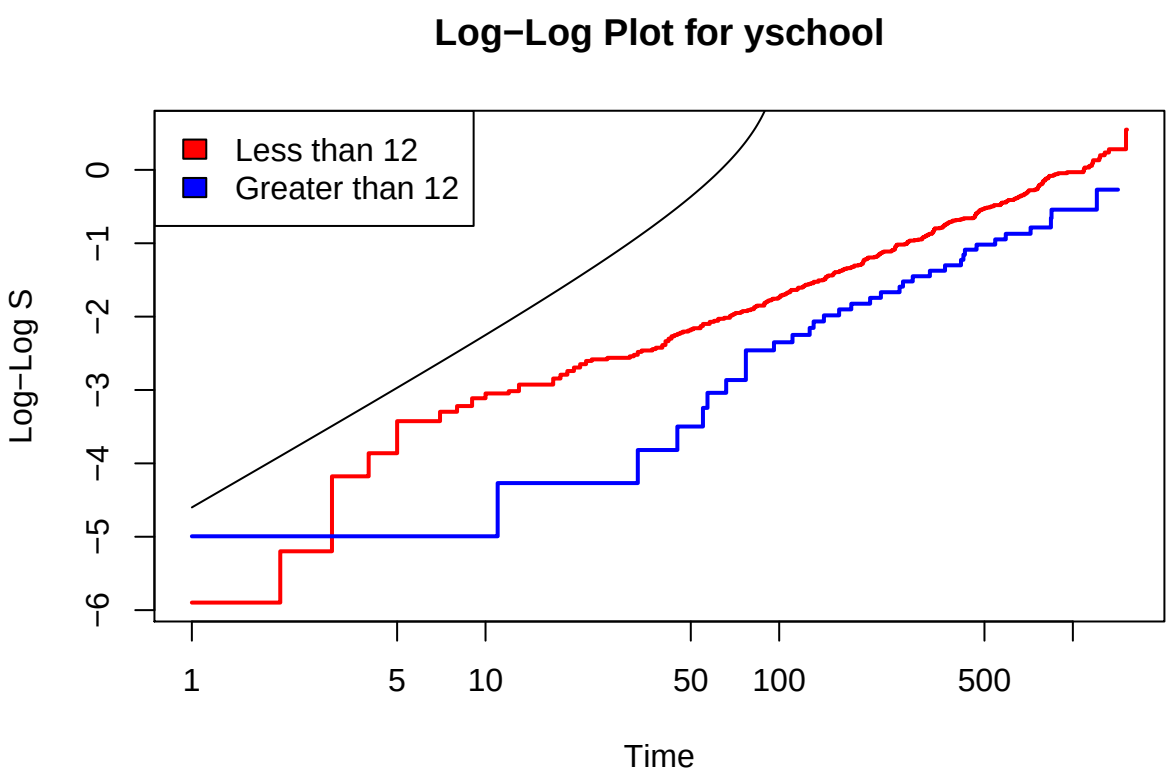
```
#plot weibull and log on histogram to see which one is better
fts <- seq(0,1529,length=200)
hist(std.new$time[std.new$rinfct ==1], breaks=15,col="white", xlab="Time to Failure",
     main="Histogram of Events", probability = TRUE, xlim=c(0,1550))
lines(fts, dlnorm(fts, meanlog= log.mod$icoef[1], sdlog=exp(log.mod$icoef[2])), lwd=2, col=2)
lines(fts, dweibull(fts, shape= 1/w.mod$scale, scale=exp(w.mod$icoef[1])), col=3)
legend("topright",c("Log-Normal Curve", "Weibull Curve"),fill=c(2,3))
```



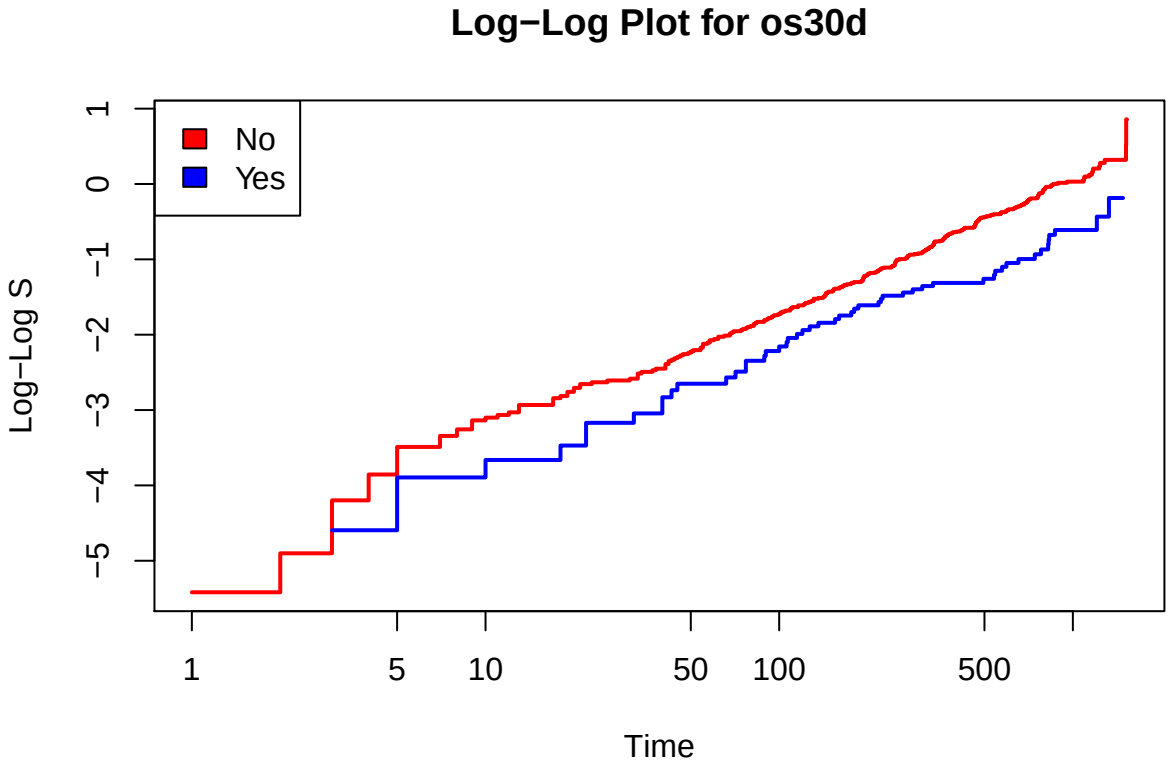

```
#clogclog plot with all the variables
plot(survfit(std.surv~yschool.f, data = std.new), lwd=2, col=c(2,4,3,5), fun="cloglog",
     xlab="Time", ylab="Log-Log S", main= "Log-Log Plot for all covariates ")
legend("topleft",c("Less than 12", "Greater than 12"),fill=c(2,4,3,5))
```



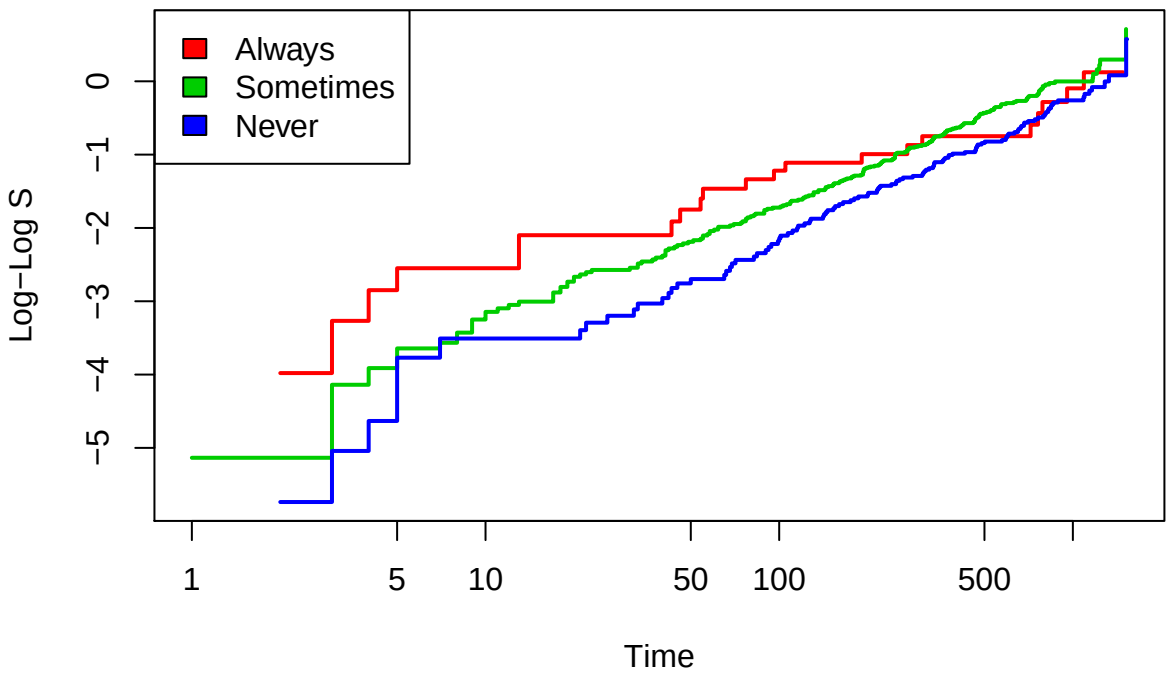
```r
#creating two models with lognormal and weibull distribution
amlfit <- survreg(Surv(time,rinfct)~yschool.f,data=std.new,dist="lognormal")
amlwfit <- survreg(Surv(time,rinfct)~yschool.f,data=std.new,dist="weibull")
# clog-log plots to check weibull vs log distribution (one for each covariate)
ps <- seq(0.01,0.99,by=0.01)

#cloglog for yschool
plot(survfit(Surv(time,rinfct)~yschool.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog", xlab="Time
legend("topleft",c("Less than 12", "Greater than 12"),fill=c(2,4))

#breaking apart the lines to troubleshoot
lines(log(-log(1-ps))) #this plots as the log curve
#these 2 lines plot as a straight line on the y=0 axis and are not plotting correctly
lines(predict(amlwfit,data.frame(x=yschool.f[1:877]),type="quantile",p=ps), col=3)
lines(predict(amlwfit,data.frame(x=yschool.f[1:877]),type="quantile",p=ps), col=6)
```



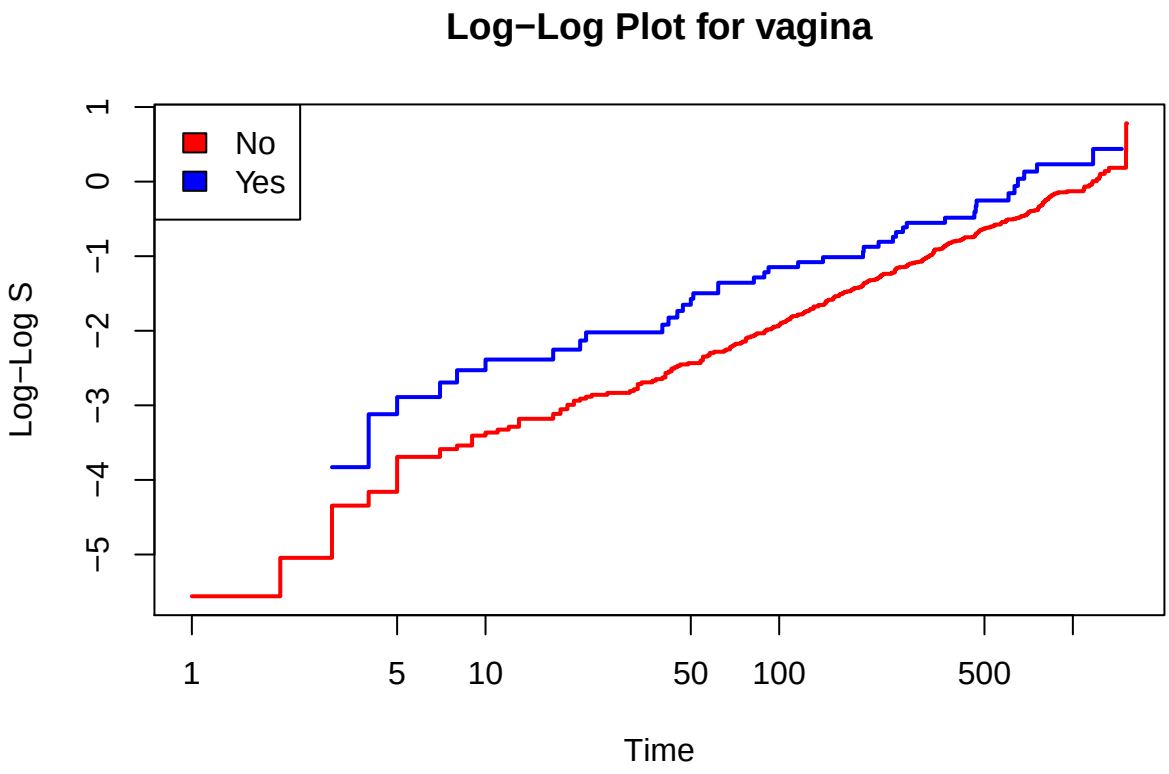

```r
#lines(predict(amlwfit,data.frame(x=std.new$yschool),type="quantile",p=ps),log(-log(1-ps)),lwd=3,col="p
#lines(predict(amlwfit,data.frame(x=std.new$yschool),type="quantile",p=ps),log(-log(1-ps)),lwd=3,col="g

plot(survfit(std.surv~os30d.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog", xlab="Time", ylab="Log
legend("topleft",c("No", "Yes"),fill=c(2,4))
```




```
plot(survfit(std.surv~condom.f, data = std.new), lwd=2, col=c(2,3,4), fun="cloglog", xlab="Time", ylab=
legend("topleft",c("Always", "Sometimes","Never"),fill=c(2,3,4))
```




```
plot(survfit(std.surv~vagina.f, data = std.new), lwd=2, col=c(2,4), fun="cloglog", xlab="Time", ylab="L
legend("topleft",c("No", "Yes"),fill=c(2,4))
```




```
w.mod.condom <- survreg(Surv(time, rinfct) ~ condom, data = std.new, dist = "weibull")
w.mod <- survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f + condom.f, data = std.new, dist =
w.mod2 <- survreg(formula = std.surv ~ yschool + os30d + vagina + condom, data = std.new, dist = "weibu
log.mod <- survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f + condom.f, data = std.new, dist
```

To compare the Weibull and log-normal model, we plotted the functions over the histogram of our uncensored observations and saw that the log-normal curve followed the distribution better than the Weibull curve, therefore we decided that the log-normal model fit our data the best.

```
#further analysis with best parametric model
summary(log.mod)
```

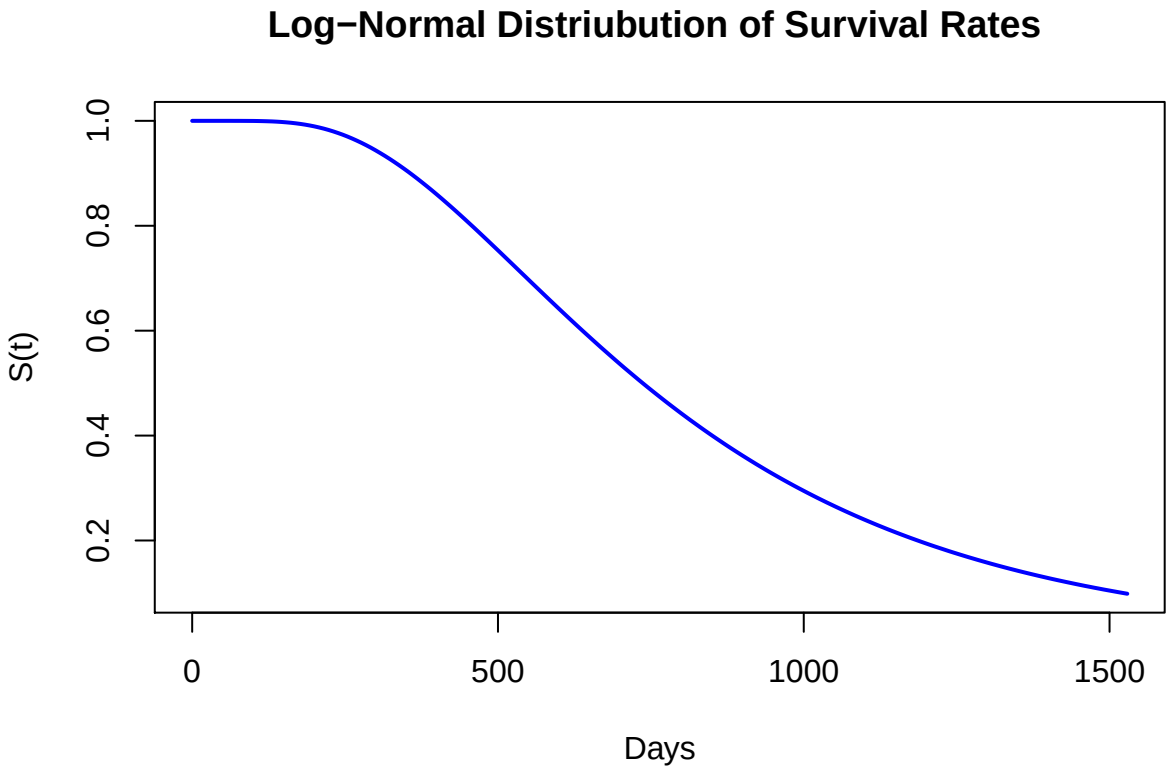
```
##
## Call:
## survreg(formula = std.surv ~ yschool.f + os30d.f + vagina.f +
##     condom.f, data = std.new, dist = "lognormal")
##                             Value Std. Error      z        p
## (Intercept)                 5.900     0.3530 16.713 1.05e-62
## yschool.fgreater than 12    0.767     0.2595  2.957 3.11e-03
## os30d.fyes                  0.702     0.2206  3.181 1.47e-03
## vagina.fyes                -0.731     0.2712 -2.695 7.04e-03
## condom.fsometimes           0.353     0.3617  0.977 3.29e-01
## condom.fnever               0.808     0.3741  2.160 3.07e-02
## Log(scale)                  0.729     0.0397 18.371 2.23e-75
##
## Scale= 2.07
##
## Log Normal distribution
## Loglik(model)= -2689.1   Loglik(intercept only)= -2709.4
##  Chisq= 40.67 on 5 degrees of freedom, p= 1.1e-07
## Number of Newton-Raphson Iterations: 4
```

```
## n= 877
```

```
anova(log.mod)
```

```
##           Df  Deviance Resid. Df    -2*LL     Pr(>Chi)
## NULL      NA        NA       875 5418.806           NA
## yschool.f  1 12.909610       874 5405.897 0.0003268993
## os30d.f    1 12.572490       873 5393.324 0.0003914667
## vagina.f   1  6.995263       872 5386.329 0.0081725718
## condom.f   2  8.189218       870 5378.140 0.0166622621
```

```
mean.log <- mean(log.mod$linear.predictors)
sd.log <- sd(log.mod$linear.predictors)
plot(fts,plnorm(fts,meanlog=mean.log ,sdlog=sd.log, lower.tail = FALSE),
xlab="Days",ylab="S(t)", main="Log-Normal Distriubution of Survival Rates ", type="l", lwd=2, col=4)
```



After determining that the log-normal parametric model fit our data the best, we plotted the estimated survival curve under the model.

Citations

Klein, J. P., & Moeschberger, M. L. (1997). Survival analysis: Techniques for censored and truncated data. New York: Springer.

Cox Proportional-Hazards Model. (2016, December 12). Retrieved from https://www.r-bloggers.com/cox-proportional-hazards-model/