Survival Analysis - Project

Vincent Buekers - r0754046

June 13, 2019

Contents

1 Introduction

The data at hand consists of patients registered on a waiting list for organ transplants. More specifically, UK residents diagnosed with either Chronic Obstructive Pulmonary Disease (COPD) or Fibrosis, both of which are forms of lung diseases. Additional measurements on the patients include age at registration, gender, Body Mass Index, survival time, death or censoring status, and of lesser interest also their ID-number. Logically, it is very valuable to indicate the survival time of such subjects as to examine the urgency of the organ transplant. R has been used to conduct the analysis, of which the code has been added in appendix I. Note that throughout this report, a significance level of $\alpha = 0.05$ is used.

2 Question 1

2.1 Kaplan-Meier Estimate of Survival function

In order to gain some preliminary insights into the survival of the patients on the waiting list, an non-parametric estimate of the survival function is obtained by the Kaplan-Meier Estimator. This estimator evaluates the conditional probability of surviving until t_{j+1}, given that a patient has survived until t_j. Note that since there are censored observations in the data, the K-M estimator does not reduce to the empirical survival function. The curve corresponding to these estimates is graphically represented in Figure 1.

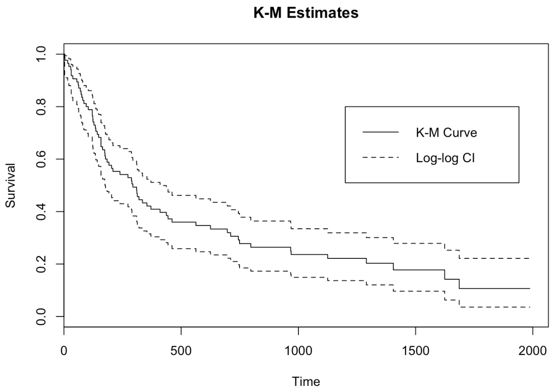


Figure 1: Kaplan-Meier estimate of survival curve with pointwise log-log confidence band

The probability to survive seems to quickly decrease within the first 500 days of enlistment, whereas the decline becomes less steep for patients who have been on the list for longer periods of time. The time at which only 35% of patients are still alive is presumed to be at 563 days on the waiting list, according to the Kaplan-Meier estimates.

Based on the Kaplan-Meier curve obtained above, an estimate for the residual lifetime can also be computed. Of particular interest is the probability that a patient survives for 300 days given that they have already been on the waiting list for 500 days. Assuming these estimates are independent at different time points, this probability is estimated to be about 0.097 with an associated approximate standard error of 0.045.

3 Question 2

3.1 Gender effect on Survival time

In order to assess the effect of gender on a patients survival time, it is possible to compare the survival function between males and females. The following hypothesis is tested, which is approximately χ^2 distributed with df = 1 under H_0.

$$H_0 : S_1(t) = S_2(t) = ... = S_K(t), 0 < t < \tau$$

In doing so, the log-rank test is often used. However, as it is desired to discriminate based on earlier differences between the genders, a more appropriate choice is the Wilcoxon-Gehan or Peto-Peto test. These are essentially weight adjusted modifications of the log-rank test. According to the Peto-peto test, the difference in survival probability between males and females is considered to be significantly different from zero ($\chi^2 = 4.6, df = 1, p = 0.03$). These R output can be found in appendix II.1. Hence, gender does impact the survival probability of the enlisted patients. Graphically, this can also be shown, as is done in Figure 2 below. Clearly, the survival probability is higher for females across the entire timeframe, increasingly so for larger waiting times.

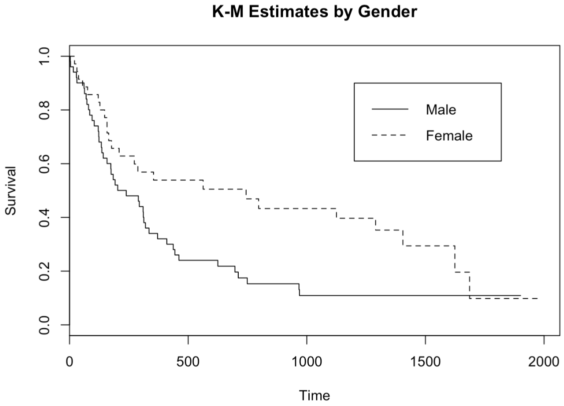


Figure 2: Kaplan-Meier estimate of the empirical survival curve by gender

For instance, the effect of gender is indeed already significant after 1 year of registration (i.e. 365 days assuming a non-leap year). This can be derived from the results obtained from a Peto-peto test on patients who have been enlisted for over a year ($\chi^2 = 5.1, df = 1, p = 0.02$). Thus, the difference in survival probability seems to manifest itself quite early on. Results are included in appendix II.2.

3.2 Disease as a Confounding Variable

Whereas in the previous section, the effect of gender was examined, now also the type of lung disease is adjusted for. Consequently, the shape of the survival function is different for the types of disease. Also, this leads to seperate hypotheses being tested for COPD and Fibrosis, namely:

$$H_0 : S_{1m}(t) = S_{2m}(t) = ... = S_{Km}(t), 0 < t < \tau, m = 1, 2$$

Although the disease corrected effect of gender is still considered significant ($\chi^2 = 4.4, df = 1, p = 0.04$), the graphical assessment does reveal a difference in this effect between COPD and fibrosis as shown in figure 3. The gender effect seems to be a lot larger for the COPD patients as opposed to for the fibrosis patients. R output is given in appendix II.3.

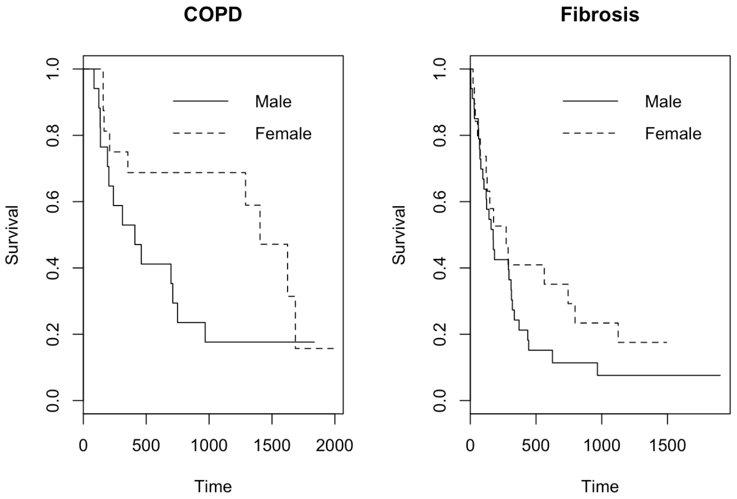


Figure 3: Gender differences in survival curve corrected for disease

4 Question 3

4.1 Cox Regression

In this section, the effect of gender is studied using semi-parametric methods, by means of a proportional hazard model. Several other covariates such as age, bmi and disease are also considered. The functional form of the model is taken to be additive in this case (1). Although the previous section hinted at an interaction between gender and disease, it is not considered significant in an interactive model (appendix II).

$$\lambda(t|age, gender, bmi, disease) = \lambda_0(t)e^{\beta_1 age+\beta_2 gender+\beta_3 bmi+\beta_4 disease} \tag{1}$$

The results corresponding to (1) are that both gender and disease significantly impact the proportional hazard (appendix II.4). This was also indicated by the non-parametric methods discussed in preceding sections. The coefficient estimates are respectively $\beta_{gender} = 0.55$ and $\beta_{disease} = 0.93$. With parsimonious modelling in mind, one could remove the covariates age and bmi. The estimates will change sightly, namely to $\beta_{gender} = 0.52$ and $\beta_{disease} = 0.73$. The hazard ratio for male patients in comparison to female patients is $HR(t) = e^{0.55} = 1.741$. This means that males have a significantly larger hazard to die than females, proportional to the baseline. Note that in order to obtain this estimate, one has to make sure females are taken as the reference class. Without specifying this in R, the reference class would have been male patients. By this logic, COPD is taken as the reference class for disease since left unaltered.

4.2 Prediction

Of interest are the predictions for the median survival times for two different patients. The first patient is a male fibrosis patient aged 60 with a bmi of 23, while the second patient is a female COPD patient aged 40 with a bmi of 19. The predicted median survival times are 148 and 1125 days respectively. Graphically, these predictions correspond to the survival curves shown in figure 4. Clearly, patient 1 is at severe risk compared to patient 2 and should be scheduled for organ transplant as soon as possible.

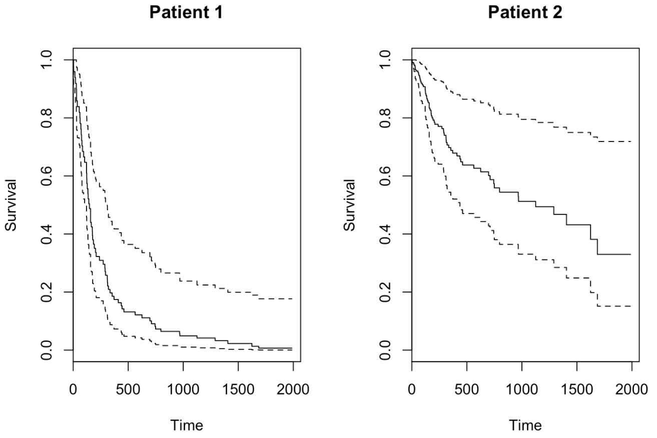


Figure 4: Survival curves or patient 1 and patient 2

4.3 Residual Diagnostics

Next, it is recommendable to take a look at the residuals of (1) as to assess whether or not the Cox regression model is appropriate for the patient data. To that end, the martingale residuals are examined. Figure 5 displays these residuals and the model fit appears adequate by looking at the upper plots. There seems to be no correlation among them and they do sum to zero overall. However, there seems to be a slight problem with heteroscedasticity, possibly induced by some of the outliers.

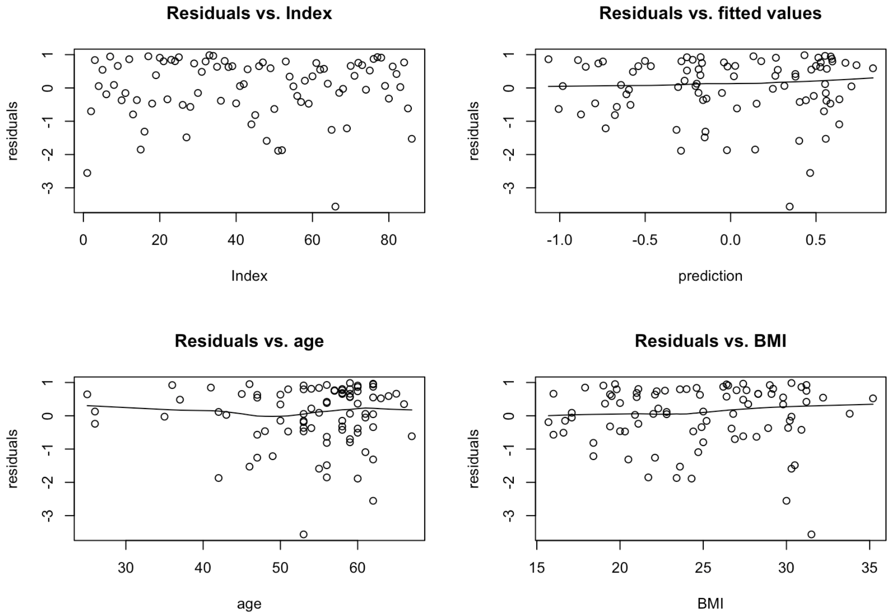


Figure 5: Martingale residuals. Top left: vs. index, top right: vs. fitted values, bottom: vs. continuous covariates

Next, the proportional hazard assumption remains to be tested for the discrete covariates gender and disease. To that end, a graphical check can be performed by plotting $log(-log(S(t|X = x))$ vs. $log(t)$ as is done in figure 6. With respect to the covariate gender, this assumption seems to be satisfied as the curves are approximately parallel. Yet the assumption does not seem to hold for the covariate disease in the earlier time periods. In an attempt to solve this, one can adopt an alternative modelling strategy by letting the baseline hazard differ for both COPD and fibrosis. This is known as a stratified Cox model (appendix II).

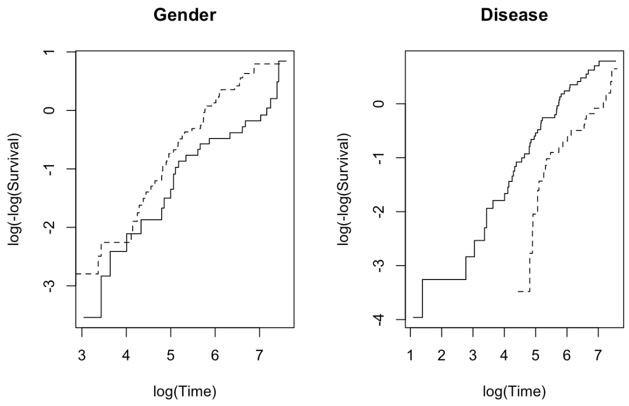


Figure 6: Proportional hazard verification plots

5 Question 4

5.1 Accelerated Failure Time model

After the non-parametric and semi-parametric sections discussed above, a last step in this analysis involves estimating a model making parametric assumptions on the error distribution similar to those in classical linear regression. Once again, the considered predictor variables are age, gender, BMI and disease. Also different distributions are considered as candidate for the baseline hazard function, namely the Weibull, exponential and Log-normal distribution. Based on figure 7, it is quite evident that the exponential distribution is the best choice for this model.

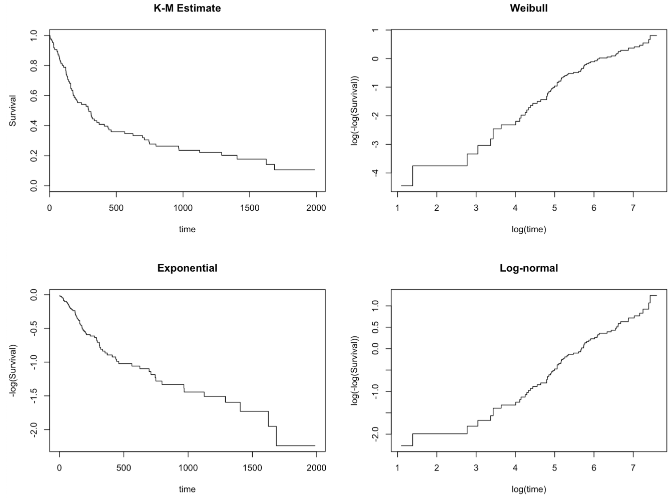


Figure 7: Distributional fit

Consequently, the functional form of the exponential accelerated failure time model is as follows:

$$log(T) = \mu + \beta_1 age + \beta_2 gender + \beta_3 bmi + \beta_4 disease + E \tag{2}$$

The results for (2) are in fact similar to those obtained for the regular Cox regression model in section 4. Again, gender and disease are considered significant predictors with respect to the log-scale of time to event $log(T)$. The coefficient estimates are $\beta_{gender} = -0.73$ and $\beta_{disease} = -1.08$, with corresponding p-values of $p = 0.006$ and $p = 0.009$ respectively. As before, age and BMI do not offer predictive performance to the model. Note the reference class for gender is taken to be the female class again for comparative purposes, while the reference class for disease also remains COPD. R output can be found in appendix II.5.

The indirect effect disease has on the survival function through its effect on the time scale can be computed as $exp(-\beta_{disease}) = exp(-1.08) = 2.95$. This means that the patients diagnosed with fibrosis, the survival process is expected to accelerate by nearly triple. Hence the survival time for fibrosis patients shrinks drastically compared to the COPD patients.

5.2 Prediction

Once again a prediction is made for the 60 year old male Fibrosis patient with a BMI of 23 (patient 1). Yet now the prediction estimates are based on the exponential accelerated failure time model discussed above. The obtained curve is shown in figure 8. This result is quite different from the estimate in section 4.2. The time to event, or death rather, is assumed to be a lot earlier on as the survival probability reaches 0 much sooner in these estimates. The rapid decline within the first 500 days does bear resemblance to the curve from section 4.2.

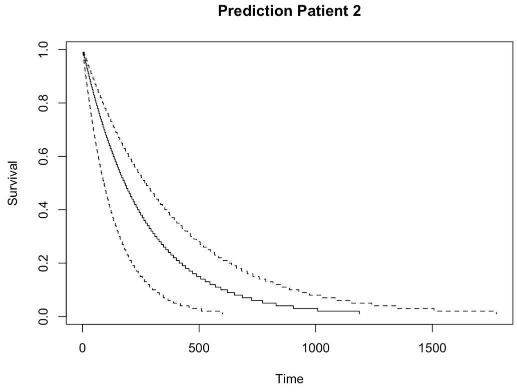


Figure 8: Prediction for patient 1, with pointwise confidence bands

6 Appendix I - R code

```
setwd("~/Desktop/Survival␣&␣Reliability")
library(survival)

data = read.table('LungNew.txt', header = TRUE, dec = '.')
n = dim(data)[1]
p = dim(data)[2]

#-----------
# Question 1
#-----------
# Kaplan-Meier Estimate
KM = survfit(Surv(time, status)~1,conf.type='log-log', data)
summary(KM)

plot(KM, main="K-M␣Estimates", xlab = "Time", ylab = "Survival")
legend(1200, 0.8, lty = c(1,2), legend=c("K-M␣Curve","Log-log␣CI"))

# time at which only 35% of patients are still alive
 according to KM estimates
KM$time[which(KM$surv<=.35)][1]

# Residual lifetime
plot(res.lifetime, type = "s")

res_lifetime = KM$surv[which(KM$time>500)][1] -
KM$surv[which(KM$time>800)][1]

se = KM$std.err[which(KM$time>800)][1] -
KM$std.err[which(KM$time>500)][1]

#-----------
# Question 2
#-----------
# Kaplan-Meier Estimates based on gender
KM2 = survfit(Surv(time, status)~gender,data)
summary(KM2)

plot(KM2, main="K-M␣Estimates␣by␣Gender"
, xlab = "Time", ylab = "Survival", lty=c(1,2))
legend(1200, .9, lty = c(1,2), legend=c("Male","Female"))

survdiff(Surv(time, status)~gender, rho=1 ,data)
```

```
# Patients on list for over a year
survdiff(Surv(time[which(time>365)],
status[which(time>365)])~gender[which(time>365)], rho=1 ,data)

# Stratified by disease
KM3 = survfit(Surv(time, status)~gender+strata(disease),data)
summary(KM3)

par(mfrow=c(1,2))
plot(KM3[c(1,3)], lty = c(1,2), main="COPD", xlab="Time", ylab = "Survival")
legend(500, 1, legend=c("Male", "Female"),lty = c(1,2), bty = "n")
plot(KM3[c(2,4)], lty=c(1,2), main="Fibrosis", xlab="Time", ylab = "Survival")
legend(500, 1, legend=c("Male", "Female"),lty = c(1,2), bty="n")

# correct for disease
survdiff(Surv(time, status)~gender+strata(disease),data)

# trend test
trend = survdiff(Surv(time, status)~disease, data)
OB = trend$obs
EX = trend$exp
V = trend$var
a = c(1,2)

test = a%*%(OB-EX)
stderror<-sqrt(t(a)%*%V%*%a)
zscore = test/stderror
Pvalue = 2*pnorm(abs(zscore),lower.tail=FALSE)
data.frame(test,stderror,zscore,Pvalue)

#--------------
# Question 3
#--------------
# Cox Regression Model (Additive)
fit1 = coxph(Surv(time, status)~ age + (gender==1) + bmi + disease, data)
cox.summary = summary(fit1)
cox.summary

# Parsimonious model
fit2 = coxph(Surv(time, status)~(gender==1) + disease, data)
summary(fit2)
anova(fit1, fit2)

# Interaction model
cox.int = coxph(Surv(time, status)~ age + gender*disease + bmi, data)
summary(cox.int)
```

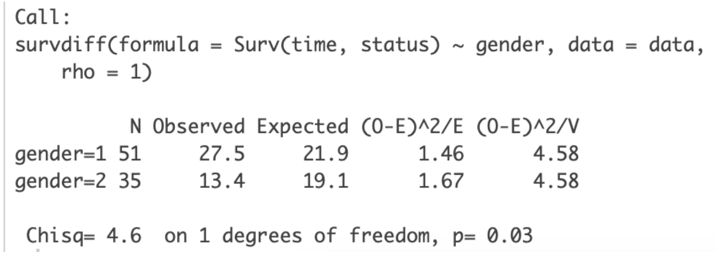
```
# Hazard Ratio Males
HR_Male = cox.summary$coefficients[2,2]
HR_Male

# Predictions for patient 1 & 2
patient1 = survfit(coxph(Surv(time, status)~age+gender+bmi+disease, data)
                   , type='breslow'
                   , newdata = data.frame(gender=1, disease=2, age=60, bmi=23))
patient1
plot(patient1, main="Patient␣1",  xlab = "Time", ylab = "Survival")

patient2 = survfit(coxph(Surv(time, status)~age+gender+bmi+disease, data)
                   , type='breslow'
                   , newdata = data.frame(gender=2, disease=1, age=40, bmi=19))
patient2
plot(patient2, main="Patient␣2",  xlab = "Time", ylab = "Survival")

# Residual diagnostics
fit.res = residuals(fit1,type="martingale")
fit.pred = predict(fit1)

par(mfrow=c(2,2))
plot(fit.res, ylab="residuals", main="Residuals␣vs.␣Index")

plot(fit.pred,fit.res,xlab="prediction",ylab="residuals",
main="Residuals␣vs.␣fitted␣values")
lines(lowess(fit.pred,fit.res))

plot(data$age,fit.res,xlab="age",ylab="residuals", main='Residuals␣vs.␣age')
lines(lowess(data$age,fit.res))

plot(data$bmi,fit.res,xlab="BMI",ylab="residuals", main='Residuals␣vs.␣BMI')
lines(lowess(data$bmi,fit.res))

# proportional hazard assumption
fit.gender = survfit(Surv(time,status) ~ gender, data)

par(mfrow=c(1,2))
plot(log(fit.gender[2]$time),log(-log(fit.gender[2]$surv)),xlab="
log(Time)",ylab="log(-log(Survival)",
     type="s", main = "Gender")
lines(log(fit.gender[1]$time),log(-log(fit.gender[1]$surv)),lty=2,type="s")

fit.disease = survfit(Surv(time,status) ~ disease, data)
```

```
plot(log(fit.disease[2]$time),log(-log(fit.disease[2]$surv)),xlab
="log(Time)",ylab="log(-log(Survival)", type="s", main="Disease")
lines(log(fit.disease[1]$time),log(-log(fit.disease[1]$surv)),lty
=2,type="s")

# stratified cox by strata of disease
cox.strat = coxph(Surv(time, status)~ age + gender + bmi +
strata(disease), data)
summary(cox.strat)

#-----------
# Question 4
#-----------
# Accelarated failure time models
aft_weibull = survreg(Surv(time, status)~ age+gender+bmi+disease, data)
summary(aft_weibull)

aft_exp = survreg(Surv(time, status)~
age+(gender==1)+bmi+disease, dist = 'exponential', data)
summary(aft_exp)

exp(-aft_exp$coefficients[5])

aft_lognorm = survreg(Surv(time, status)~ age+gender+bmi+disease,
dist = 'lognormal', data)
summary(aft_lognorm)

# Distributional fit
fit.dist = survfit(Surv(time, status)~1, conf.type="none", data)
par(mfrow=c(2,2))
plot(fit.dist, xlab = "time", ylab = "Survival", main="K-M Estimate")

plot(log(fit.dist$time), log(-log(fit.dist$surv)), type="s"
      , xlab = "log(time)", ylab="log(-log(Survival))", main="Weibull")

plot(fit.dist$time, log(fit.dist$surv), type="s"
      , xlab = "time", ylab="-log(Survival)", main='Exponential')

plot(log(fit.dist$time), qnorm(1-fit.dist$surv), type="s"
      , xlab = "log(time)", ylab="log(-log(Survival))", main="Log-normal")

# Prediction
pct = 1:99/100
ptime = predict(aft_exp, newdata = list(age=60, bmi=23, gender=1,
disease=2), type='quantile',
                    p=pct, se=TRUE)
```

```
matplot(cbind(ptime$fit, ptime$fit + 1.96*ptime$se.fit, ptime$fit
- 1.96*ptime$se.fit), 1-pct, xlab=”Time”, ylab=”Survival”,
type='s', lty=c(1,2,2),col=1, main='Prediction␣Patient␣2')
```

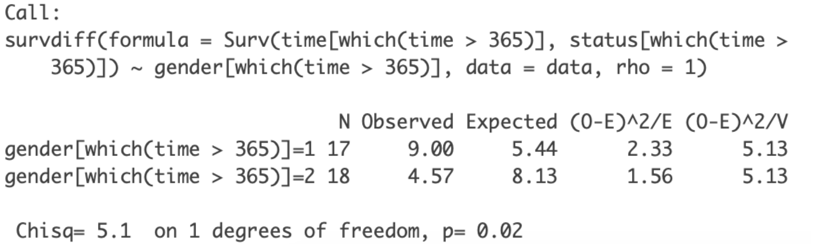
7 Appendix II - Output

7.1 Gender based differences



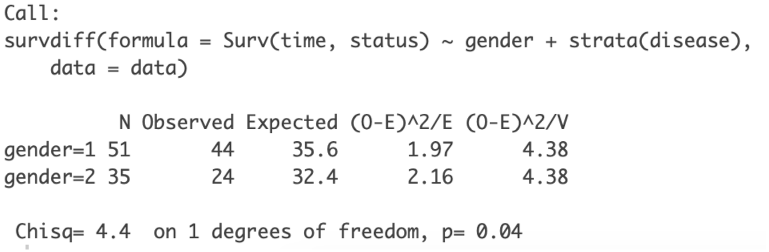
```
Call:
survdiff(formula = Surv(time, status) ~ gender, data = data,
    rho = 1)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gender=1 51     27.5     21.9      1.46      4.58
gender=2 35     13.4     19.1      1.67      4.58

 Chisq= 4.6  on 1 degrees of freedom, p= 0.03
```

7.2 Gender based differences after 1 year of enlistment



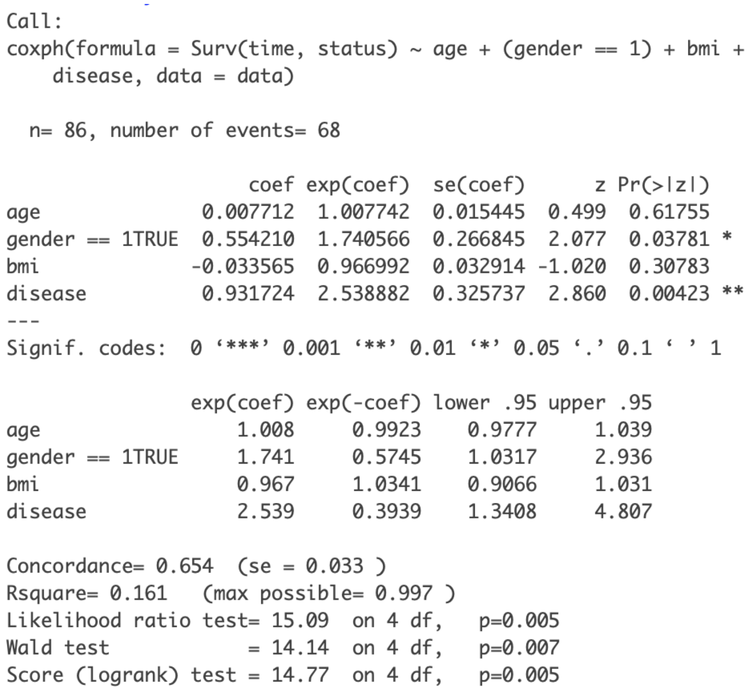
```
Call:
survdiff(formula = Surv(time[which(time > 365)], status[which(time >
    365)]) ~ gender[which(time > 365)], data = data, rho = 1)

                             N Observed Expected (O-E)^2/E (O-E)^2/V
gender[which(time > 365)]=1 17     9.00     5.44      2.33      5.13
gender[which(time > 365)]=2 18     4.57     8.13      1.56      5.13

 Chisq= 5.1  on 1 degrees of freedom, p= 0.02
```

7.3 Gender based differences corrected for disease



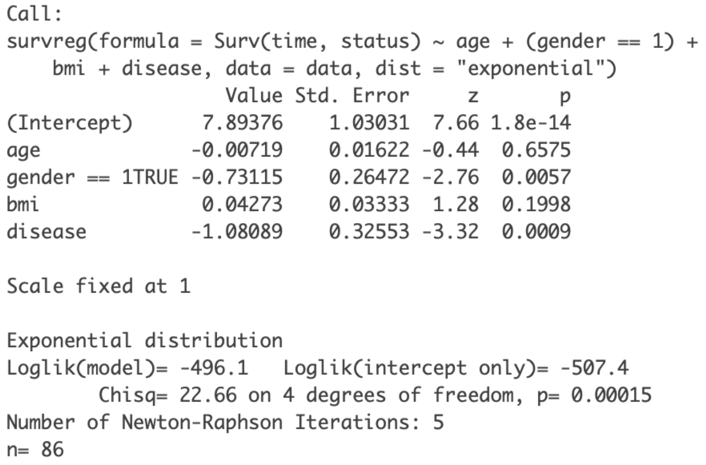
```
Call:
survdiff(formula = Surv(time, status) ~ gender + strata(disease),
    data = data)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gender=1 51       44     35.6      1.97      4.38
gender=2 35       24     32.4      2.16      4.38

 Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

7.4 Cox's Regression Model



```
Call:
coxph(formula = Surv(time, status) ~ age + (gender == 1) + bmi +
    disease, data = data)

  n= 86, number of events= 68

                    coef exp(coef)  se(coef)      z Pr(>|z|)
age             0.007712  1.007742  0.015445  0.499  0.61755
gender == 1TRUE  0.554210  1.740566  0.266845  2.077  0.03781 *
bmi            -0.033565  0.966992  0.032914 -1.020  0.30783
disease         0.931724  2.538882  0.325737  2.860  0.00423 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
age                 1.008     0.9923    0.9777     1.039
gender == 1TRUE     1.741     0.5745    1.0317     2.936
bmi                 0.967     1.0341    0.9066     1.031
disease             2.539     0.3939    1.3408     4.807

Concordance= 0.654  (se = 0.033 )
Rsquare= 0.161   (max possible= 0.997 )
Likelihood ratio test= 15.09  on 4 df,   p=0.005
Wald test            = 14.14  on 4 df,   p=0.007
Score (logrank) test = 14.77  on 4 df,   p=0.005
```

7.5 Exponential Accelarated Failure Time Model



```
Call:
survreg(formula = Surv(time, status) ~ age + (gender == 1) +
    bmi + disease, data = data, dist = "exponential")
                   Value Std. Error     z       p
(Intercept)      7.89376    1.03031  7.66 1.8e-14
age             -0.00719    0.01622 -0.44  0.6575
gender == 1TRUE -0.73115    0.26472 -2.76  0.0057
bmi              0.04273    0.03333  1.28  0.1998
disease         -1.08089    0.32553 -3.32  0.0009

Scale fixed at 1

Exponential distribution
Loglik(model)= -496.1   Loglik(intercept only)= -507.4
        Chisq= 22.66 on 4 degrees of freedom, p= 0.00015
Number of Newton-Raphson Iterations: 5
n= 86
```