**Warsaw University of Life Sciences**

**( WULS – SGGW )**

**Faculty Applied Informatics and Mathematics Event History**

**Final Project Report**

**Lung Cancer**

**Prof. Dr. Joanna Landmesser**

**Cumali Bereket**

**Album number 211060**

**Warsaw, year 2023**

# **Introduction:**

Lung cancer is a major global health issue and remains one of the leading causes of cancer-related deaths worldwide. Survival analysis plays a crucial role in understanding the factors that influence the duration of survival for individuals diagnosed with lung cancer. By analyzing the time until an event, such as death or disease progression, survival analysis provides valuable insights into prognosis, treatment effectiveness, and patient outcomes.

In this project, we aim to conduct a comprehensive survival analysis using a lung cancer dataset. The dataset contains information on patients diagnosed with lung cancer, including demographic characteristics, clinical features, and treatment modalities. By applying various survival analysis techniques, we seek to identify significant prognostic factors and develop predictive models to estimate patient survival times.

Survival analysis allows us to account for censoring, a common feature in longitudinal studies where the event of interest (e.g., death) has not yet occurred for some individuals. Censoring can occur due to a variety of reasons, such as loss to follow-up or patients still being alive at the end of the study period. By considering censored observations, survival analysis provides a more accurate estimation of survival probabilities and enables the assessment of time-dependent covariates.

The primary objectives of this project are twofold: first, to investigate the factors associated with lung cancer survival and their impact on patient outcomes, and second, to develop and compare survival models to predict individual survival times. By identifying prognostic factors and constructing reliable predictive models, clinicians and researchers can better understand the disease progression, tailor treatment strategies, and improve patient care.

The results of this study have the potential to enhance the understanding of lung cancer prognosis and assist in clinical decision-making. By gaining insights into the factors that influence survival and developing accurate prediction models, healthcare professionals can optimize treatment plans and improve patient outcomes. Ultimately, this project contributes to the ongoing efforts to combat lung cancer and improve the quality of life for affected individuals.

# **Introduction of Survival Analysis**

Generally, **survival analysis** is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.

In the medical world, we typically think of survival analysis literally – tracking time until death. But, it’s more general than that – survival analysis models time until an event occurs (any event). This might be death of a biological organism. But it could also be the time until a hardware failure in a mechanical system, time until recovery, time someone remains unemployed after losing a job, time until a ripe tomato is eaten by a grazing deer, time until someone falls asleep in a workshop, etc. Survival analysis also goes by reliability theory in engineering, duration analysis in economics, and event history analysis in sociology.

Type of events: death, disease, relapse, recover. The goal of survival analysis is to:

1: To estimate and interpret survivor and/or hazard functions from survival data. 2: To compare survivor and/or hazard functions.

3: To assess the relationship of explanatory variables to survival time.

# **Description of dataset:**

**NCCTG Lung Cancer Data**

## **Description**

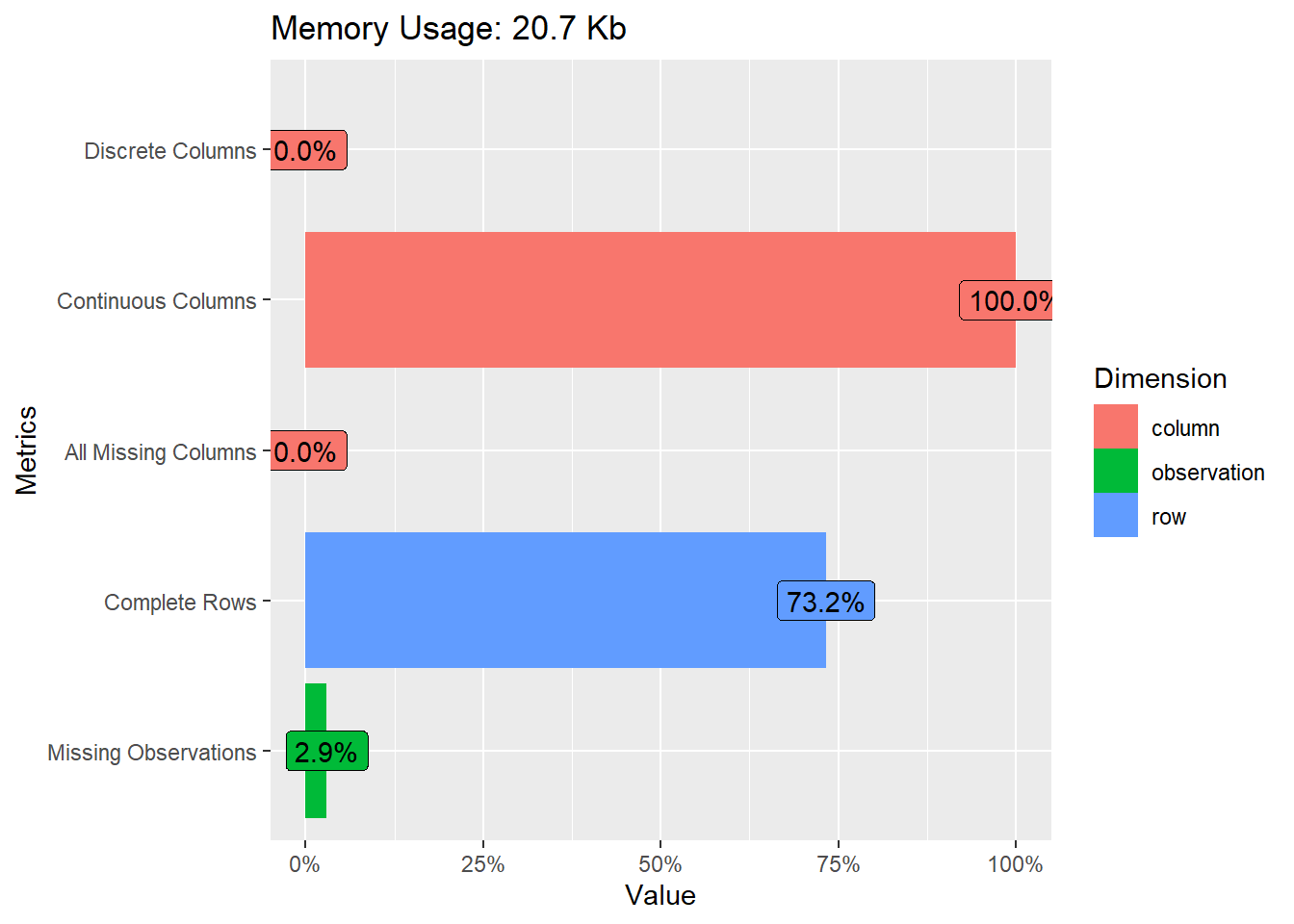
Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

We use the **lung dataset** available from the survival package survival. The data contain subjects with advanced lung cancer from the North Central Cancer Treatment Group. It includes the 10 following variables:

* inst:  Institution code
* time:  Survival time in days
* status:  censoring status 1=censored, 2=dead
* age:  Age in years
* sex:  Male=1 Female=2
* ph.ecog:  ECOG performance score as rated by the physician. 0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed <50% of the day, 3= in bed > 50% of the day but not bedbound, 4 = bedbound
* ph.karno:  Karnofsky performance score (bad=0-good=100) rated by physician
* pat.karno:  Karnofsky performance score (0 = bad, 100 = good) as rated by patient
* meal.cal:  Calories consumed at meals
* wt.loss:  Weight loss in last six months

|  |  |
| --- | --- |
|  |  |

plot\_intro(lung)



We have 2280 observations. All the variables are treated as continuous. Therefore, we transform the variable sex into a factor. We can also see that only 167/228= 73% of the rows are complete.

**Kaplan-Meier model**

The Kaplan-Meier method is the most common way to estimate survival times and probabilities. It is a non-parametric approach that results in a step function, where there is a step down each time an event occurs.

The remaining survival probabilities are computed such that we count the number of subjects surviving past the specified time being considered and divide by the number of subjects at the start of follow-up.

S^(tf)=S^(tf−1)P^r(T≥t(f)|T≥t(f))=∏i=1fP^r(T≥t(i)|T≥t(i))

This formula gives the probability of surviving past the previous failure time tf−1 multiplied by the conditional probability of surviving past time tf given survival to at least time tf. Each term in the product is the probability of exceeding a specific ordered failure time tf given that a subject survives up to that failure time.

## **Compute the Kaplan-Meier model**

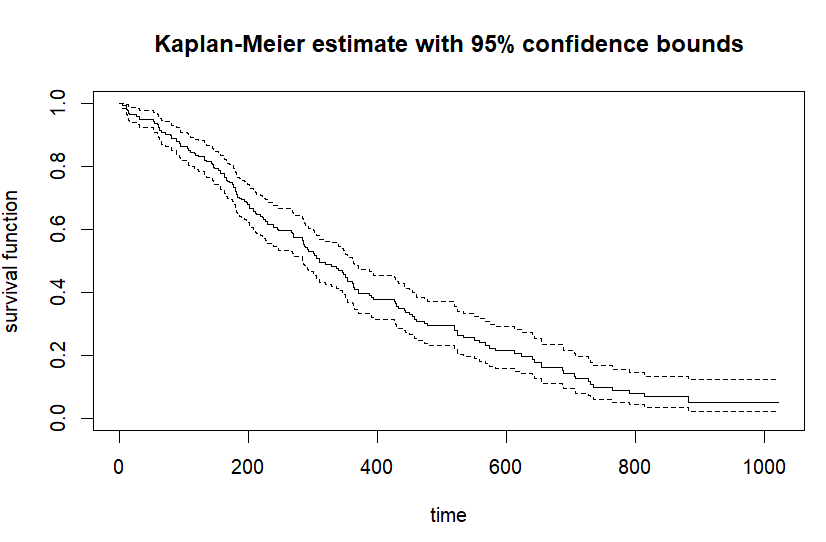
With the survfit() function, we create a simple survival curve that doesn’t consider any different groupings, so we’ll specify just an intercept (e.g., ~1) in the formula that survfit expects.

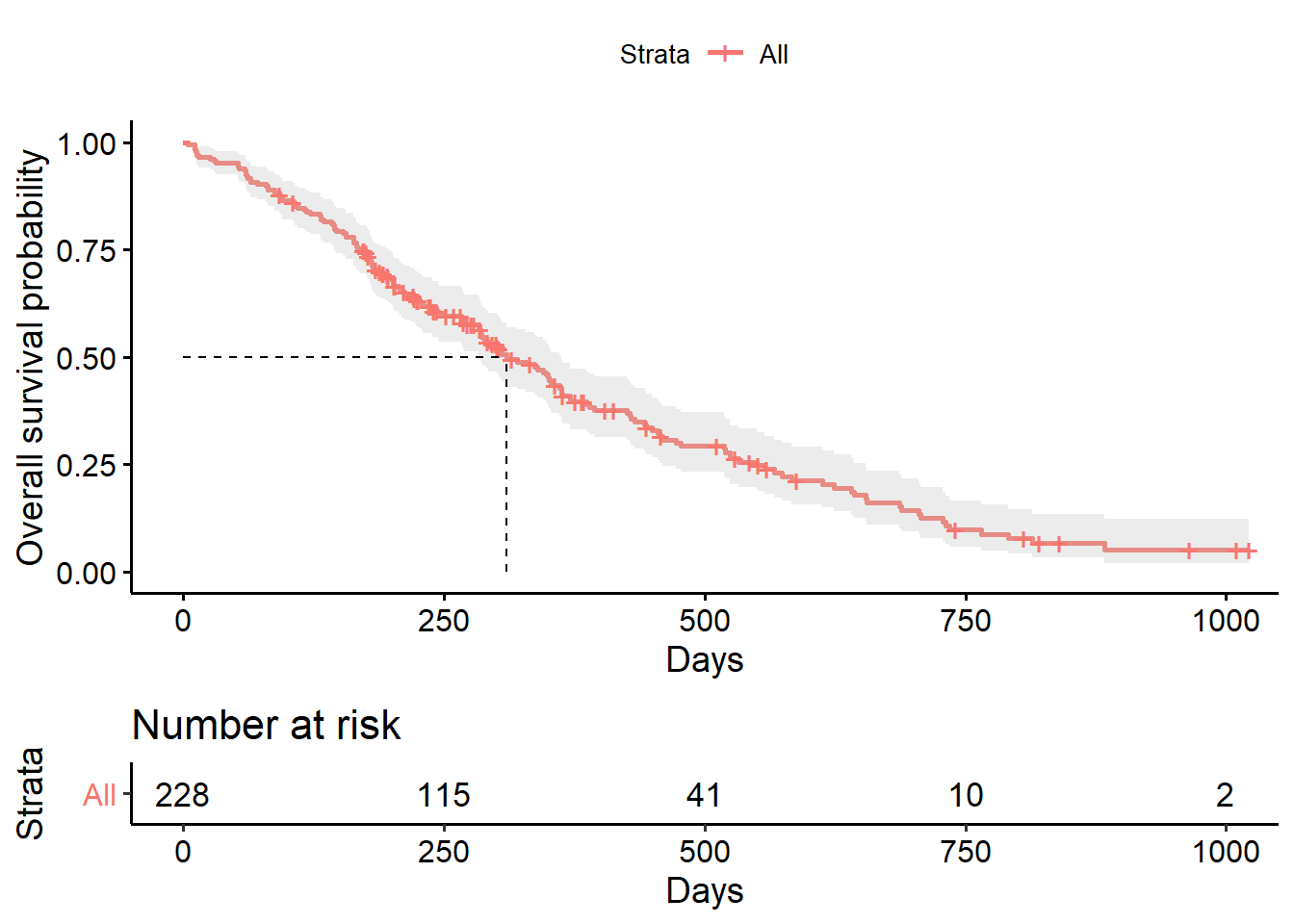
By default, the function print() shows a short summary of the survival curves. It prints the number of observations, number of events, the median survival and the confidence limits for the median. We can note that the median survival time is 310 days.

The function survfit() returns a list of variables, including the following components:

* n: total number of subjects in each curve.
* time: the time points on the curve.
* n.risk: the number of subjects at risk at time t
* n.event: the number of events that occurred at time t.
* n.censor: the number of censored subjects, who exit the risk set, without an event, at time t.
* lower,upper: lower and upper confidence limits for the curve, respectively.

We use the function ggsurvplot() [in Survminer R package] to produce the survival curves for the two groups of subjects.



****

## **Using Variables: Kaplan Meier Model With Sex**

#### We model something besides just an intercept. Let’s fit survival curves separately by sex.

# **Comparing Survival Curves**

The **log-rank** test can be used to evaluate whether or not KM curves for two or more groups are statistically equivalent. The null hypothesis is that there is no difference in survival between the groups.

The log rank test is a **non-parametric test**, which makes no assumptions about the survival distributions. It is approximately distributed as a chi-square test statistic.

The function survdiff() [in survival package] can be used to compute log-rank test comparing males and females survival curves.

survdiff(Surv(time, had\_event) ~ sex, lung)

## Call:

## survdiff(formula = Surv(time, had\_event) ~ sex, data = lung)

##

## N Observed Expected (O-E)^2/E (O-E)^2/V

## sex=M 138 112 91.6 4.55 10.3

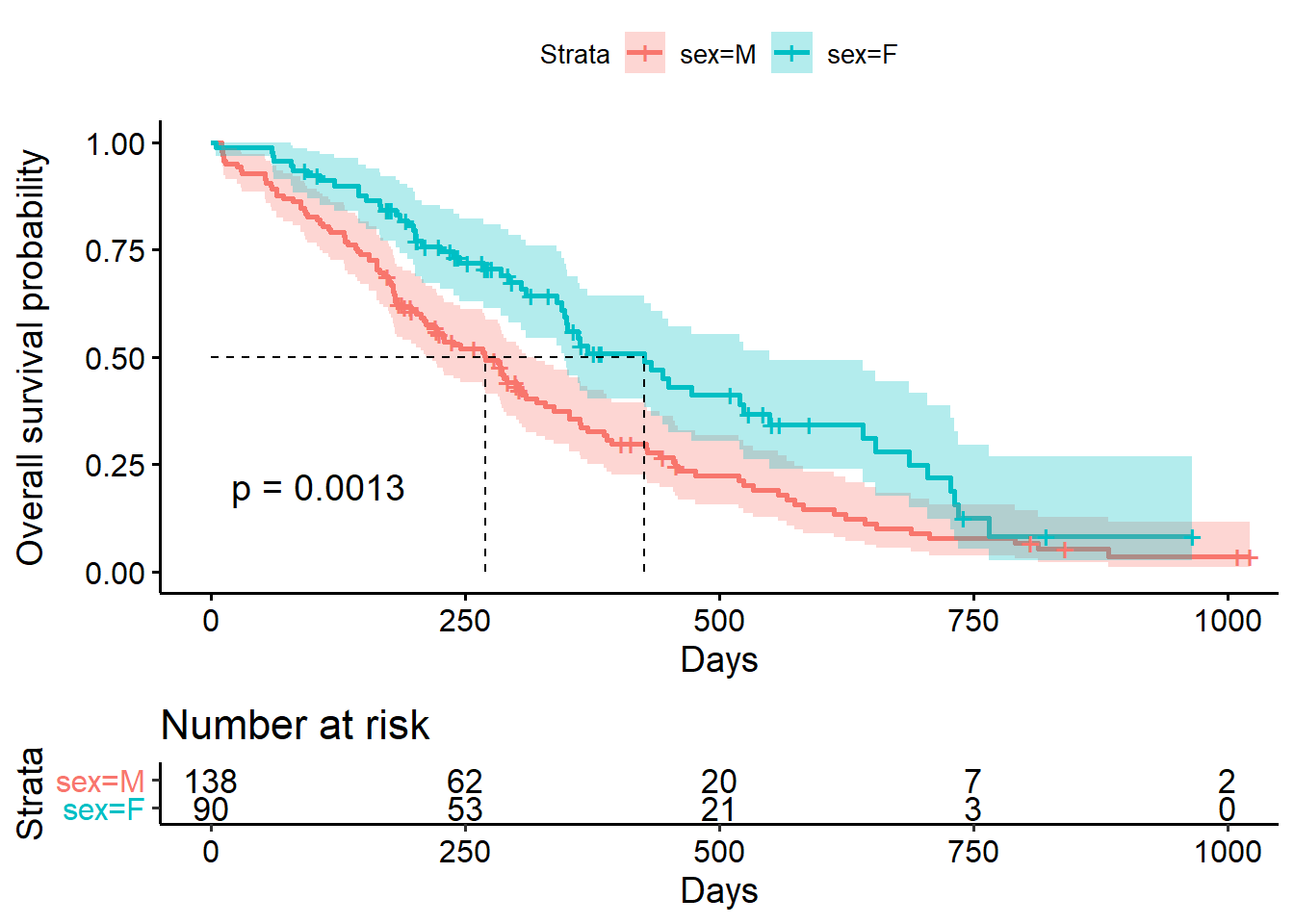
## sex=F 90 53 73.4 5.68 10.3

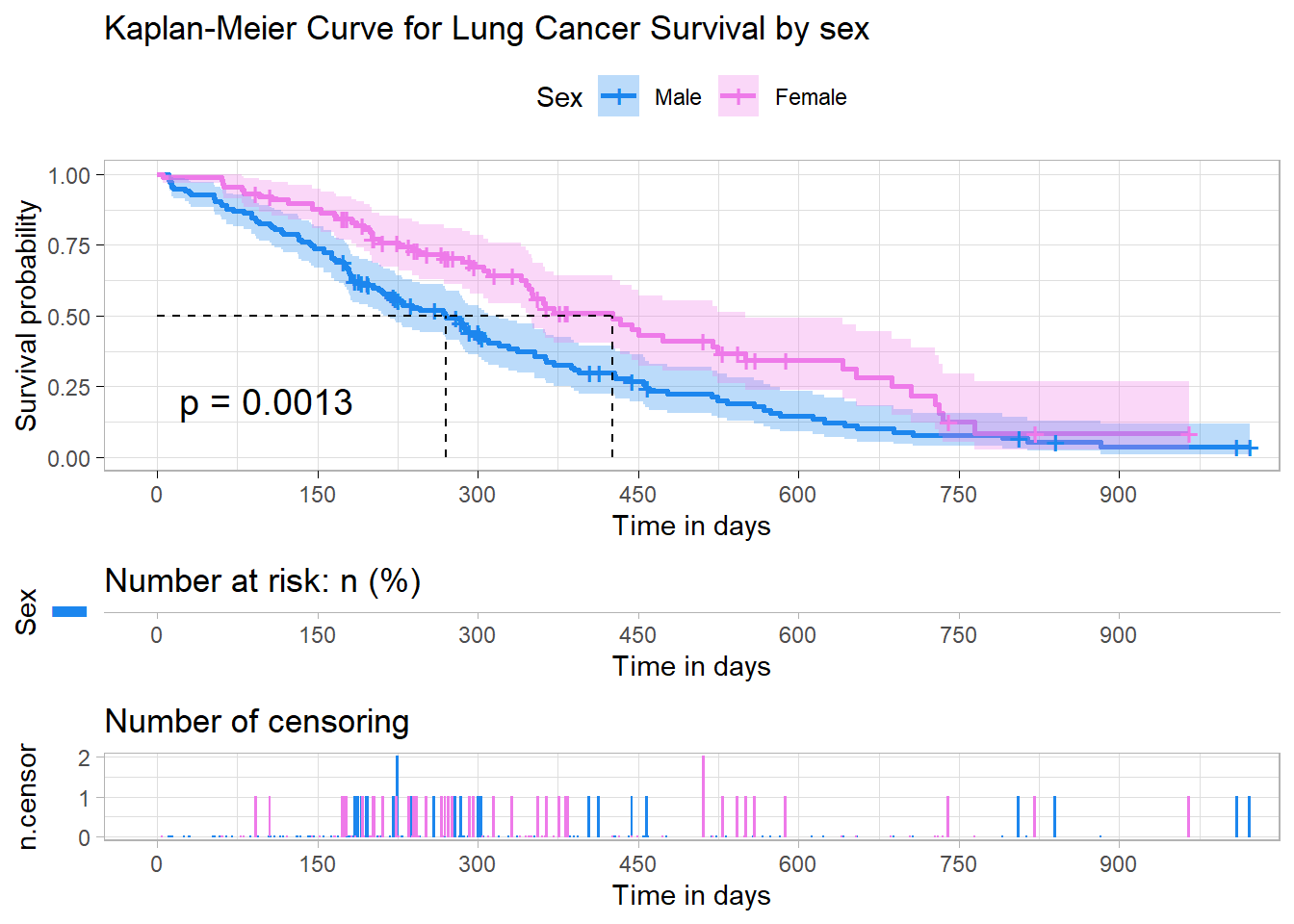
##

## Chisq= 10.3 on 1 degrees of freedom, p= 0.001

The log rank test for difference in survival gives a p-value of p = 0.001, indicating that the sex groups differ significantly in survival.

There are several alternatives to the log rank test designed to test the hypothesis that two or more survival curves are equivalent called the Wilcoxon, the Tarone-Ware, the Peto, and the Flemington-Harrington test. These test statistics are variations of the log rank test statistic and are derived by applying different weights at the f-th failure time (as shown on the left for two groups).

****

****

# **Comparing Survival Curves**

The **log-rank** test can be used to evaluate whether or not KM curves for two or more groups are statistically equivalent. The null hypothesis is that there is no difference in survival between the groups.

The log rank test is a **non-parametric test**, which makes no assumptions about the survival distributions. It is approximately distributed as a chi-square test statistic.

X2= (−10.26) ^2 / 19.26+(10.26) ^2 / 10.74 = 15.276

The calculation of the approximate formula is shown here for the remission data. The expected values are 19.26 and 10.74 for groups 1 and 2, respectively. The chi-square value obtained is 15.276, which is slightly smaller than the log–rank statistic of 16.793.

Although the same tabular layout can be used to carry out the calculations when there are more than two groups, the test statistic is more complicated mathematically, involving both ariances and covariances of summed observed minus expected scores for each group.

he function survdiff() [in survival package] can be used to compute log-rank test comparing males and females survival curves.

survdiff(Surv(time, had\_event) ~ sex, lung)

## Call:

## survdiff(formula = Surv(time, had\_event) ~ sex, data = lung)

##

## N Observed Expected (O-E)^2/E (O-E)^2/V

## sex=M 138 112 91.6 4.55 10.3

## sex=F 90 53 73.4 5.68 10.3

##

## Chisq= 10.3 on 1 degrees of freedom, p= 0.001

The log rank test for difference in survival gives a p-value of p = 0.001, indicating that the sex groups differ significantly in survival.

There are several alternatives to the log rank test designed to test the hypothesis that two or more survival curves are equivalent called the Wilcoxon, the Tarone-Ware, the Peto, and the Flemington-Harrington test. These test statistics are variations of the log rank test statistic and are derived by applying different weights at the f-th failure time (as shown on the left for two groups).

## **Transforming The Survival Curve**

There is another major argument of the ggsurvplot() function, “fun”, which proposes alternative transformations:

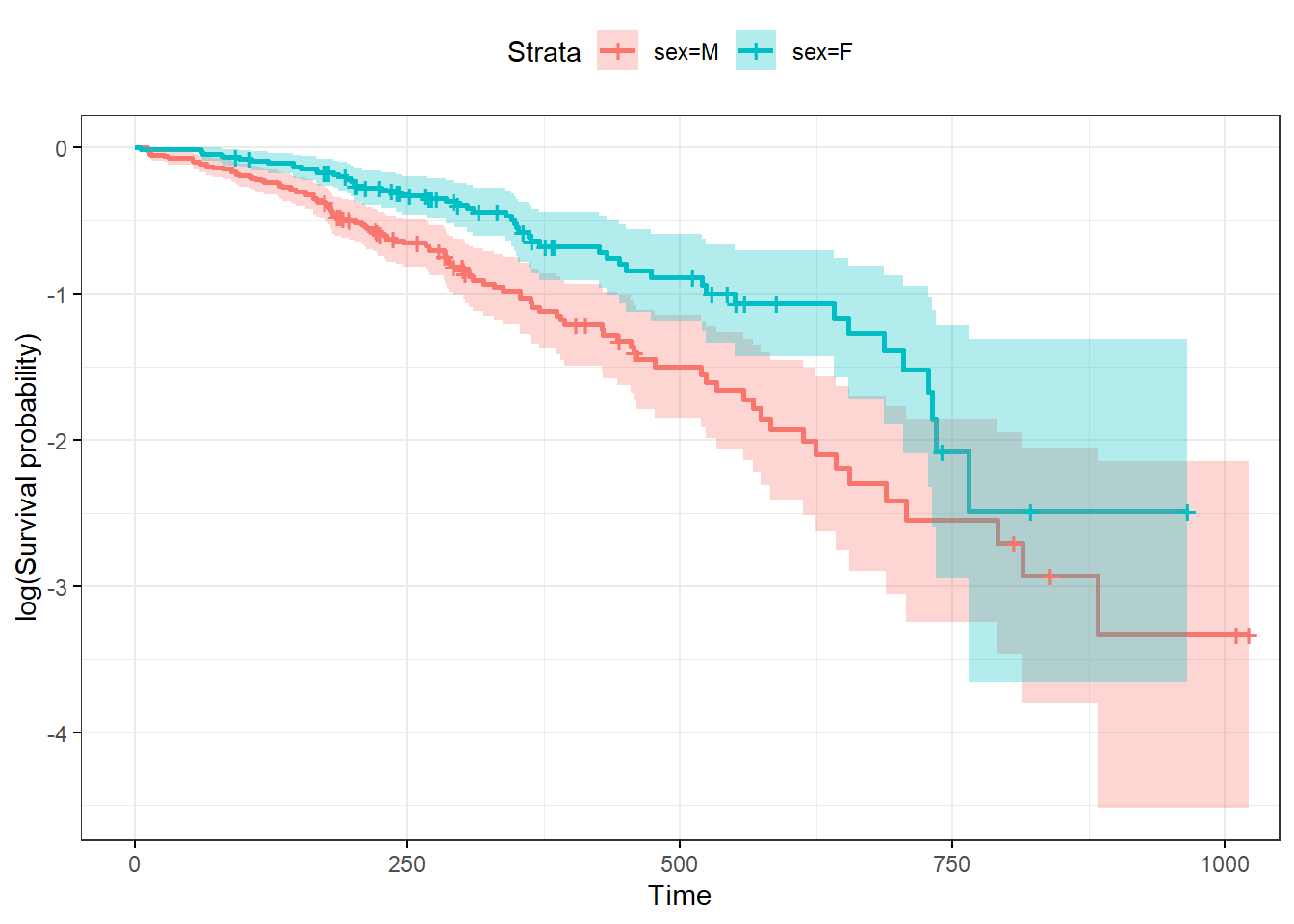
* “log”: **log transformation** of the survivor function

ggsurvplot(survfit(Surv(time, event=had\_event) ~ sex, data=lung),

conf.int = TRUE,

ggtheme = theme\_bw(),

fun = "log")



**Cox PH model**

Kaplan-Meier curves are good for visualizing differences in survival between two categorical groups, and the log-rank test you get when you ask for pval = TRUE is useful for asking if there are differences in survival between different groups. But this doesn’t generalize well for assessing the effect of quantitative variables.

Another method, called **Cox PH** regression, can assess the effect of both categorical and continuous variables, and can model the effect of multiple variables at once.

Cox PH General model

The Cox model is expressed by the hazard function denoted by h(t). It is a semi-parametric model that can be used to fit univariable and multivariable regression models that have survival outcomes. The hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

h(t,Xi)=h0(t)e∑pj=1βjXi,j=h0(t)exp(β1Xi,1+...+ betapXi,p)

where:

h(t)

is the hazard, the instantaneous rate at which events occur.

h0(t)

is called the baseline hazards (when all X’s are equal to 0), depends on t

X=(X1,X2,...,Xp)

explanatory/predictor variables

e∑pi=1βiXi

, depends only on X’s, called .

Because the baseline hazard h0(t)

is an unspecified function, the Cox model us a semiparametric model.

Advantages of the model: “robust” model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables Xi, with the baseline hazard, h0(t), being an ‘intercept’ term that varies with time.

log(h(t,Xi))=log(h0(t))+∑j=1pβjXi,j

We can compute the hazard ratio, which is the ratio of hazards between two groups at any particular point in time: “hazard for one individual divided by the hazard for a different individual”.

HR^=h^(t,X∗)h^(t,X)=e∑pi=1βi(X∗i−Xi)

with:

X∗: set of predictors for one individual

X: set of predictors for the other individual

This model shows that the hazard ratio is equal to e∑pi=1βi(X∗i−Xi)

, and remains constant over time t (hence the name proportional hazards regression). In this sense, we do not need the baseline hazard because we can interpret coefficients as hazard ratios.

A hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

HR=1 : No effect

HR<1: Reduction in the hazard

HR>1: Increase in Hazard

As a note, in cancer studies, a covariate with hazard ratio :

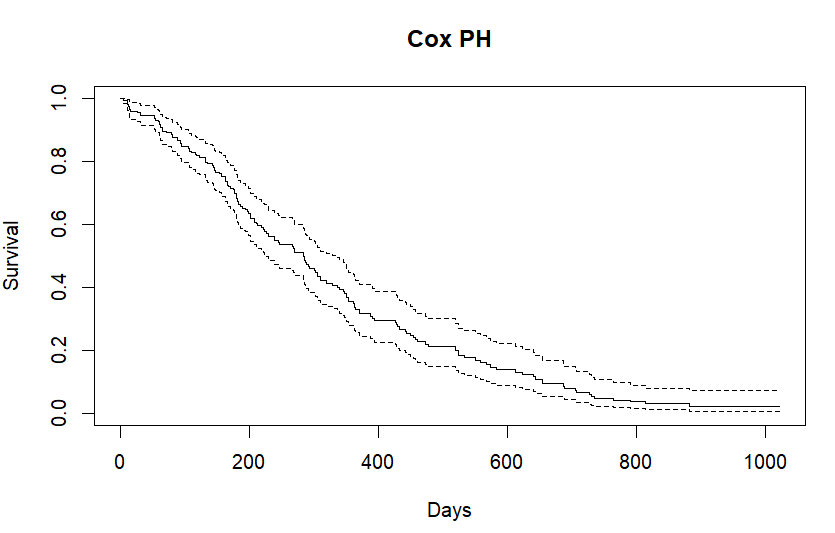
greater than 1 (i.e.: b>0) is called bad prognostic factor.

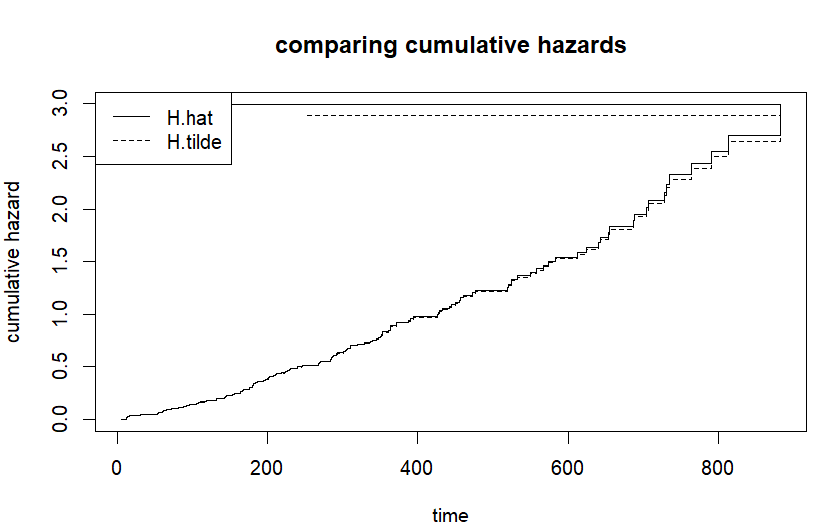
smaller than 1 (i.e.: b<0) is called good prognostic factor.

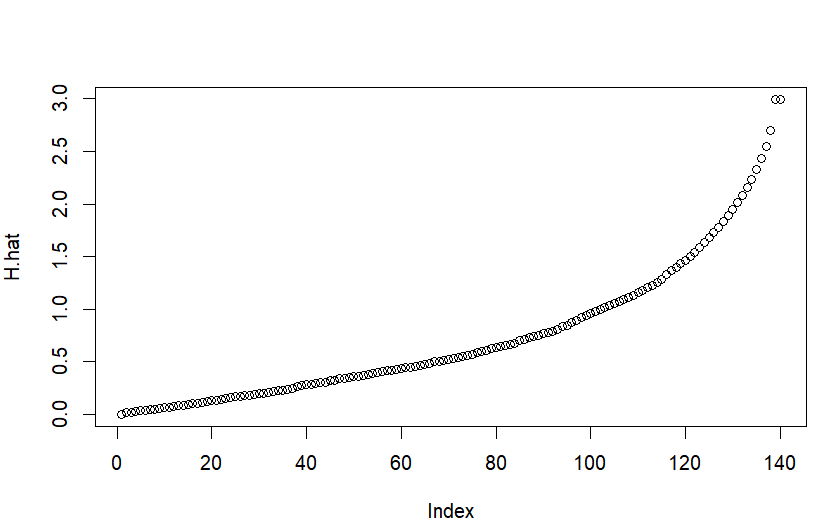
As a consequence, a major assumption of this model is that the HR is constant over time because it is independent of time. Or equivalently that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

It is possible, nevertheless, to consider X’s which do involve t. Such X’s are called time-dependent variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the extended Cox model.

**Compute the Cox Model**

The coxph() function uses the same syntax as lm(), glm(), etc. The response variable you create with Surv() goes on the left hand side of the formula, specified with a ~. Explanatory variables go on the right side.****





**COX PH Model With Sex Variable Only**

cox\_sex = coxph(Surv(time, had\_event) ~ sex, data=lung)

print(cox\_sex)

## Call:

## coxph(formula = Surv(time, had\_event) ~ sex, data = lung)

##

## coef e xp(coef) se(coef) z p

## sexF -0.5310 0.5880 0.1672 -3.176 0.00149

##

## Likelihood ratio test=10.63 on 1 df, p=0.001111

## n= 228, number of events= 165

The effect of sex is significantly related to survival (p-value = 0.00149), with better survival in females in comparison to males (hazard ratio of dying = 0.588).

The model is statistically significant. That 0.00111 p-value of the model is really close to the p = 0.00131 p-value we saw on the Kaplan-Meier nodel as well as the likelihood ratio test = 10.63 is close to the log-rank chi-square (10.3) in the Kaplan-Meier model.

eβ1

= e−0.531

= 0.5880 is the hazard ratio - the multiplicative effect of that variable on the hazard rate (for each unit increase in that variable). Females have 0.588 (~ 60%) times the hazard of dying in comparison to males. So, for a categorical variable like sex, going from male (baseline) to female results in approximately ~40% reduction in hazard.

We could also flip the sign on the coef column, and take e0.531

= 1/0.588 = 1.7, which we can interpret as being male resulting in a 1.7-fold increase in hazard, or that males die as approximately 1.7 x the rate per unit time as females (females die at 0.588x the rate per unit time as males).

We can directly calculate the log-rank test p-value using survdiff(). Cox regression and the logrank test from survdiff will to give you similar results most of the time. The log-rank test is asking if survival curves differ significantly between two groups. Cox regression is asking which of many categorical or continuous variables significantly affect survival.

Call:

The log-rank test tests the following hypothesis:

H0: There is no difference in the survival function when comparing males to females after adjusting for age.

Ha: There is a difference in the survival function when comparing males to females after adjusting for age.

survdiff(Surv(time, status)~sex, data=lung)

## Call:

## survdiff(formula = Surv(time, status) ~ sex, data = lung)

##

## N Observed Expected (O-E)^2/E (O-E)^2/V

## sex=M 138 112 91.6 4.55 10.3

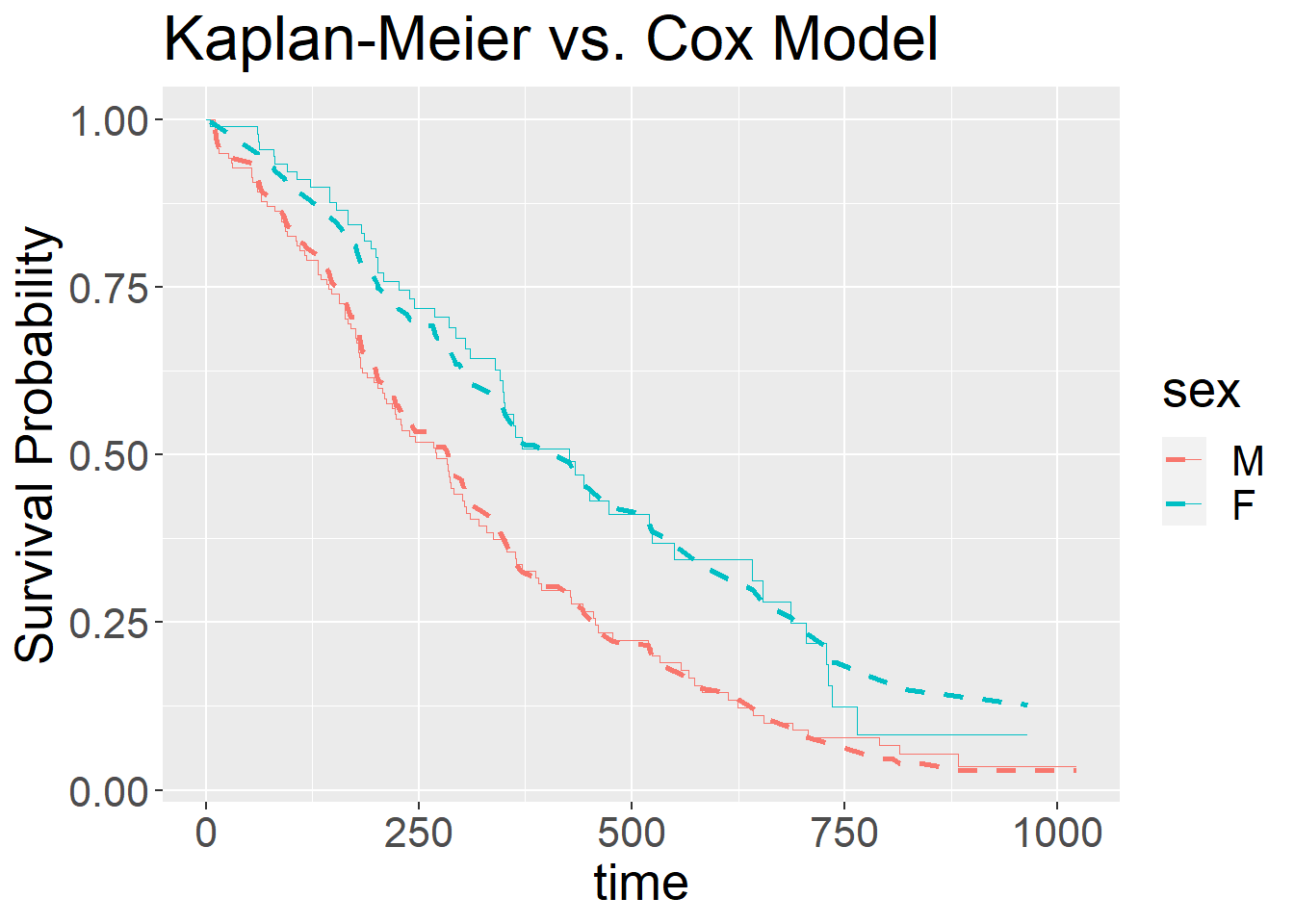
## sex=F 90 53 73.4 5.68 10.3

##

## Chisq= 10.3 on 1 degrees of freedom, p= 0.001

# **Comparison of the KM Model and COX PH Model With the Variable Sex**

We can then compare the survival curves from KM (km\_sex) and Cox PH model (cox\_sex).

****

## **COX PH Model With Age Variable Only**

cox\_age = coxph(Surv(time, had\_event) ~ age, data=lung)

print(cox\_age)

## Call:

## coxph(formula = Surv(time, had\_event) ~ age, data = lung)

##

## coef exp(coef) se(coef) z p

## age 0.018720 1.018897 0.009199 2.035 0.0419

##

## Likelihood ratio test=4.24 on 1 df, p=0.03946

## n= 228, number of events= 165

The effect of age is significantly related to survival (p-value = 0.03946). eβ1

= e0.018720

= 1.018897 is the hazard ratio - the multiplicative effect of that variable on the hazard rate (for each unit increase of age). We have a better survival in younger people (hazard ratio of dying = 1.018897).

**AFT MODELS**

**Parametric Proportional Hazard (PH) and AFT (Acceleration Failure Time)**

Call:

coxph(formula = Surv(time, status) ~ sex + age + ph.ecog + ph.karno +

pat.karno + meal.cal + wt.loss, data = lung)

n= 228, number of events= 165

exp(coef) exp(-coef) lower .95 upper .95

sex 0.5572 1.7947 0.3991 0.7778

age 1.0125 0.9877 0.9939 1.0314

ph.ecog 1.7398 0.5748 1.2164 2.4885

ph.karno 1.0127 0.9874 0.9943 1.0315

pat.karno 0.9875 1.0127 0.9741 1.0010

meal.cal 1.0000 1.0000 0.9996 1.0005

wt.loss 0.9894 1.0107 0.9767 1.0022

Concordance= 0.648 (se = 0.024 )

Likelihood ratio test= 36.16 on 7 df, p=7e-06

Wald test = 35.44 on 7 df, p=9e-06

Score (logrank) test = 36.4 on 7 df, p=6e-06

**Peto&Peto modification of the Gehan-Wilcoxon test**

Call:

survdiff(formula = Surv(time, status) ~ sex, rho = 0)

N Observed Expected (O-E)^2/E (O-E)^2/V

sex=0 138 112 91.6 4.55 10.3

sex=1 90 53 73.4 5.68 10.3

Chisq= 10.3 on 1 degrees of freedom, p= 0.001

> survdiff(Surv(time, status) ~ sex, rho=1) #Peto&Peto modification of the Gehan-Wilcoxon test

**Log-rank or Mantel-Haenszel test**

Call:

survdiff(formula = Surv(time, status) ~ sex, rho = 1)

N Observed Expected (O-E)^2/E (O-E)^2/V

sex=0 138 70.4 55.6 3.95 12.7

sex=1 90 28.7 43.5 5.04 12.7

Chisq= 12.7 on 1 degrees of freedom, p= 4e-04

> survdiff(Surv(time, status) ~ ph.ecog, rho=0) #log-rank or Mantel-Haenszel test

Call:

survdiff(formula = Surv(time, status) ~ ph.ecog, rho = 0)

**1-Weibull**

> #AFT models

> srFitWeib <- survreg(Surv(time, status) ~ sex+age+ph.ecog+ph.karno+pat.karno+meal.cal+wt.loss, dist="weibull")

> summary(srFitWeib)

Call:

survreg(formula = Surv(time, status) ~ sex + age + ph.ecog +

ph.karno + pat.karno + meal.cal + wt.loss, dist = "weibull")

Value Std. Error z p

(Intercept) 6.82e+00 8.80e-01 7.74 9.7e-15

sex 4.21e-01 1.23e-01 3.42 0.00063

age -8.41e-03 6.75e-03 -1.25 0.21262

ph.ecog -4.09e-01 1.28e-01 -3.19 0.00143

ph.karno -1.02e-02 6.56e-03 -1.55 0.12069

pat.karno 9.23e-03 4.93e-03 1.87 0.06118

meal.cal -1.32e-05 1.70e-04 -0.08 0.93772

wt.loss 7.58e-03 4.68e-03 1.62 0.10547

Log(scale) -3.32e-01 6.17e-02 -5.38 7.6e-08

Scale= 0.718

Weibull distribution

Loglik(model)= -1135.7 Loglik(intercept only)= -1153.9

Chisq= 36.2 on 7 degrees of freedom, p= 6.6e-06

Number of Newton-Raphson Iterations: 6

n= 228

**2-Exponential**

Call:

survreg(formula = Surv(time, status) ~ sex + age + ph.ecog +

ph.karno + pat.karno + meal.cal + wt.loss, dist = "exponential")

Value Std. Error z p

(Intercept) 6.77e+00 1.23e+00 5.49 4.1e-08

sex 5.36e-01 1.70e-01 3.16 0.0016

age -1.06e-02 9.30e-03 -1.14 0.2557

ph.ecog -4.46e-01 1.79e-01 -2.49 0.0128

ph.karno -8.82e-03 9.46e-03 -0.93 0.3511

pat.karno 9.77e-03 6.65e-03 1.47 0.1416

meal.cal 2.09e-05 2.32e-04 0.09 0.9281

wt.loss 8.07e-03 6.41e-03 1.26 0.2078

Scale fixed at 1

Exponential distribution

Loglik(model)= -1148 Loglik(intercept only)= -1162.3

Chisq= 28.62 on 7 degrees of freedom, p= 0.00017

Number of Newton-Raphson Iterations: 4

n= 228

**3-Log-Normal**

Call:

survreg(formula = Surv(time, status) ~ sex + age + ph.ecog +

ph.karno + pat.karno + meal.cal + wt.loss, dist = "lognormal")

Value Std. Error z p

(Intercept) 5.734731 1.169921 4.90 9.5e-07

sex 0.575232 0.153831 3.74 0.00018

age -0.016264 0.008418 -1.93 0.05336

ph.ecog -0.266880 0.164045 -1.63 0.10377

ph.karno -0.001424 0.009609 -0.15 0.88223

pat.karno 0.009603 0.005896 1.63 0.10340

meal.cal 0.000295 0.000211 1.40 0.16200

wt.loss 0.005958 0.005757 1.04 0.30065

Log(scale) 0.015042 0.055793 0.27 0.78747

Scale= 1.02

Log Normal distribution

Loglik(model)= -1150.4 Loglik(intercept only)= -1169.3

Chisq= 37.77 on 7 degrees of freedom, p= 3.4e-06

Number of Newton-Raphson Iterations: 4

n= 228

**4-Log-Logistic**

Call:

survreg(formula = Surv(time, status) ~ sex + age + ph.ecog +

ph.karno + pat.karno + meal.cal + wt.loss, dist = "loglogistic")

Value Std. Error z p

(Intercept) 4.703009 1.083119 4.34 1.4e-05

sex 0.523936 0.135941 3.85 0.00012

age -0.005886 0.007516 -0.78 0.43351

ph.ecog -0.233764 0.155006 -1.51 0.13153

ph.karno 0.004581 0.009290 0.49 0.62191

pat.karno 0.009675 0.005207 1.86 0.06315

meal.cal 0.000201 0.000191 1.05 0.29244

wt.loss 0.005728 0.005024 1.14 0.25427

Log(scale) -0.638174 0.065811 -9.70 < 2e-16

Scale= 0.528

Log logistic distribution

Loglik(model)= -1141 Loglik(intercept only)= -1160.9

Chisq= 39.78 on 7 degrees of freedom, p= 1.4e-06

Number of Newton-Raphson Iterations: 4

n= 228

**Interpretation:**

I prefer to choose Weibull Model according to resulst of Loglikelihood values after survival analysis. So Weibull model is the best model.

For Weibull Model = -1135.7

For Exponential Model = -1148

For Log Normal Model = -1150.4

For Log Logistic Model = -1141