Εργασία 1:Ανάλυση Επιβίωσης και Στατιστική Θεωρία Αξιοπιστίας

Σουσάνης Ανδρέας

ΠΜΣ: Υπολογιστική και Στατιστική Αναλυτική στην Επιστήμη των Δεδομένων.

Μάθημα:Ανάλυση Επιβίωσης και Στατιστική Θεωρία Αξιοπιστίας

Ακαδημαϊκό Έτος: 2023-24

Contents

1 Περιγραφή Δεδομένων

Στην παρούσα εργασία θα μελετήσουμε το σύνολο δεδομένων Cancer από το πακέτο Survival της R. Το dataset "cancer" περιλαμβάνει πληροφορίες για ασθενείς με καρκίνο του πνεύμονα. Καταγράφονται ηλικία, φύλο, αξιολόγηση από τον ιατρό και τον ασθενή για την απόδοση και άλλοι παράγοντες όπως ο χρόνος επιβίωσης και η κατάσταση (επιζών/μη επιζών). Τα δεδομένα παρέχουν πληροφορίες για τους παράγοντες που επηρεάζουν την επιβίωση και την ποιότητα ζωής των ασθενών. Η αρχική πηγή της έρευνας είναι: Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7,1994. Χρησιμοποιώντας τις παρακάτω εντολές θα πάρουμε σημαντικές πληροφορίες για κάθε μεταβλητή των δεδομένων.

```
library(survival)
library(survminer)
library(ggplot2)
library(gridExtra)
library(fitdistrplus)
data <- cancer
summary(data)
```

Χρόνος(time): Ο χρόνος μέχρι να συμβεί το γεγονός

Κατάσταση(status): Επιβίωση:1 Μη επιβίωση:2

Ηλικία (age): Εύρος: 39-82, Μέσος όρος: 62

Φύλο (sex): Ανδρες: 1, Γυναίκες: 2

ECOG (ph.ecog): Εύρος: 0-4, Μέσος όρος: 0.9515

0: Ασυμπτωματικός

1: Συμπτωματικός αλλά πλήρως ικανός

2: Στο κρεβάτι $<$ από τη μισή μέρα

3: Στο κρεβάτι $>$ από τη μισή μέρα

4: Ακίνητος στο κρεβάτι

ph.karno: Εύρος: 50-100, Μέσος όρος: 81.94

pat.karno: Εύρος: 30-100, Μέσος όρος: περίπου 79.96

Θερμίδες στα γεύματα (meal.cal): Εύρος: 96-2600, Μέσος όρος: 928.8

Απώλεια βάρους (wt.loss): Εύρος: -24 εως 68, Μέσος όρος: 9.832

Στη συνέχεια θα εξετάσουμε το ποσοστό ανδρών και γυναικών στο δείγμα.Στο γράφημα 1.1 φαίνεται ότι το ποσοστό των ανδρών στο δείγμα είναι μεγαλύτερο.

```
percentage_male <- sum(data$sex == "1") / nrow(data) * 100
percentage_female <- sum(data$sex == "2") / nrow(data) * 100
barplot(c(percentage_male, percentage_female),
        names.arg = c("Male", "Female"),
        ylab = "Percentage (%)",
        main = "Percentage of Men and Women in the Sample",
        border = "black",
        width = 0.8,
        ylim = c(0, max(c(percentage_male, percentage_female)) * 1.1),
        col = "red"
        )
```

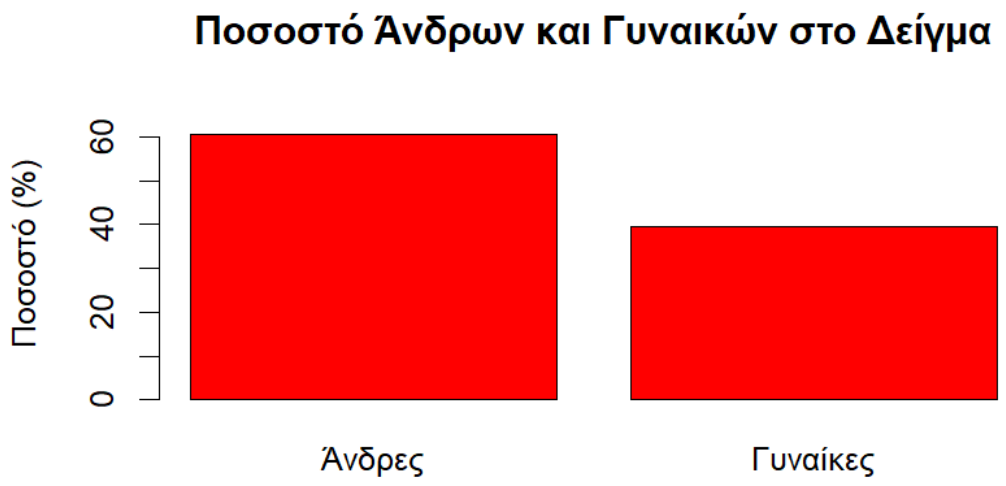
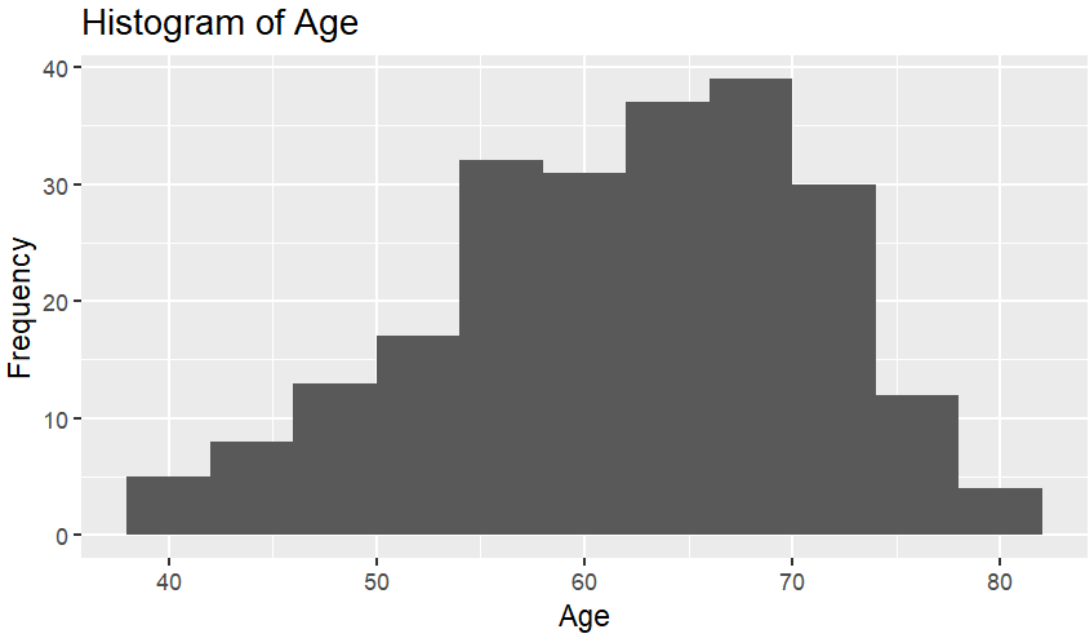


Figure 1: 1.1

```
ggplot(data, aes(data$age)) +
geom_histogram()
```



Στο γράφημα 1.2 παρουσιάζεται πως κατανέμονται οι ηλικίες των ατόμων του δείγματος.Παρατηρούμε ότι τα περισσότερα άτομα του δείγματος είναι από 60 εώς 70 ετών. Παρατηρούμε οτι το σχήμα μας δεν είναι συμμετρικό. Στις ηλικίες 37 εως 53 έχουμε μικρό αριθμό ασθενών, όπως αντίστοιχα και στις ηλικίες 76 και άνω.Αντίθετα παρατηρούμε οτι το μεγαλύτερο πλήθος ασθενών κυμαίνεται ανάμεσα στα έτη 53 και 76 με τους περισσότερους ασθενείς (22) να είναι στην ηλικία των 69 ετών(κορυφή).

```r
count_death_2 <- sum(data$status == 2)
count_death_1 <- sum(data$status == 1)

barplot(c(count_death_1, count_death_2),
        names.arg = c("censored (1)", "Not censored (2)"),
        ylab = "Number of Observations",
        main = "Number of censored and not censored observations",
        col = c("red", "green"),
        border = "blue",
        ylim = c(1, max(count_death_1, count_death_2) * 1.1)
        )
count_death_1
count_death_2
```

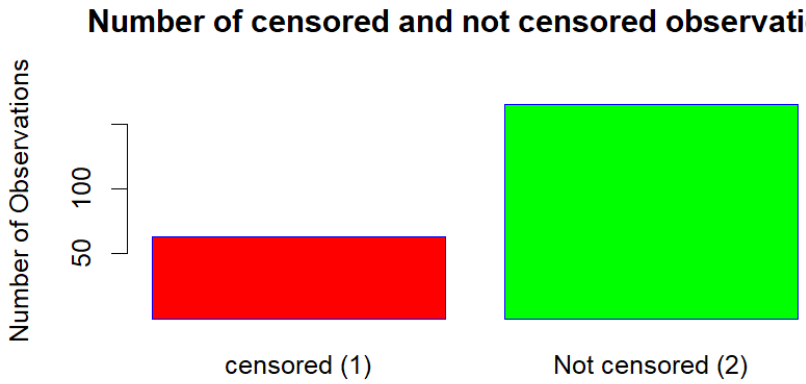


Figure 2: 1.3

Στο γράφημα 1.3 παρουσιάζονται ο αριθμός των ατόμων που επιβίωσαν και αυτών που δεν επιβίωσαν μέχρι το τέλος της έρευνας.Δεξιά αποκομμένες παρατηρήσεις είναι όσοι επιβίωσαν.

- Αριθμός αποκομμένων παρατηρήσεων: 63 (27.63%)
- Αριθμός μη αποκομμένων παρατηρήσεων: 165 (72.37%)

2 Σύγκριση των Χρόνων Επιβίωσης και Εκτίμηση Kaplan-Meier

Εκτίμηση Kaplan-Meier: Η εκτίμηση Kaplan-Meier χρησιμοποιείται για να αξιολογήσει την πιθανότητα επιβίωσης σε διάφορα χρονικά σημεία σε ομάδες ασθενών. Αρχικά, δημιουργείται μια Kaplan-Meier εκτιμήτρια με βάση τα δεδομένα επιβίωσης. Στη συνέχεια, η εκτιμήτρια αυτή αναπαρίσταται γραφικά με τη χρήση του πακέτου survminer στη γλώσσα προγραμματισμού R.

```r
# Kaplan-Meier
km <- survfit(Surv(time, status) ~ 1, data = cancer)
ggsurvplot(km, data = cancer)
summary(km)
km$surv
```

Listing 1: Kaplan-Meier estimation and graphical representation

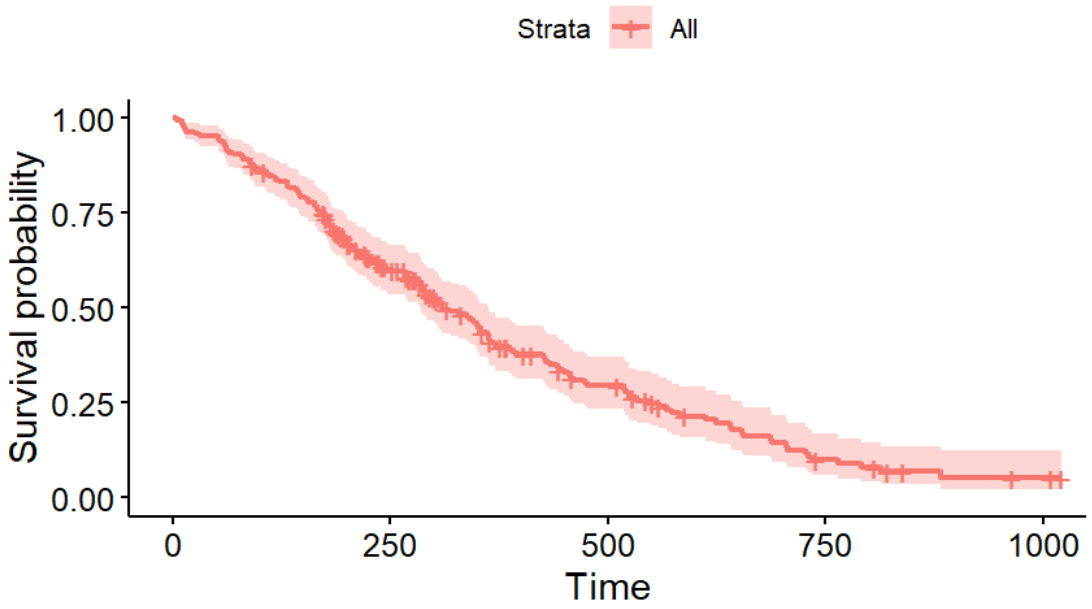


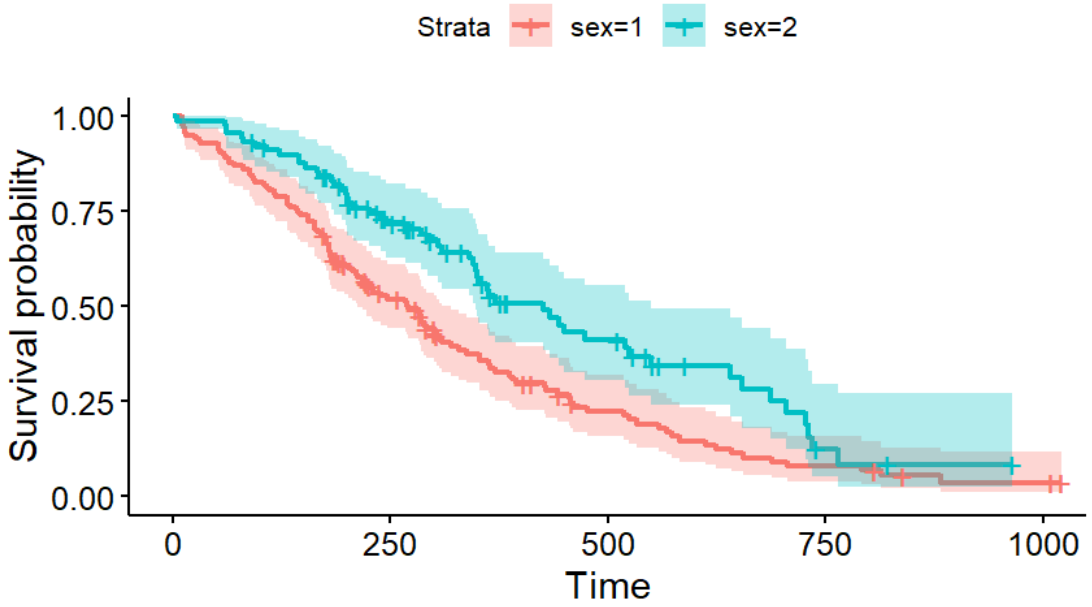
Figure 3: Kaplan-Meier estimation

Στο παραπάνω γράφημα βλέπουμε οτι όσο περνάει ο χρόνος οι πιθανότητες επιβίωσης μειώνονται.Παρατηρούμε οτι ο ρυθμός με τον οποίο μειώνεται η καμπύλη δεν είναι ίδιος σε όλο το σχήμα. Όσο ο χρόνος(μέρες) πηγαίνει απο το 0 μέχρι 350 η μείωση ειναι μεγαλύτερη απο αυτήν μετά. Στο σχήμα φαίνεται επίσης το διάστημα εμπιστοσύνης 95% για κάθε χρονική στιγμή ti(κόκκινη περιοχή δίπλα απο την καμπύλη).

2.1 Εκτιμήτρια Kaplan-Meier για τη μεταβλητή φύλο (sex)

```
#Kaplan-Meier
km <- survfit(Surv( time , status )~sex, data = cancer )
ggsurvplot ( km, data = cancer,conf.int = TRUE )
summary(km)
km$surv
```

Listing 2: Kaplan-Meier estimation and graphical representation



Στο παραπάνω γράφημα παρατηρούμε ότι υπάρχουν διαφορές ανάλογα με το φύλο στη πιθανότητα επιβίωσης όσο περνάει ο χρόνος.Στο γράφημα φαίνονται τα 95% διαστήματα εμπιστοσύνης για την εκτίμηση κάθε κατηγορίας (Γυναίκες- Άνδρες) και φαίνεται ότι υπάρχει διαφορά ανάμεσα τους.Οι γυναίκες φαίνεται να έχουν μεγαλύτερη πιθανότητα επιβίωσης από ότι οι άντρες καθώς η καμπύλη τους είναι διαρκώς πάνω από αυτή των αντρών.Η μεγαλύτερη διαφορά εντοπίζεται στο διάστημα που ο χρόνος είναι απο 230 εώς 750.Για να επιβεβαιώσουμε το συμπέρασμα αυτό θα κάνουμε log-rank έλεγχο.

2.2 Έλεγχος logrank

```r
survsex <- Surv(cancer$time/365, cancer$status)
km_1 <- survfit( survsex ~ cancer$sex, data =
    cancer,conf.type = "log-log")
logrank_test <- survdiff(survsex ~ sex, data = cancer, rho=0)

print(logrank_test)
```

Listing 3: logrank

Table 1: Αποτελέσματα του ελέγχου log-rank

	N	Observed	Expected	$\frac{(O-E)^2}{E}$
sex=1	138	112	91.6	4.55
sex=2	90	53	73.4	5.68
Chisq= 10.3 on 1 degrees of freedom, p= 0.001				

Στον πίνακα 1 απεικονίζονται τα αποτελέσματα του ελέγχου logrank. Παρατηρούμε ότι το p-value έχει τιμή $0.001 < 0.05$, άρα επιβεβαιώνεται η αρχική εκτίμηση απο το γράφημα της kaplan-meier, δηλαδή η συνάρτηση επιβίωσης διαφέρει ανάμεσα στους άνδρες και τις γυναίκες του δείγματος.

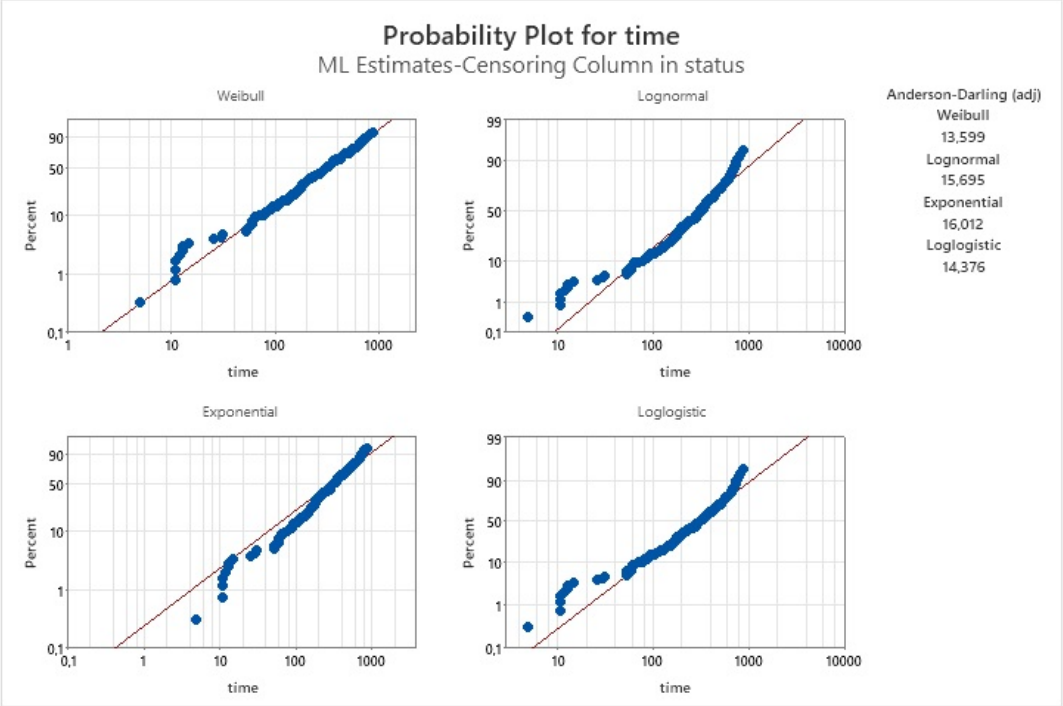
3 Goodness of fit test

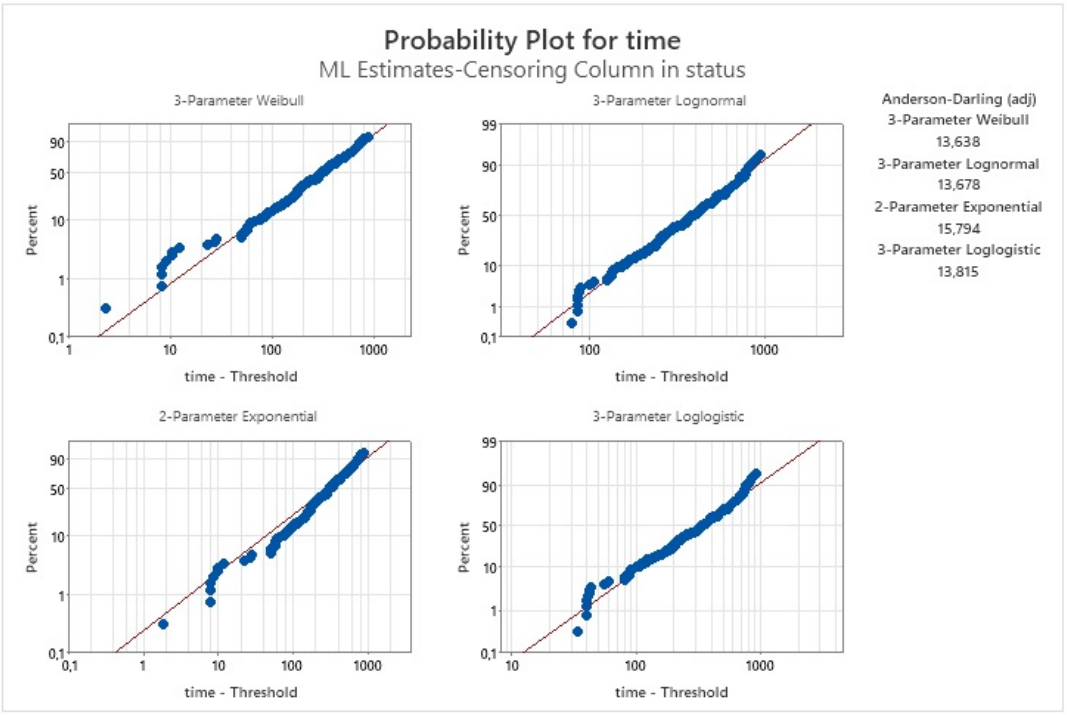
Παρακάτω θα κάνουμε τον έλεγχο καλής προσαρμογής Anderson-Darling ώστε να δούμε ποιά κατανομή προσεγγίζει καλύτερα τα δεδομένα μας.Επιλέγουμε την κατανομή με το χαμηλότερο δείκτη Anderson-Darling.

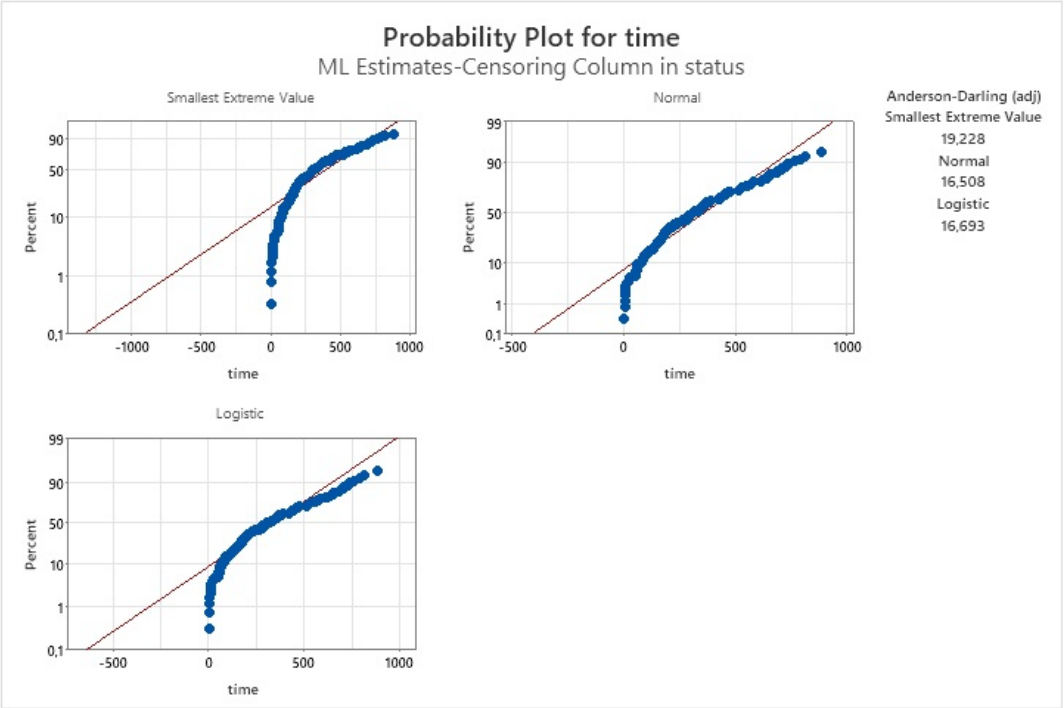
Goodness-of-Fit

Distribution	Anderson-Darling (adj)
Weibull	13,599
Lognormal	15,695
Exponential	16,012
Loglogistic	14,376
3-Parameter Weibull	13,638
3-Parameter Lognormal	13,678
2-Parameter Exponential	15,794
3-Parameter Loglogistic	13,815
Smallest Extreme Value	19,228
Normal	16,508
Logistic	16,693

Η κατανομή που προσεγγίζει καλύτερα τα δεδομένα είναι η Weibull με δείκτη Andrerson-Darling 13.599.Παρακάτω θα δούμε τις γραφικές παραστάσεις όλων των κατανομών που εξετάζονται στον έλεγχο.Αυτό που παρατηρήσαμε απο το δείκτη Andrerson-Darling επιβεβαιώνεται και απο τη γραφική παράσταση, καθώς φαίνεται ότι η γραφική παράσταση της κατανομής Weibull είναι πιο κοντά στα δεδομένα του dataset μας.(Φαίνεται οι κουκίδες να πέφτουν καλύτερα πάνω και γύρω απο τη γραμμή)




Probability Plot for time
ML Estimates-Censoring Column in status
3-Parameter Weibull
3-Parameter Lognormal
2-Parameter Exponential
3-Parameter Loglogistic
Percent
time - Threshold
Anderson-Darling (adj)
3-Parameter Weibull
13,638
3-Parameter Lognormal
13,678
2-Parameter Exponential
15,794
3-Parameter Loglogistic
13,815


Probability Plot for time
ML Estimates-Censoring Column in status
Smallest Extreme Value
Normal
Logistic
Percent
time
Anderson-Darling (adj)
Smallest Extreme Value
19,228
Normal
16,508
Logistic
16,693

4 Εκτιμητές Μέγιστης Πιθανοφάνειας (MLEs)

Σε αυτή την ενότητα θα υπολογίσουμε τη συνάρτηση μέγιστης πιθανοφάνειας της Weibull κατανομής με τη χρήση του λογισμικού Minitab και στη συνέχεια στην R και θα συγκρίνουμε τα αποτελέσματα.

4.1 Minitab

Distribution Analysis: time

Variable: time

Censoring

Censoring Information	Count
Uncensored value	165
Right censored value	63

Censoring value: status = 1

Estimation Method: Maximum Likelihood

Distribution: Weibull

Parameter Estimates

		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Shape	1,31684	0,0822107	1,16518	1,48824
Scale	417,759	24,7045	372,039	469,096

Log-Likelihood = -1153,851

Goodness-of-Fit

Anderson-Darling (Adjusted)
13,599

Σύμφωνα με το Minitab οι εκτιμίσεις της Weibull είναι:

- shape=1.31684
- scale=417.759
- Log-Likelihood = -1153.851

4.2 R

```
data(cancer , package = "survival")
cancer$status<- ifelse(cancer$status == 1, 0, 1)


log_Lik_Weib <- function(par, dataT,dataD){
  shape <- par[1]
  scale <- par[2]
  x <- dataT
  d <- dataD
  loglik <-  sum(d*(-(x**shape/scale**shape) - log(scale) +log(shape) + (-1 +
shape)*(-log(scale) + log(x)))+(1-d)*(-(x**shape/scale**shape)))
  return(-loglik)
}

optim(c(1,400),log_Lik_Weib,dataT=cancer$time, dataD=cancer$status)
```

Parameter	**Value**
shape	1.316813
scale	417.758143
Optimization Results	
Value	1153.851
Counts (Function)	73
Counts (Gradient)	NA
Convergence	0
Message	NULL

Εισάγω την εντολή ifelse ώστε να μετατρέψω τις τιμές της μεταβλητής status απο 1 και 2 σε 0 και 1 αντίστοιχα διότι με βολεύει στον υπολογισμό της Log-Likelihood.

5 Εκτίμηση Μοντέλου Αναλογικών Κινδύνων (Μοντέλο Cox)

```
cmx<-coxph(Surv(cancer$time,cancer$status)~.,data=cancer)
summary(cmx)
ggforest(cmx,data=cancer)
```

Listing 4: Cox Model

Μεταβλητή	coef	exp(coef)	se(coef)	z	Pr(>|z|)	lower .95	upper .95
inst	-0.0304	0.9701	0.0131	-2.315	0.0206	0.9455	0.9954
age	0.0128	1.0130	0.0119	1.073	0.2834	0.9895	1.0369
sex	-0.5666	0.5674	0.2014	-2.814	0.0049	0.3824	0.8420
ph.ecog	0.9074	2.4778	0.2386	3.803	0.0001	1.5523	3.9552
ph.karno	0.0266	1.0269	0.0116	2.286	0.0222	1.0038	1.0506
pat.karno	-0.1091	0.9891	0.0081	-1.340	0.1802	0.9735	1.0051
meal.cal	2.60×10^{-6}	1.0000	2.68×10^{-4}	0.010	0.9922	0.9995	1.0005
wt.loss	-0.0167	0.9834	0.0079	-2.112	0.0346	0.9683	0.9988

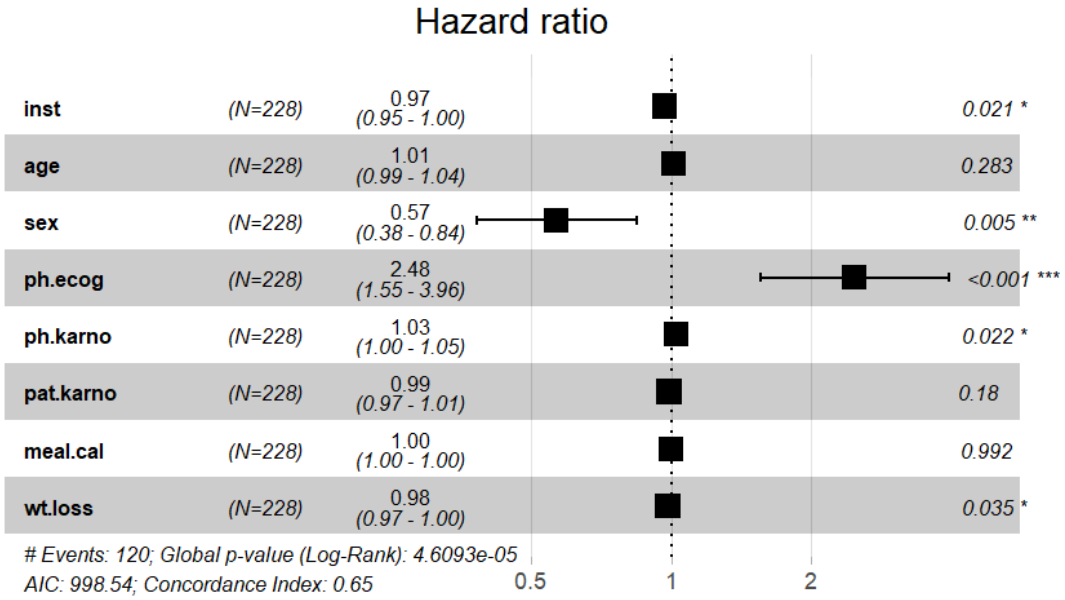
Table 2: Στατιστική ανάλυση Cox Proportional Hazards

Στατιστικά Τεστ		
Concordance	0.648	$(se = 0.03)$
Likelihood ratio test	33.7	$(8 \text{ df}, p = 5 \times 10^{-5})$
Wald test	31.72	$(8 \text{ df}, p = 1 \times 10^{-4})$
Score (logrank) test	32.51	$(8 \text{ df}, p = 8 \times 10^{-5})$

Table 3: Στατιστικά τεστ

Στατιστικά σημαντικές φαίνεται να είναι οι μεταβλητές inst, sex, ph.ecog, ph.karno, wt.loss. Επειδή η εκτίμηση του συντελεστή της μεταβλητής inst είναι (-0.03037) αρνητική, έχουμε ότι κάθε μοναδιαία αύξηση της μεταβλητής αυτής, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές, συνεπάγεται μείωση του κινδύνου κατά έναν συντελεστή της τάξης του 0.9701. Η εκτίμηση του συντελεστή της μεταβλητής sex είναι (-0.5666) αρνητική, έχουμε ότι κάθε μοναδιαία αύξηση της μεταβλητής αυτής, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές, συνεπάγεται μείωση του κινδύνου κατά έναν συντελεστή της τάξης του 0.5674. Η εκτίμηση του συντελεστή της μεταβλητής ph.ecog είναι (0.9074) θετική, έχουμε ότι κάθε μοναδιαία αύξηση της μεταβλητής αυτής, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές, συνεπάγεται αύξηση του κινδύνου κατά έναν συντελεστή της τάξης

του 2.4778. Η εκτίμηση του συντελεστή της μεταβλητής ph.karno είναι (0.02658) θετική, έχουμε ότι κάθε μοναδιαία αύξηση της μεταβλητής αυτής, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές, συνεπάγεται αύξηση του κινδύνου κατά έναν συντελεστή της τάξης του 2.4778 και τέλος η εκτίμηση του συντελεστή της μεταβλητής wt.loss είναι (-0.01671) αρνητική, έχουμε ότι κάθε μοναδιαία αύξηση της μεταβλητής αυτής, δεδομένου ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές, συνεπάγεται μείωση του κινδύνου κατά έναν συντελεστή της τάξης του 0.9834. Τέλος παρατηρούμε οτι και οι τρείς έλεγχοι το Likelihood ratio test, Wald test, Score (logrank) test τα οποία ελέγχουν τη μηδενική υπόθεση ότι όλοι οι συντελεστές βi είναι μηδέν έναντι της εναλλακτικής, ότι υπάρχει τουλάχιστον ένας συντελεστής διαφορετικός του μηδενός. Και οι τρεις έλεγχοι συμφωνούν στην απόρριψη, σε όλα τα συνήθη επίπεδα σημαντικότητας, της μηδενικής υπόθεσης, αφού οι p-value τιμές τους είναι εξαιρετικά μικρές.



Το διάγραμμα παρουσιάζει τις αναλογίες κινδύνου (Hazard Ratios) και τα αντίστοιχα διαστήματα εμπιστοσύνης για διάφορους παράγοντες που περιλαμβάνονται σε ένα μοντέλο Cox για την ανάλυση επιβίωσης. Ακολουθεί ο σχολιασμός του διαγράμματος:

1. **Γενική Εικόνα**:

   - Το διάγραμμα παρουσιάζει τις αναλογίες κινδύνου (HR) για οκτώ διαφορετικούς παράγοντες.
   - Οι αναλογίες κινδύνου συνοδεύονται από 95% διαστήματα εμπιστοσύνης (95% ΔΕ).
   - Οι παράγοντες που εμφανίζονται περιλαμβάνουν το νοσοκομείο (inst), την ηλικία (age), το φύλο (sex), το ph.ecog, το ph.karno, το pat.karno, την κατανάλωση θερμίδων στο γεύμα (meal.cal) και την απώλεια βάρους (wt.loss).

2. **Ερμηνεία των Αναλογιών Κινδύνου**:

   - **inst**: Η αναλογία κινδύνου είναι 0.97 με 95% ΔΕ 0.95 - 1.00, με p-value 0.021. Αυτό δείχνει ότι ο παράγοντας αυτός έχει στατιστικά σημαντική επίδραση στη θνησιμότητα.
   - **age**: Η αναλογία κινδύνου είναι 1.01 με 95% ΔΕ 0.99 - 1.04, με p-value 0.283. Αυτό υποδηλώνει ότι η ηλικία δεν έχει στατιστικά σημαντική επίδραση στην επιβίωση.
   - **sex**: Η αναλογία κινδύνου είναι 0.57 με 95% ΔΕ 0.38 - 0.84, με p-value 0.005. Οι άνδρες φαίνεται να έχουν χαμηλότερο κίνδυνο θανάτου σε σύγκριση με τις γυναίκες, κάτι που είναι στατιστικά σημαντικό.
   - **ph.ecog**: Η αναλογία κινδύνου είναι 2.48 με 95% ΔΕ 1.55 - 3.96, με p-value <0.001. Αυτός ο παράγοντας έχει ισχυρή και στατιστικά σημαντική επίδραση στον κίνδυνο θανάτου.
   - **ph.karno**: Η αναλογία κινδύνου είναι 1.03 με 95% ΔΕ 1.00 - 1.05, με p-value 0.022. Εδώ, υπάρχει μια μικρή αλλά στατιστικά σημαντική επίδραση.
   - **pat.karno**: Η αναλογία κινδύνου είναι 0.99 με 95% ΔΕ 0.97 - 1.01, με p-value 0.18. Αυτή η επίδραση δεν είναι στατιστικά σημαντική.
   - **meal.cal**: Η αναλογία κινδύνου είναι 1.00 με 95% ΔΕ 1.00 - 1.00, με p-value 0.992. Αυτός ο παράγοντας δεν έχει καμία επίδραση στην επιβίωση.
   - **wt.loss**: Η αναλογία κινδύνου είναι 0.98 με 95% ΔΕ 0.97 - 1.00, με p-value 0.035. Η απώλεια βάρους έχει μια μικρή αλλά στατιστικά σημαντική επίδραση στη θνησιμότητα.

3. **Στατιστική Σημαντικότητα**:

   - Οι παράγοντες με p-value κάτω από 0.05 θεωρούνται στατιστικά σημαντικοί. Αυτοί είναι οι: inst, sex, ph.ecog, ph.karno, και wt.loss.

- Τα επίπεδα σημαντικότητας είναι σημειωμένα με *, **, και *** ανάλογα με την τιμή του p-value. Το ph.ecog είναι εξαιρετικά σημαντικό με p-value <0.001 (***).

4. **Συνολική Επίδοση του Μοντέλου**:

   - Το μοντέλο περιλαμβάνει 120 events και έχει συνολικό p-value (Log-Rank) 4.6093e-05, που δείχνει ότι το μοντέλο είναι στατιστικά σημαντικό συνολικά.
   - Η AIC (998.54) και η Concordance Index (0.65) δίνουν μια ιδέα για την καλή προσαρμογή και την προβλεπτική ικανότητα του μοντέλου αντίστοιχα.

Συμπεράσματα:

- Ορισμένοι παράγοντες έχουν σημαντική επίδραση στην επιβίωση, όπως το φύλο και η κλίμακα ph.ecog, υποδεικνύοντας ότι πρέπει να ληφθούν σοβαρά υπόψη στην ανάλυση κινδύνου.
- Άλλοι παράγοντες, όπως η ηλικία και η κατανάλωση θερμίδων, δεν φαίνεται να έχουν στατιστικά σημαντική επίδραση στην επιβίωση, υποδεικνύοντας ότι μπορεί να μην είναι τόσο κρίσιμοι για το συγκεκριμένο μοντέλο.

6 Εύρεση του 'Καλύτερου' Μοντέλου

```
cancer_nomissing<- na.omit(cancer)
cxx<-coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)~.
,data=cancer_nomissing)
library(MASS)
All_cox<-
    coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)
    ~ ., data=cancer_nomissing)
fit0 <- coxph(Surv(cancer_nomissing$time,
    cancer_nomissing$status) ~ 1, data=cancer_nomissing)
fitf <- stepAIC(fit0, scope=formula(All_cox),
    direction="forward", k=2)
summary(fitf)
cmxx<-
    coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)~.
,data=cancer_nomissing)
summary(cmxx)
All_cox_back <- stepAIC(cmxx,scope=formula(All_cox),
    direction="backward", k=2)
fit <- stepAIC(All_cox, direction="both", k=2)
summary(fit)
```

Listing 5: Model Selection

AIC FORWARD METHOD

```
     Start:  AIC=1016.23
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ 1

              Df     AIC
+ ph.ecog      1  1005.8
+ pat.karno    1  1009.4
+ sex          1  1012.0
+ age          1  1014.7
+ ph.karno     1  1015.0
<none>           1016.2
+ inst         1  1017.0
+ meal.cal     1  1018.0
+ wt.loss      1  1018.2

Step:  AIC=1005.82
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog

              Df     AIC
+ sex          1  1000.8
+ inst         1  1004.1
+ ph.karno     1  1005.7
<none>           1005.8
+ wt.loss      1  1006.3
+ pat.karno    1  1006.5
+ age          1  1007.0
+ meal.cal     1  1007.8

Step:  AIC=1000.75
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex
              Df      AIC
+ inst         1   999.31
+ ph.karno     1  1000.07
+ wt.loss      1  1000.17
<none>           1000.75
+ pat.karno    1  1001.86
+ age          1  1002.22
+ meal.cal     1  1002.66
```

```
Step:  AIC=999.31
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + inst

             Df     AIC
+ wt.loss     1  997.70
+ ph.karno    1  997.77
<none>           999.31
+ pat.karno   1 1000.66
+ age         1 1000.75
+ meal.cal    1 1001.15

Step:  AIC=997.7
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + inst + wt.loss

             Df     AIC
+ ph.karno    1  995.79
<none>           997.70
+ pat.karno   1  998.41
+ age         1  999.27
+ meal.cal    1  999.58

Step:  AIC=995.79
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + inst + wt.loss + ph.karno

             Df     AIC
+ pat.karno   1  995.74
<none>           995.79
+ age         1  996.38
+ meal.cal    1  997.60
```

```
Step:  AIC=995.74
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + inst + wt.loss + ph.karno + pat.karno

            Df     AIC
<none>         995.74
+ age        1 996.54
+ meal.cal   1 997.71
```

AIC BACKWARD METHOD

```
Start:  AIC=998.54
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss

             Df     AIC
- meal.cal    1  996.54
- age         1  997.71
- pat.karno   1  998.33
<none>           998.54
- wt.loss     1 1001.28
- inst        1 1002.07
- ph.karno    1 1002.12
- sex         1 1004.88
- ph.ecog     1 1011.46

Step:  AIC=996.54
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + wt.loss

             Df     AIC
- age         1  995.74
- pat.karno   1  996.38
<none>           996.54
- wt.loss     1  999.28
- inst        1 1000.08
- ph.karno    1 1000.12
- sex         1 1003.02
- ph.ecog     1 1009.46
```

```
Step:  AIC=995.74
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + ph.karno + pat.karno + wt.loss

             Df     AIC
<none>           995.74
- pat.karno  1  995.79
- ph.karno   1  998.41
- wt.loss    1  998.70
- inst       1  998.95
- sex        1 1002.17
- ph.ecog    1 1008.23
```

AIC BOTH METHOD

```
Start:  AIC=998.54
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss

             Df     AIC
- meal.cal   1  996.54
- age        1  997.71
- pat.karno  1  998.33
<none>           998.54
- wt.loss    1 1001.28
- inst       1 1002.07
- ph.karno   1 1002.12
- sex        1 1004.88
- ph.ecog    1 1011.46
```

```
Step:  AIC=996.54
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + wt.loss

             Df     AIC
- age         1  995.74
- pat.karno   1  996.38
<none>           996.54
+ meal.cal    1  998.54
- wt.loss     1  999.28
- inst        1 1000.08
- ph.karno    1 1000.12
- sex         1 1003.02
- ph.ecog     1 1009.46

Step:  AIC=995.74
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + ph.karno + pat.karno + wt.loss

             Df     AIC
<none>           995.74
- pat.karno   1  995.79
+ age         1  996.54
+ meal.cal    1  997.71
- ph.karno    1  998.41
- wt.loss     1  998.70
- inst        1  998.95
- sex         1 1002.17
- ph.ecog     1 1008.23
```

Εισάγουμε όλες τις διαθέσιμες μεταβλητές στο μοντέλο. Αφαιρούμε τη λιγότερο σημαντική μεταβλητή, δηλαδή αυτήν που η αφαίρεσή της προκαλεί τη μεγαλύτερη μείωση του AIC. Προσαρμόζουμε εκ νέου ένα μοντέλο παλινδρόμησης στα δεδομένα, παραλείποντας τη μεταβλητή που αφαιρέσαμε. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι η αφαίρεση μιας οποιασδήποτε μεταβλητής να έχει ως συνέπεια την αύξηση της τιμής του AIC οπότε και σταματάμε τη διαδικασία. Αρχικά το AIC με την forward μέθοδο είναι 1016.23.Παρατηρούμε ότι η καλύτερη μεταβλητή για να βάλουμε στη αρχή είναι η ph.ecog 'οπου κάνει το AIC 1005.82.Αμέσως μετά βάζουμε τη μεταβλητή sex όπου κάνει το AIC 1000.75.Αμέσως μετά βάζουμε τη μεταβλητή inst και το AIC γίνεται 999.31.Ακολουθεί η μεταβλητή wt.loss και το AIC γίνεται 997.7 μετά προσθέτουμε τη μεταβλητή ph.karno και το AIC γίνεται 995.79 και τέλος προσθέ-

τουμε τη μεταβλητή pat.karno και το AIC γίνεται 995.74.Απο κει και πέρα η καλύτερη επιλογή είναι να μην βάλουμε άλλη μεταβλητή καθώς αν προστεθεί άλλη θα ανέβει ο δείκτης AIC.Το AIC με την backward μέθοδο αρχικά είναι 998.54 πρώτα αφαιρούμε τη μεταβλητή meal.cal και το AIC γίνεται 996.54 έπειτα αφαιρούμε τη μεταβλητή age και το AIC γίνεται 995.74.Πλέον δε χρειάζεται να βγάλουμε άλλη μεταβλητή καθώς η αφαίρεση οποιασδήποτε άλλης μεταβλητής θα αυξήσει το AIC.Με τη μέθοδο both παρατηρούμε ότι γίνεται ακριβώς η ίδια διαδικασία με τη μέθοδο backward.

7 Έλεγχος της Υπόθεσης Αναλογικών Κινδύνων

```
res<- residuals(cmx,type='schoenfeld')
summary(res)
(zph1<- cox.zph(All_cox_back,transform = 'identity'))
ggcoxzph(zph1)
```

Listing 6: Proportional Hazards Assumption

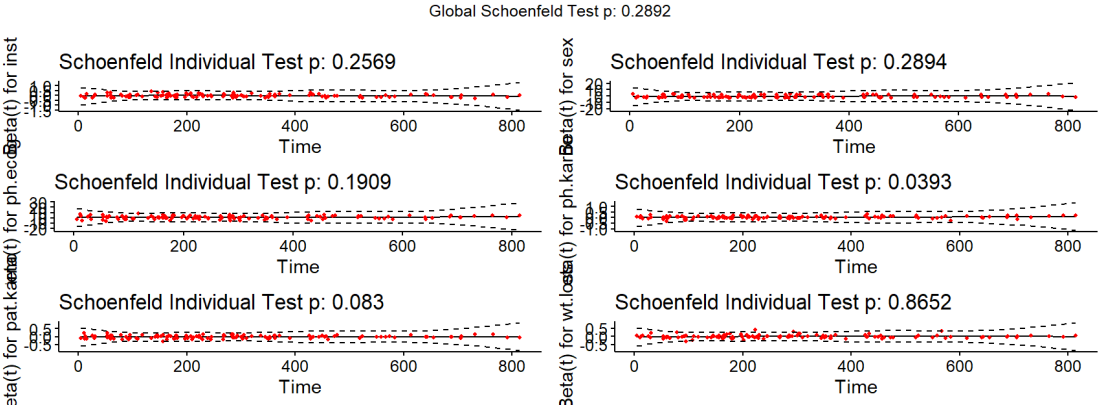
Πληροφορίες για τα υπόλοιπα

inst	age	sex
Min. :−11.8780	Min. :−19.990	Min. :−0.3866
1st Qu.: −6.7458	1st Qu.: −6.405	1st Qu.:−0.3272
Median : 0.4254	Median : 1.059	Median :−0.2836
Mean : 0.0000	Mean : 0.000	Mean : 0.0000
3rd Qu.: 4.8131	3rd Qu.: 6.856	3rd Qu.: 0.6403
Max. : 22.3206	Max. : 17.869	Max. : 0.7650
ph.ecog	ph.karno	pat.karno
Min. :−1.30491	Min. :−30.0693	Min. :−46.218
1st Qu.:−0.23612	1st Qu.:−10.8517	1st Qu.:−10.064
Median :−0.00615	Median : −0.7902	Median : 0.777
Mean : 0.00000	Mean : 0.0000	Mean : 0.000
3rd Qu.: 0.72652	3rd Qu.: 8.8267	3rd Qu.: 10.730
Max. : 1.80655	Max. : 22.0114	Max. : 26.102
meal.cal	wt.loss	
Min. :−755.085	Min. :−33.327	
1st Qu.:−242.551	1st Qu.: −9.252	
Median : 9.469	Median : −1.630	
Mean : 0.000	Mean : 0.000	
3rd Qu.: 239.861	3rd Qu.: 5.480	
Max. :1537.222	Max. : 57.163	

Έλεγχος Αναλογικής Διακινδύνευσης

	chisq	df	p
inst	1.2856	1	0.257
sex	1.1223	1	0.289
ph.ecog	1.7105	1	0.191
ph.karno	4.2479	1	0.039
pat.karno	3.0053	1	0.083
wt.loss	0.0288	1	0.865
GLOBAL	7.3560	6	0.289

Υπόλοιπα Schoenfeld



Χρησιμοποιώντας το μοντέλο που επιλέχθηκε με τη διαδικασία της διαδοχικής αφαίρεσης, παρατηρούμε οτι η υπόθεση αναλογικής διακινδύνευσης ισχύει για τις μεταβλητές inst,sex,ph.ecog,pat.karno,wt.loss και για το GLOBAL ενώ η υπόθεση απορρίπτεται για τη μεταβλητή ph.karno.
Παραπάνω βλέπουμε επίσης την εικόνα με τα υπόλοιπα schoenfeld για τα οποία βλέπουμε ότι πρέπει να είναι μέσα σε κάποια όρια.

Για να επιλύσουμε το πρόβλημα με τη μεταβλητή ph.karno επιλέγουμε να χωρίσουμε τη μεταβλητή ph.karno σε δύο υποκατηγορίες.Η πρώτη θα περιλαμβάνει τις τιμές που είναι απο 0 εως 50 και η δεύτερη τις τιμές που είναι απο 51 εως 100.Ουσιαστικα σε μια νεα μεταβλητή την cat.karno θα βάλουμε A για την πρώτη κατηγορία και B για την δεύτερη.

```r
cancer$cat.karno <- as.factor(ifelse(cancer$ph.karno <=
    50,'A',
                                ifelse(cancer$ph.karno > 50,
                                    'B', NA)))

head(cancer)
```

Listing 7: Categorizing Karnofsky Score

```r
cancer_nomissing<- na.omit(cancer)
cmxx<-coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)~.
,data = cancer_nomissing)
library(MASS)
All_cox <-
    coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)
    ~ . , data=cancer_nomissing)
fit0 = coxph(Surv(cancer_nomissing$time,
    cancer_nomissing$status) ~ 1, data=cancer_nomissing)
fitf = stepAIC(fit0, scope=formula(All_cox),
    direction="forward", k=2)
summary(fitf)

cmxxx<-coxph(Surv(cancer_nomissing$time,cancer_nomissing$status)~.
,data=cancer_nomissing)
summary(cmxxx)
All_cox_back = stepAIC(cmxxx,scope=formula(cmxxx),
    direction="backward", k=2)

fits = stepAIC(All_cox, direction="both", k=2)
summary(fits)
```

Listing 8: Refitting Model

AIC FORWARD METHOD

```
     Start:  AIC=1016.23
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ 1

              Df     AIC
+ ph.ecog      1  1005.8
+ pat.karno    1  1009.4
+ sex          1  1012.0
+ age          1  1014.7
+ ph.karno     1  1015.0
<none>            1016.2
+ inst         1  1017.0
+ cat.karno    1  1017.1
+ meal.cal     1  1018.0
+ wt.loss      1  1018.2

Step:  AIC=1005.82
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog

              Df     AIC
+ sex          1  1000.8
+ inst         1  1004.1
+ cat.karno    1  1004.8
+ ph.karno     1  1005.7
<none>            1005.8
+ wt.loss      1  1006.3
+ pat.karno    1  1006.5
+ age          1  1007.0
+ meal.cal     1  1007.8
```

```
Step:  AIC=1000.75
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex

             Df     AIC
+ cat.karno  1  997.75
+ inst       1  999.31
+ ph.karno   1 1000.07
+ wt.loss    1 1000.17
<none>         1000.75
+ pat.karno  1 1001.86
+ age        1 1002.22
+ meal.cal   1 1002.66

Step:  AIC=997.75
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + cat.karno

             Df     AIC
+ wt.loss    1  996.46
+ inst       1  996.62
<none>         997.75
+ age        1  998.84
+ pat.karno  1  999.01
+ meal.cal   1  999.55
+ ph.karno   1  999.58

Step:  AIC=996.46
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + cat.karno + wt.loss

             Df     AIC
+ inst       1  994.23
<none>         996.46
+ pat.karno  1  996.88
+ age        1  997.66
+ meal.cal   1  998.26
+ ph.karno   1  998.35
```

```
Step:  AIC=994.23
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ ph.ecog +
    sex + cat.karno + wt.loss + inst

             Df     AIC
<none>           994.23
+ pat.karno   1 994.93
+ age         1 995.31
+ ph.karno    1 995.68
+ meal.cal    1 995.91
```

AIC BACKWARD METHOD

```
Start:  AIC=998.87
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss +
    cat.karno

             Df     AIC
- meal.cal    1  996.88
- age         1  997.93
- ph.karno    1  998.15
- pat.karno   1  998.28
- cat.karno   1  998.54
<none>           998.87
- inst        1 1001.72
- wt.loss     1 1001.91
- sex         1 1006.34
- ph.ecog     1 1007.21
```

```
Step:  AIC=996.88
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + wt.loss + cat.karno

             Df     AIC
- age         1  996.00
- ph.karno    1  996.21
- pat.karno   1  996.40
- cat.karno   1  996.54
<none>           996.88
- inst        1  999.72
- wt.loss     1  999.93
- sex         1 1004.46
- ph.ecog     1 1005.35

Step:  AIC=996
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + ph.karno + pat.karno + wt.loss + cat.karno

             Df     AIC
- ph.karno    1  994.93
- pat.karno   1  995.68
- cat.karno   1  995.74
<none>           996.00
- inst        1  998.53
- wt.loss     1  999.22
- sex         1 1003.55
- ph.ecog     1 1004.13
```

```
Step:  AIC=994.93
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + pat.karno + wt.loss + cat.karno

             Df     AIC
- pat.karno   1  994.23
<none>           994.93
- inst        1  996.88
- wt.loss     1  998.05
- cat.karno   1  998.41
- sex         1 1003.03
- ph.ecog     1 1005.64

Step:  AIC=994.23
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + wt.loss + cat.karno

             Df     AIC
<none>           994.23
- inst        1  996.46
- wt.loss     1  996.62
- cat.karno   1  997.70
- sex         1 1002.70
- ph.ecog     1 1014.94
```

AIC BOTH METHOD

```
Start:  AIC=998.87
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss +
    cat.karno

             Df     AIC
- meal.cal    1  996.88
- age         1  997.93
- ph.karno    1  998.15
- pat.karno   1  998.28
- cat.karno   1  998.54
<none>           998.87
- inst        1 1001.72
- wt.loss     1 1001.91
- sex         1 1006.34
- ph.ecog     1 1007.21

Step:  AIC=996.88
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    age + sex + ph.ecog + ph.karno + pat.karno + wt.loss + cat.karno

             Df     AIC
- age         1  996.00
- ph.karno    1  996.21
- pat.karno   1  996.40
- cat.karno   1  996.54
<none>           996.88
+ meal.cal    1  998.87
- inst        1  999.72
- wt.loss     1  999.93
- sex         1 1004.46
- ph.ecog     1 1005.35
```

```
Step:  AIC=996
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + ph.karno + pat.karno + wt.loss + cat.karno

             Df     AIC
- ph.karno    1  994.93
- pat.karno   1  995.68
- cat.karno   1  995.74
<none>           996.00
+ age         1  996.88
+ meal.cal    1  997.93
- inst        1  998.53
- wt.loss     1  999.22
- sex         1 1003.55
- ph.ecog     1 1004.13

Step:  AIC=994.93
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + pat.karno + wt.loss + cat.karno

             Df     AIC
- pat.karno   1  994.23
<none>           994.93
+ ph.karno    1  996.00
+ age         1  996.21
+ meal.cal    1  996.81
- inst        1  996.88
- wt.loss     1  998.05
- cat.karno   1  998.41
- sex         1 1003.03
- ph.ecog     1 1005.64
```

```
Step:  AIC=994.23
Surv(cancer_nomissing$time, cancer_nomissing$status) ~ inst +
    sex + ph.ecog + wt.loss + cat.karno

             Df     AIC
<none>           994.23
+ pat.karno   1  994.93
+ age         1  995.31
+ ph.karno    1  995.68
+ meal.cal    1  995.91
- inst        1  996.46
- wt.loss     1  996.62
- cat.karno   1  997.70
- sex         1 1002.70
- ph.ecog     1 1014.94
```

Ξανακάνουμε AIC και με τις 3 μεθόδους(forward,backward,both) με την ίδια λογική όπως και πριν καταλήγοντας στο τελικό μοντέλο με AIC 994.23 και οι μεταβλητές που μένουν είναι inst,sex,wt.loss,cat.karno,ph.ecog.

```r
(zph2<- cox.zph(All_cox_back,transform = 'identity'))
```

Listing 9: Testing Proportional Hazards

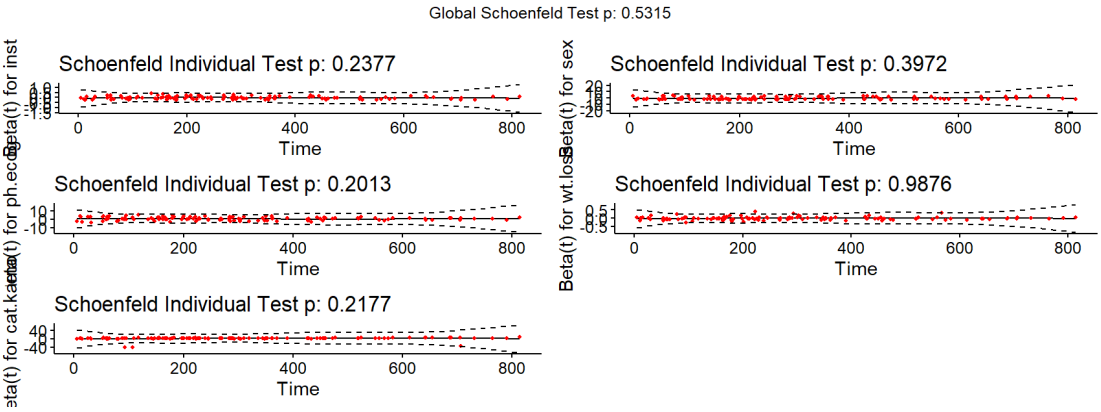
Τα αποτελέσματα είναι:

	chisq	df	p
inst	1.393998	1	0.24
sex	0.716728	1	0.40
ph.ecog	1.632736	1	0.20
wt.loss	0.000241	1	0.99
cat.karno	1.519435	1	0.22
GLOBAL	4.125187	5	0.53

Πλέον στην υπόθεση αναλογικής διακινδύνευσης ισχύει για όλες τις μεταβλητές (inst,sex,wt.loss,cat.karno,ph.ecog) αφού το p-value τους είναι μεγαλύτερο του 0.05 (0.24,0.4,0.99,0.22,0.20) όπως και το p-value του GLOBAL (0.53).

```
1 ggcoxzph(zph2)
```

Listing 10: Graphical Diagnostics

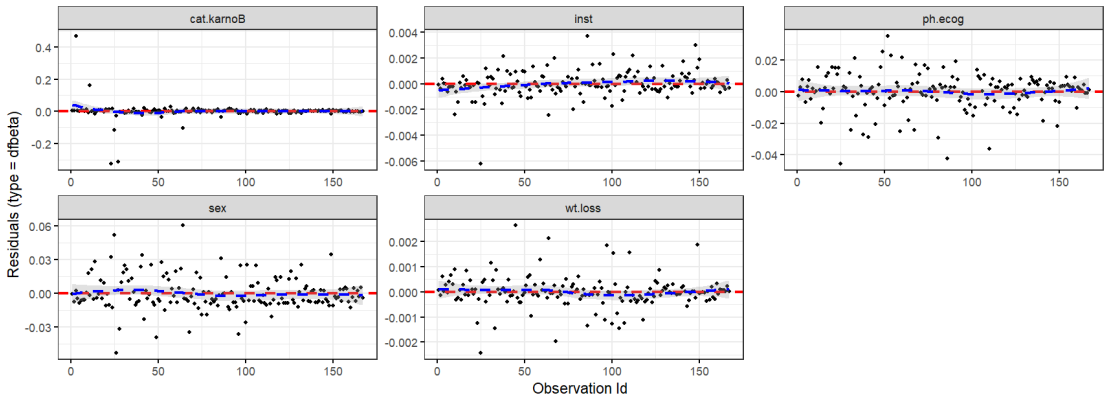


8 Έλεγχος για Επηρεαστικές Παρατηρήσεις

```
ggcoxdiagnostics(fits, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme =
                    theme_bw())
```

Listing 11: Influential Observations

Με αυτή την εντολή κάνω τα διαγράμματα των υπολοίπων με τη μέθοδο dfbeta.



Για τη μεταβλητή cat.karno έχουμε:

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,5] > 0.1)
cancer[which(resid_dfbeta[,5] > 0.1),]
```

Listing 12: Influential Observations

```
 6 18
```

Για την 6η παρατήρηση όπως και για την 18η παρατηρούμε οτι παρα την χαμηλή βαθμολογία που τους έχει δώσει ο γιατρός 50 και στους δύο έχουν ζησει πολύ καιρό

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,5] < -0.03)
cancer[which(resid_dfbeta[,5] < -0.03),]
```

Listing 13: Influential Observations

```
 34  37  39  71  85 118
```

Για τους ασθενείς 34,37,39 πατηρούμε χαμηλή βαθμολογία απο τον γιατρό αλλα και απο τον εαυτό τους ο ασθενής 37 παρότι δεν ήταν σε καλή κατάσταση επιβίωσε πολύ καιρό ενώ οι άλλοι 2 πέθαναν σε πολυ μικρό χρονικό διάστημα.

Για τη μεταβλητή inst έχουμε:

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,1] > 0.002)
cancer[which(resid_dfbeta[,1] > 0.002),]
```

Listing 14: Influential Observations

```
  53   88  118  139  155  200
```

Για τον ασθενή 53 έχουμε οτι παρότι ηταν ασυμπτωματικός και είχε πολύ καλη βαθμολογία απο το γιατρό πεθανε πολύ γρήγορα(53 μέρες)

Για τον ασθενή 88 παρατηρούμε το ίδιο ακριβώς συμβάν

Για τον ασθενή 118 παρατηρούμε ότι ενα καταναλώνει 1300 θερμίδες ημερησίως έχει χάσει 30 κιλά ο ασθενής

Στον ασθενή 155 παρατηρούμε μεγαλη απόκλιση στη βαθμολογία του γιατρού σε σχέση με αυτή του ασθενή (70 ο γιατρός 30 ο ασθενής)

Για τη μεταβλητή ph.ecog έχουμε:

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,3] > 0.002)
cancer[which(resid_dfbeta[,3] > 0.002),]
```

Listing 15: Influential Observations

```
 46  68  71  73  81
```

Για τον ασθενή 73 ενώ εβαλε στον εαυτό του πολυ καλό σκορ δηλαδή ένιωθε πολυ καλά πέθανε σε μολις 11 μέρες.

Για τον ασθενή 46 έχουμε ότι ενώ δεν φαίνεται να έτρωγε πολύ ήταν σε κακή κατάσταση και πέθανε σύντομα, πήρε 24 κιλά.

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,3] < -0.02)
cancer[which(resid_dfbeta[,3] < -0.02),]
```

Listing 16: Influential Observations

```
 37   43   53   57   62   80   88  111  118  149  201
```

Ο ασθενής 43 παρατηρείται οτι έτρωγε ασυνήθιστα λίγες θερμίδες
Ο ασθενής 53 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα
Ο ασθενής 57 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα
Ο ασθενής 80 φαίνεται να έτρωγε πολύ λίγες θερμίδες
Ο ασθενής 88 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα
Ο ασθενής 111 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα
Ο ασθενής 118 παρότι έτρωγε πολλές θερμίδες έχασε πολλα κιλά
Ο ασθενής 149 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα
Ο ασθενής 201 παρότι είχε καλες αξιολογήσεις πέθανε σε πολύ μικρό χρονικό διάστημα

Για τη μεταβλητή wt.loss έχουμε:

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,4] > 0.002)
cancer[which(resid_dfbeta[,4] > 0.002),]
```

Listing 17: Influential Observations

62 85

```
resid_dfbeta <- residuals(All_cox_back, type = "dfbeta")
which(resid_dfbeta[,4] < -0.002)
cancer[which(resid_dfbeta[,4] < -0.002),]
```

Listing 18: Influential Observations

37

Ο ασθενής 37 παρόλο που δεν είχε καλές βαθμολογίες είχε χάσει αρκετά κιλά έζησε για μεγάλο χρονικό διάστημα

Για τον ασθενή 62 παρόλο που ηταν σε καλή κατάσταση έχασε 68 κιλά.

9 Συμπεράσματα Ακραίων Τιμών

Παρατηρούμε οτι οι ακραίες τιμές εμφανίζονται στις μεταβλητές ph.karno, pat.karno, wt.loss. Πολλοί απο αυτούς τους ασθενείς πέθαναν σε πολύ σύντομο χρονικό διάστημα ενω έδειχναν να είναι σε καλή κατάσταση είτε το αντίθετο. Άλλοι ενώ έτρωγαν ικανοποιητικές θερμίδες την μέρα έχαναν πολλά κιλά και άλλοι που έτρωγαν φυσιολογικά έπαιρναν πολλά κιλά. Επιπλέον σε ορισμένους ασθενείς υπάρχει μεγάλη απόκλιση μεταξύ βαθμολογίας ασθενή και γιατρού. Μπορούμε να υποθέσουμε οτι έχει γίνει κάποιο λάθος στην εισαγωγή των στοιχείων των ασθενών (wt.loss), είτε οτι υπήρχαν άλλοι παράγοντες που δυσκόλευαν την καθημερινότητα των ασθενών χωρίς να το καταλαβαίνουν οι γιατροί (ph.karno-pat.karno).

10 Συμπέρασμα

Σε αυτή την εργασία κάναμε μια έρευνα για τους ασθενείς με καρκίνο του πνεύμονα απο το dataset cancer της βιβλιοθήκης survival της R. Κάναμε έλεγχο kaplan-meier και διαπιστώσαμε διαφορά στο χρόνο επιβίωσης ανάλογα με το φύλο του ασθενή(γυναίκες έχουν μεγαλύτερη πιθανότητα επιβίωσης). Έπειτα καταλήξαμε στο συμπέρασμα οτι η κατανομή weibull προσσεγγίζει καλύτερα τα δεδομένα μας και βρήκαμε τη συγκεκριμένη αυτή μεταβλητή με τις παραμέτρους της. Στη συνέχεια κάναμε cox μοντέλο για να δούμε τι κίνδυνο παρουσιάζει η κάθε μεταβλητή καθώς και το πόσο σημαντικές ειναι οι μεταβλητές Έπειτα διαλέξαμε με τη μέθοδο AIC τις 'καλύτερες' μεταβλητές που απο αυτές παίρνουμε τις περισσότερες πληροφορίες και είναι οι μεταβλητές που μας ενδιαφέρουν και είναι σημαντικές,μετά στον έλεγχο της υπόθεσης αναλογικού κινδύνου δεχτήκαμε όλες τις μεταβλητές εκτός απο το pat.karno και γι αυτο την κατηγοριοποιήσαμε σε 2 μεταβλητες την cat.karnoA και την cat.karnoB ανάλογα με τη βαθμολογία που είχε δώσει ο κάθε ασθενής.Στο τέλος βρήκαμε τις επηρεάζουσες τιμές οι οποίες είναι τιμές που έχουν κάποια παράξενη τιμή και τις σχολιάσαμε.