Haberman Survival Analysis

Rania Ahmed

January 1, 2020

Source: Opendata source https://www.kaggle.com/gilsousa/habermans-survival-data-set

Sources: (a) Donor: Tjen-Sien Lim (limt@stat.wisc.edu) (b) Date: March 4, 1999

Past Usage:

Haberman, S. J. (1976). Generalized Residuals for Log-Linear Models, Proceedings of the 9th International Biometrics Conference, Boston, pp. 104-122. Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), Graphical Models for Assessing Logistic Regression Models (with discussion), Journal of the American Statistical Association 79: 61-83. Lo, W.-D. (1993). Logistic Regression Trees, PhD thesis, Department of Statistics, University of Wisconsin, Madison, WI. Relevant Information: **The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.**

Number of Instances: 306

Number of Attributes: 4 (including the class attribute)

Attribute Information:

1- Age of patient at time of operation (numerical)

2- Patient's year of operation (year - 1900, numerical)

3- Number of positive axillary nodes detected (numerical)

4- Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year

5- Missing Attribute Values: None

Let's begin:

```
library(knitr)
library(dplyr)

library(finalfit)
library(moments)
library(survival)
library(survminer)
```

```
library(ggplot2)
```

Summary of the data:

```
haberman <- read.csv(file.choose(), header = T)
attach(haberman)

colnames(haberman) <-
c("Age","year_operation","Axillary_nodes_detected","surv_status")
haberman <- mutate(haberman,
dummy_surv=ifelse(haberman$surv_status=="1","nodeath","death"),
                   time2=(70-haberman$year_operation),
                   dummy_age=ifelse(haberman$Age > 50,"old","young"))
summary(haberman)
```

```
##       Age        year_operation  Axillary_nodes_detected  surv_status
##  Min.   :30.00   Min.   :58.00   Min.   : 0.000          Min.   :1.000
##  1st Qu.:44.00   1st Qu.:60.00   1st Qu.: 0.000          1st Qu.:1.000
##  Median :52.00   Median :63.00   Median : 1.000          Median :1.000
##  Mean   :52.53   Mean   :62.85   Mean   : 4.036          Mean   :1.266
##  3rd Qu.:61.00   3rd Qu.:66.00   3rd Qu.: 4.000          3rd Qu.:2.000
##  Max.   :83.00   Max.   :69.00   Max.   :52.000          Max.   :2.000
##   dummy_surv            time2         dummy_age
##  Length:305         Min.   : 1.000   Length:305
##  Class :character   1st Qu.: 4.000   Class :character
##  Mode  :character   Median : 7.000   Mode  :character
##                     Mean   : 7.151
##                     3rd Qu.:10.000
##                     Max.   :12.000
```

Analysis of variances:

1-Relation between Age and survived status:

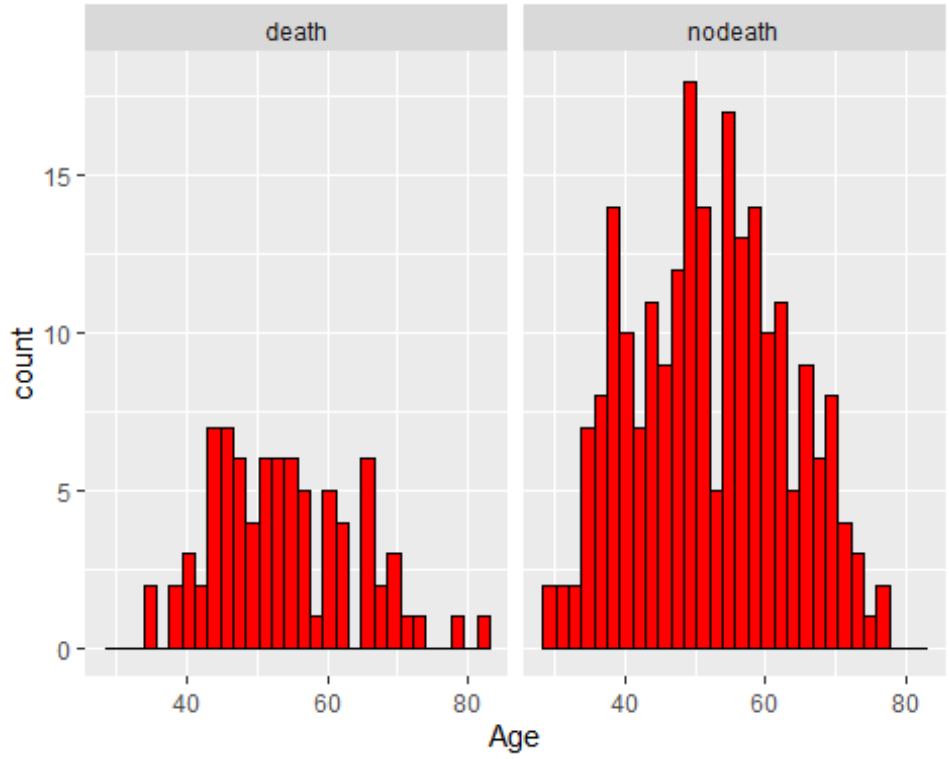
```
summary(haberman$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   44.00   52.00   52.53   61.00   83.00
```

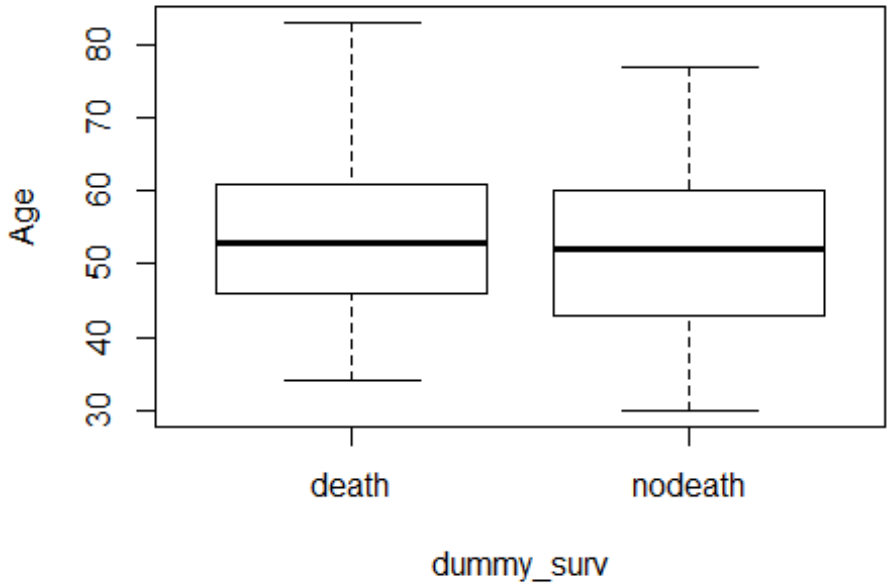
```
Agecoun <- group_by(haberman, dummy_surv)%>%
  summarise(count=n(), mean=mean(Age), sd=sd(Age), var=var(Age))
Agecoun
```

```
## # A tibble: 2 x 5
##   dummy_surv count  mean    sd   var
##   <chr>      <int> <dbl> <dbl> <dbl>
## 1 death         81  53.7  10.2  103.
## 2 nodeath      224  52.1  10.9  120.
```

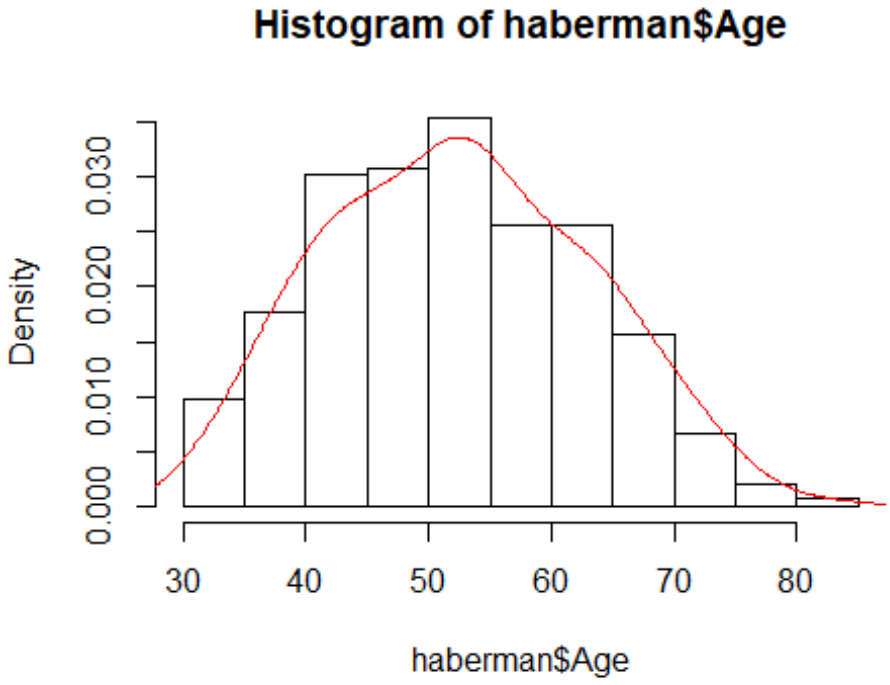
```
ggplot(haberman,aes(Age, fill= Age))+geom_histogram(stat = "bin",
color="black",
       fill="red")+facet_wrap(~haberman$dummy_surv)+theme_get()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
skewness(haberman$Age)

## [1] 0.1582031

kurtosis(haberman$Age)

## [1] 2.394681

boxplot(Age~dummy_surv, data = haberman)
```




```r
hist(haberman$Age, freq = FALSE)
lines(density(haberman$Age),col="red",lwd=1)
```



```
shapiro.test(haberman$Age)

##
##  Shapiro-Wilk normality test
##
## data:  haberman$Age
## W = 0.98898, p-value = 0.02072

wilcox.test(Age~dummy_surv, mu=0, alternative="two.sided",
            var.equal=F, conf.level=0.95, data = haberman)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Age by dummy_surv
## W = 9698, p-value = 0.3576
## alternative hypothesis: true location shift is not equal to 0
```

In data the **Min. age is 33 years** and **Max. is 83 years**. the mean of the age is 52.53 The number of dead patients is 81 patients which mean of their age is 53.68 and standard deviation (sd) is 10.17 The number of survived patients is 224 patients which mean of their age is 52.12 and sd is 10.94.

Check normality:

skewness test(0.16) is in accepted range (-1 to +1)

kurtosis test(2.39) is in accepted range (-2 to +3)

So numerically the data is normally distributed

Shapiro test(0.02) is **less** than **0.05** so **statistically the data is not normally distributed. So we will apply wilcox rank test.**

Null hypothesis(H0): mean of age of survived patients = mean of age of dead patients

in wilcox rank test p-value (0.358) is more than 0.05. so we **fail to reject null hypothesis**. there is **no significant differience** between age and survived status.

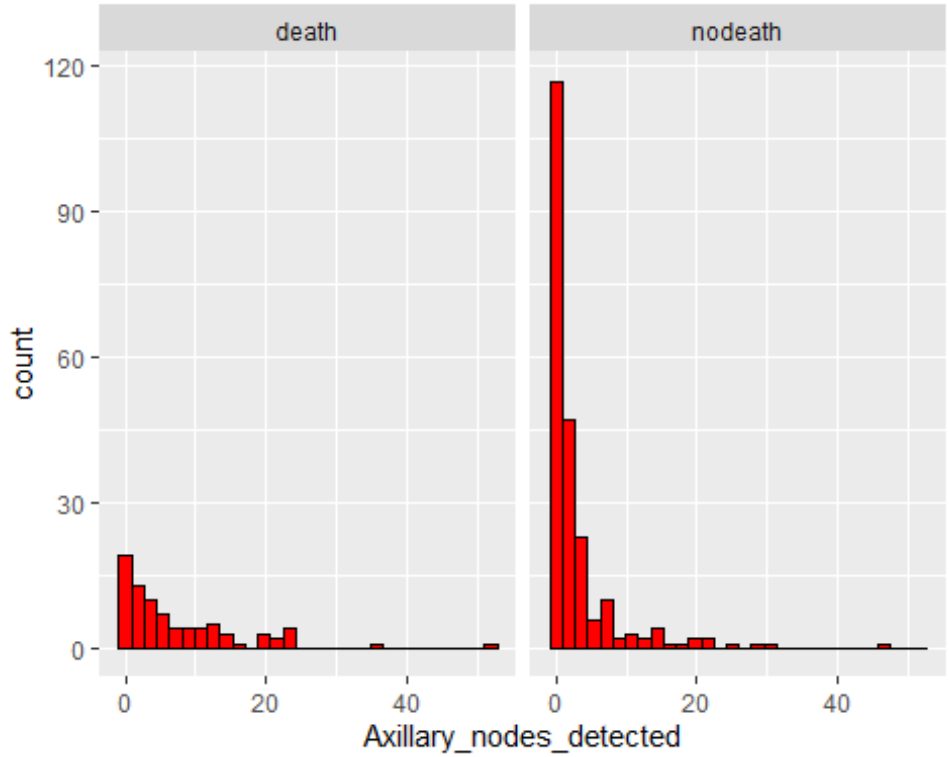
2-Relation between Axillary nodes detected and survived status:

```
summary(haberman$Axillary_nodes_detected)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   4.036   4.000  52.000

axilcoun <- group_by(haberman, dummy_surv) %>%
  summarise(count=n(), mean=mean(haberman$Axillary_nodes_detected),
            sd=sd(haberman$Axillary_nodes_detected), var=
var(haberman$Axillary_nodes_detected))
axilcoun
```

```
## # A tibble: 2 x 5
##   dummy_surv count  mean    sd   var
##   <chr>      <int> <dbl> <dbl> <dbl>
## 1 death         81  4.04  7.20  51.8
## 2 nodeath      224  4.04  7.20  51.8
```

```
ggplot(haberman, aes(Axillary_nodes_detected))+geom_histogram(stat = "bin",
                  color="black",
fill="red")+facet_wrap(~dummy_surv)+theme_get()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



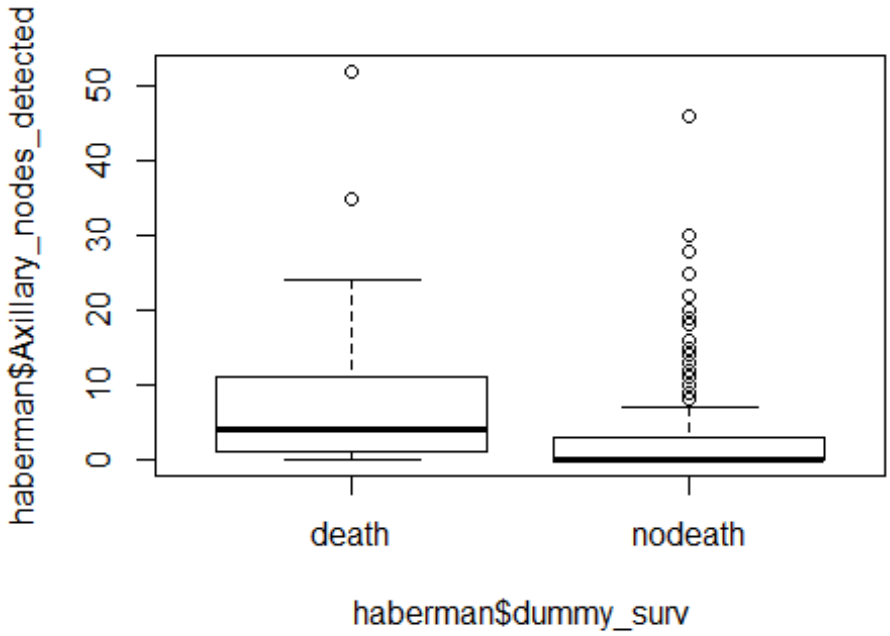

```
skewness(haberman$Axillary_nodes_detected)
```

```
## [1] 2.963017
```

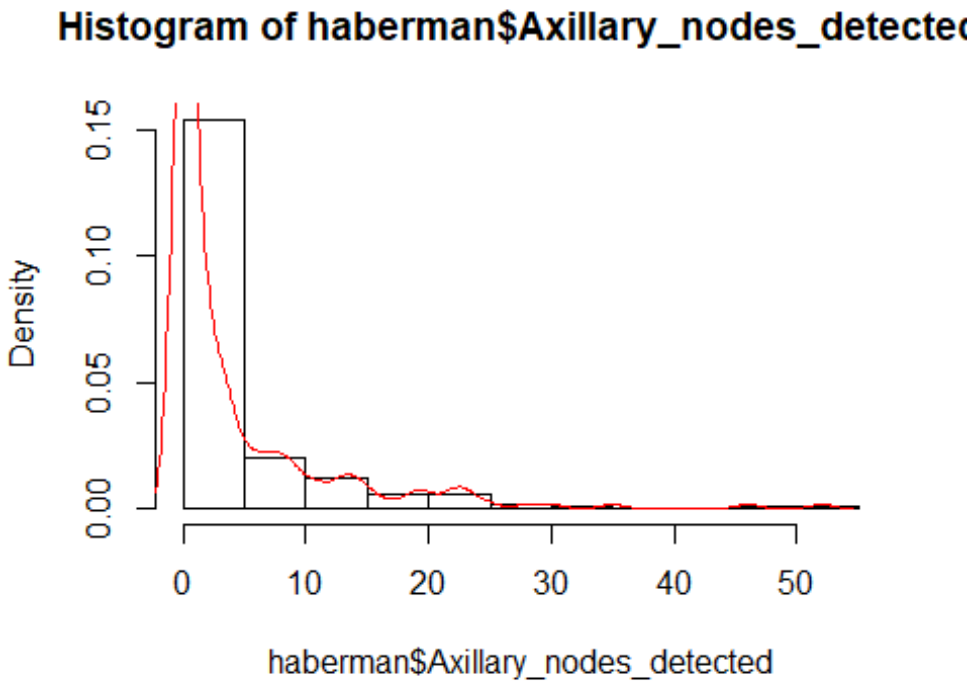
```
kurtosis(haberman$Axillary_nodes_detected)
```

```
## [1] 14.47348
```

```
boxplot(haberman$Axillary_nodes_detected~haberman$dummy_surv)
```




```r
hist(haberman$Axillary_nodes_detected, freq = FALSE)
lines(density(haberman$Axillary_nodes_detected), col="red", lwd=1)
```



```
shapiro.test(haberman$Axillary_nodes_detected)

##
##  Shapiro-Wilk normality test
##
## data:  haberman$Axillary_nodes_detected
## W = 0.61607, p-value < 2.2e-16

wilcox.test(Axillary_nodes_detected~dummy_surv, mu=0,
alternative="two.sided",
            var.equal=T, conf.level=0.95, data = haberman)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Axillary_nodes_detected by dummy_surv
## W = 12774, p-value = 1.138e-08
## alternative hypothesis: true location shift is not equal to 0
```

In data **the Min. number of axillary nodes detected (AND) is Zero** and **Max. number of AND id 52**. The mean is 4.036 The number of dead patients is 81 patients which mean of number of AND is 4.036 and sd is 7.199. The number of survived patients is 224 patients which mean of number AND is 4.036 and sd is 7.199.

Check normailty:

skewness test(2.96) is out of acceptable range.

kurtosis test(14.47) is out of acceptable range.

so numerically data is not normally distributed.

Shapiro test p-value(2.2e-16) is **less** than 0.05. so statistically the data is **not normally distributed.**

Visually the data is **not normally** distributed.

So we will apply wilcox rank test

Null hypothesis(H0): mean of number axillary nodes detected in survived patients = mean of number axillary nodes detected in dead patients

in wilcox test p-value(1.138e-08) is less than 0.05. so we **reject null hypothesis**. there is **significant differience** between number of AND and survived status.

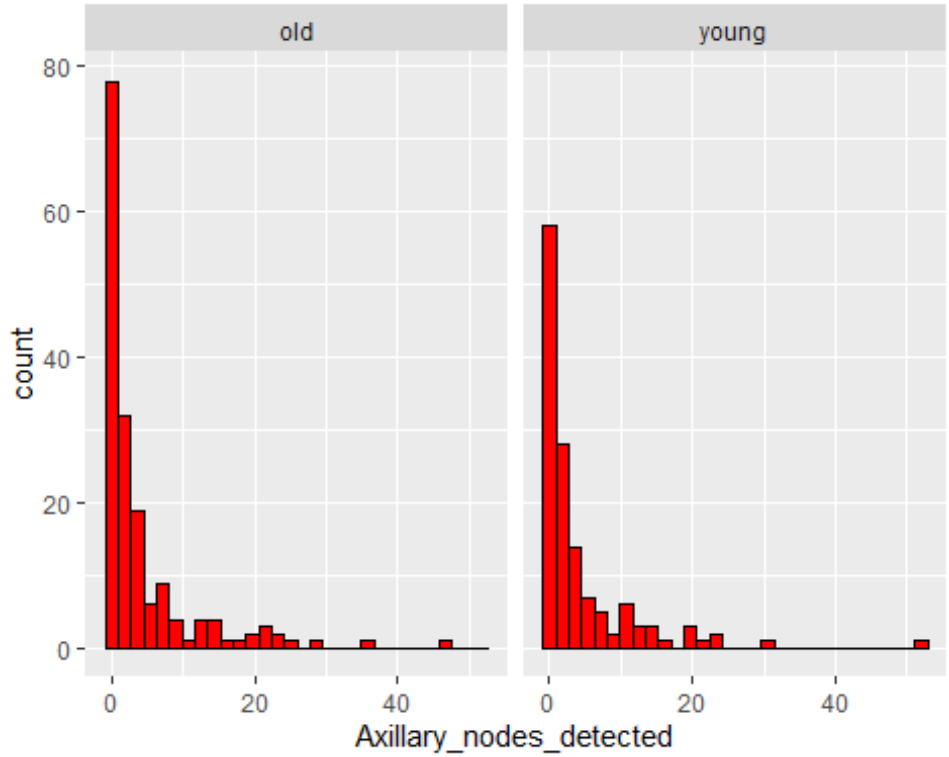
Relation between dummy_age and Axillary_nodes_detected:

```
ageper <- group_by(haberman, dummy_age) %>%
    summarise(count=n(), mean=mean(Axillary_nodes_detected),
sd=sd(Axillary_nodes_detected), var=var(Axillary_nodes_detected))
ageper
```

```
## # A tibble: 2 x 5
##   dummy_age count  mean    sd   var
##   <chr>     <int> <dbl> <dbl> <dbl>
## 1 old         170  3.99  7.22  52.1
## 2 young       135  4.09  7.20  51.8
```

```
ggplot(haberman,aes(Axillary_nodes_detected, fill=
Axillary_nodes_detected))+geom_histogram(stat ="bin", color="black",
fill="red")+facet_wrap(~dummy_age)+theme_get()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
wilcox.test(Axillary_nodes_detected~dummy_age, mu=0,
alternative="two.sided",var.equal=T, conf.level=0.95, data = haberman)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Axillary_nodes_detected by dummy_age
## W = 11110, p-value = 0.6176
## alternative hypothesis: true location shift is not equal to 0
```

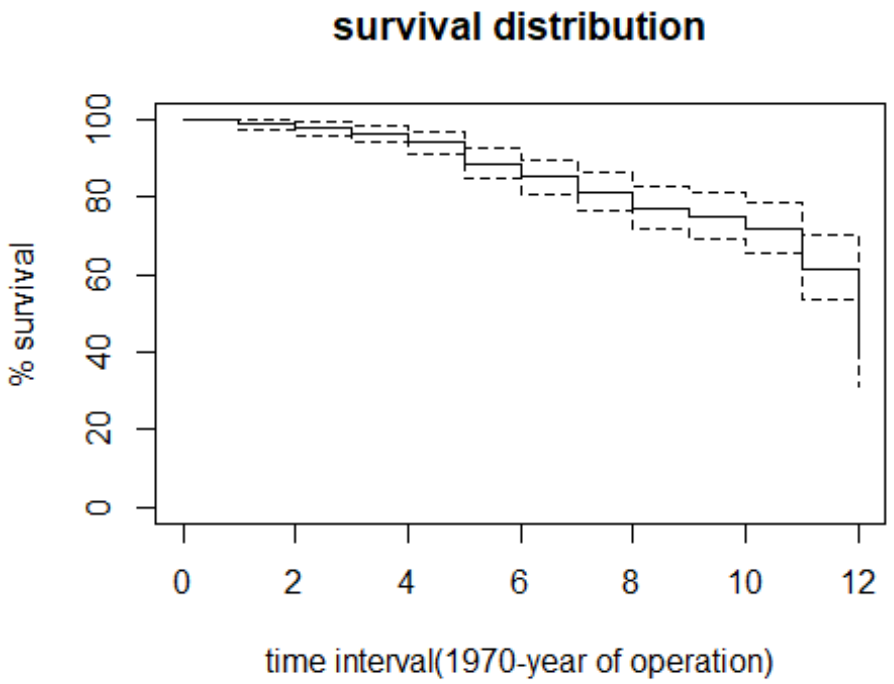
In data the number of old patients who are more than 55 years old are 170 patients and the mean of AND is 3.99 and sd is 7.22. the number of young patients who are less than 55 years old are 135 patients and the mean of AND 4.09 and sd is 7.20.

we have already checked normality of AND . it is **not normally distributed**.

In wilcox test p-value(0.618) is more than 0.05. so we **fail to reject null hypothesis**(mean of AND in young patients= mean of AND in old patients). there is **no significant differience.**

Kaplien mier analysis(Survival analysis):

```
survobj <- with(haberman, Surv(time2, surv_status))
fit0 <- survfit(survobj~1, data = haberman)
summary(fit0)

## Call: survfit(formula = survobj ~ 1, data = haberman)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    305       4    0.987 0.00651        0.974        1.000
##     2    294       3    0.977 0.00866        0.960        0.994
##     3    281       4    0.963 0.01098        0.942        0.985
##     4    256       6    0.940 0.01407        0.913        0.968
##     5    228      13    0.887 0.01961        0.849        0.926
##     6    200       8    0.851 0.02248        0.808        0.896
##     7    170       8    0.811 0.02550        0.763        0.863
##     8    140       7    0.771 0.02846        0.717        0.828
##     9    117       3    0.751 0.02993        0.694        0.812
##    10     91       4    0.718 0.03285        0.656        0.785
##    11     63       9    0.615 0.04236        0.538        0.704
##    12     36      12    0.410 0.05599        0.314        0.536

plot(fit0, xlab = "time interval(1970-year of operation)", ylab = "%
survival",
     yscale = 100, main="survival distribution")
```



As we see in summary :the clinical trial started with 305 patients in the end of trials found that **81 dead** patients and **224 survived** patients.

Survival analysis between young and old patients:

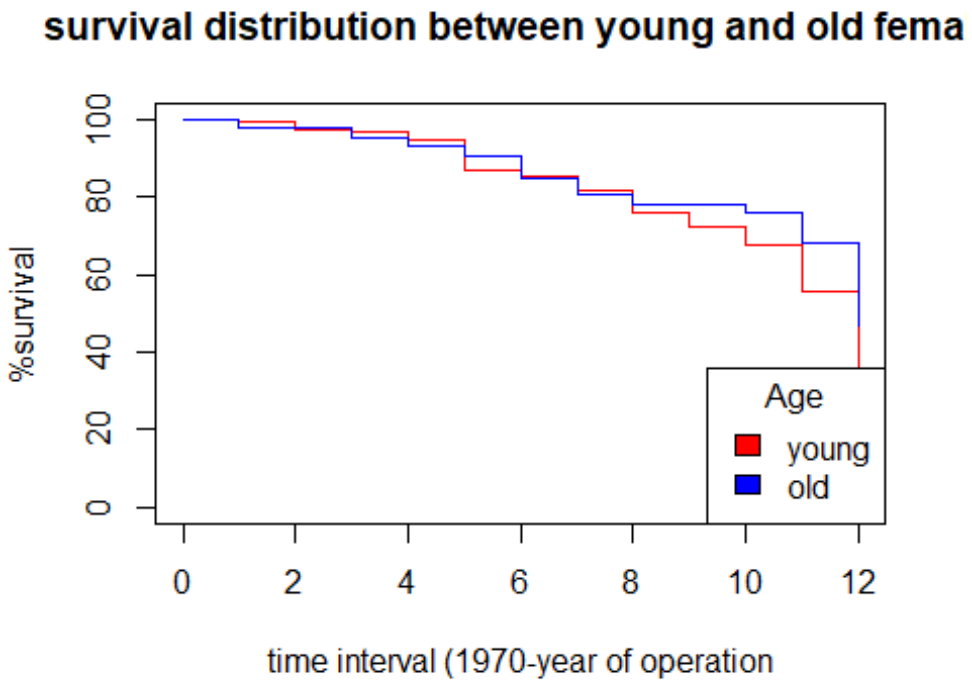
```
fitage <- survfit(survobj~dummy_age, data = haberman)
summary(fitage)

## Call: survfit(formula = survobj ~ dummy_age, data = haberman)
##
##                 dummy_age=old
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    170       1    0.994 0.00587        0.983        1.000
##     2    164       3    0.976 0.01189        0.953        1.000
##     3    153       1    0.970 0.01341        0.944        0.996
##     4    139       3    0.949 0.01775        0.914        0.984
##     5    122      10    0.871 0.02865        0.816        0.929
##     6    105       2    0.854 0.03041        0.797        0.916
##     7     89       4    0.816 0.03457        0.751        0.887
##     8     76       5    0.762 0.03977        0.688        0.844
##     9     62       3    0.725 0.04317        0.645        0.815
##    10     47       3    0.679 0.04798        0.591        0.780
##    11     34       6    0.559 0.05943        0.454        0.689
##    12     20       7    0.363 0.07106        0.248        0.533
##
##                 dummy_age=young
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

```
##     1    135       3    0.978  0.0127        0.953        1.000
##     3    128       3    0.955  0.0180        0.920        0.991
##     4    117       3    0.930  0.0224        0.887        0.975
##     5    106       3    0.904  0.0264        0.854        0.957
##     6     95       6    0.847  0.0335        0.784        0.915
##     7     81       4    0.805  0.0378        0.734        0.883
##     8     64       2    0.780  0.0406        0.704        0.864
##    10     44       1    0.762  0.0434        0.682        0.852
##    11     29       3    0.683  0.0581        0.579        0.807
##    12     16       5    0.470  0.0887        0.325        0.680
```

```
survdiff(survobj~dummy_age, data = haberman)
```

```
## Call:
## survdiff(formula = survobj ~ dummy_age, data = haberman)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## dummy_age=old   170       48     43.7     0.431      1.02
## dummy_age=young 135       33     37.3     0.504      1.02
##
##  Chisq= 1  on 1 degrees of freedom, p= 0.3
```

```
plot(fitage, xlab = "time interval (1970-year of operation", ylab =
"%survival",
     yscale =100 ,col=c("red","blue"), main="survival distribution between
young and old female")
legend("bottomright",title = "Age", c("young", "old"), fill =
c("red","blue"))
```



In survival distribution between young and old females , we found that number of **old** patients (more than 55 years old) is **170** and the number of **dead** patients is **48**. so **percentage of their survival** is **40%** and the **percentage of dead old** patients is **15.73%**. the number of **young** patients is **135** patients and the **dead** patients are **33** patients. so the **percentage of their survival** is **33.44%** and **the percentage of dead young patients** is **10.82%.**

p-value (0.3) is more than 0.05 so median time between two groups of age is not significant.

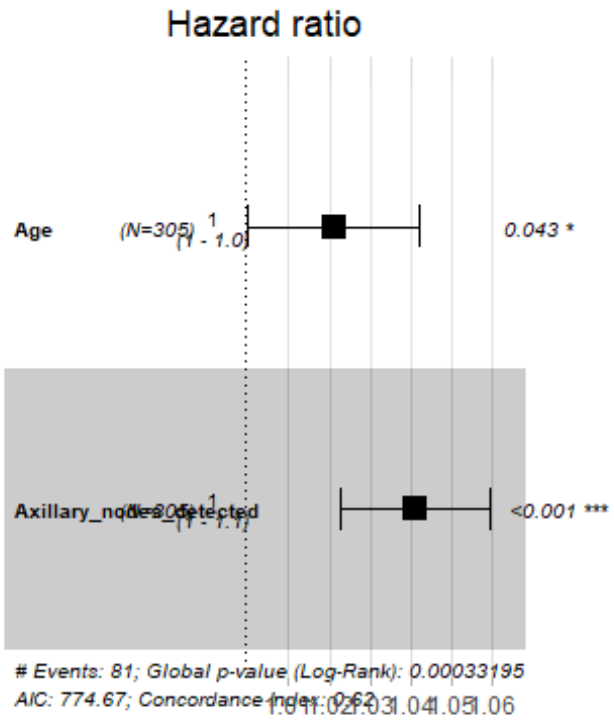
Cox regression:

Reporting hazard ratio:

```
mod1 <- coxph(survobj~Age+Axillary_nodes_detected, data = haberman)
summary(mod1)

## Call:
## coxph(formula = survobj ~ Age + Axillary_nodes_detected, data = haberman)
##
##   n= 305, number of events= 81
##
##                              coef exp(coef) se(coef)     z Pr(>|z|)
## Age                      0.020896  1.021116 0.010340 2.021   0.0433 *
## Axillary_nodes_detected 0.040055  1.040868 0.009112 4.396  1.1e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                          exp(coef) exp(-coef) lower .95 upper .95
## Age                          1.021     0.9793     1.001     1.042
## Axillary_nodes_detected      1.041     0.9607     1.022     1.060
##
## Concordance= 0.615  (se = 0.038 )
## Likelihood ratio test= 16.02  on 2 df,   p=3e-04
## Wald test            = 20.26  on 2 df,   p=4e-05
## Score (logrank) test = 21.7  on 2 df,   p=2e-05
```

```
ggforest(mod1, data = haberman)
```



1- <u>Age :</u>

- as **age increases** the hazard ratio(**risk of death**) **increases**(coeff is postive value)
- the effect size of age as a covariate(hazard ratio) is 1.02(exp(coeff)
- p value is at the margin of statistical significant 0.05 so there is **significant**.

2- <u>Axillary nodes:</u>

- as the **axillary nodes detected increase** the **hazard ratio increases**(postive coeff)

- the effect size of axillary nodes detected as a covarite is 1.04
- p-value is less than 0.05 so there is **highly significant differience**.

let's make the predication risk model:

```
mod2 <- coxph(survobj~Axillary_nodes_detected, data = haberman)
summary(mod2)
```

```
## Call:
## coxph(formula = survobj ~ Axillary_nodes_detected, data = haberman)
##
##   n= 305, number of events= 81
##
##                              coef exp(coef) se(coef)     z Pr(>|z|)
## Axillary_nodes_detected 0.035858  1.036509 0.008862 4.046  5.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## Axillary_nodes_detected     1.037     0.9648     1.019     1.055
##
## Concordance= 0.655  (se = 0.033 )
## Likelihood ratio test= 11.96  on 1 df,   p=5e-04
## Wald test            = 16.37  on 1 df,   p=5e-05
## Score (logrank) test = 17.61  on 1 df,   p=3e-05
```

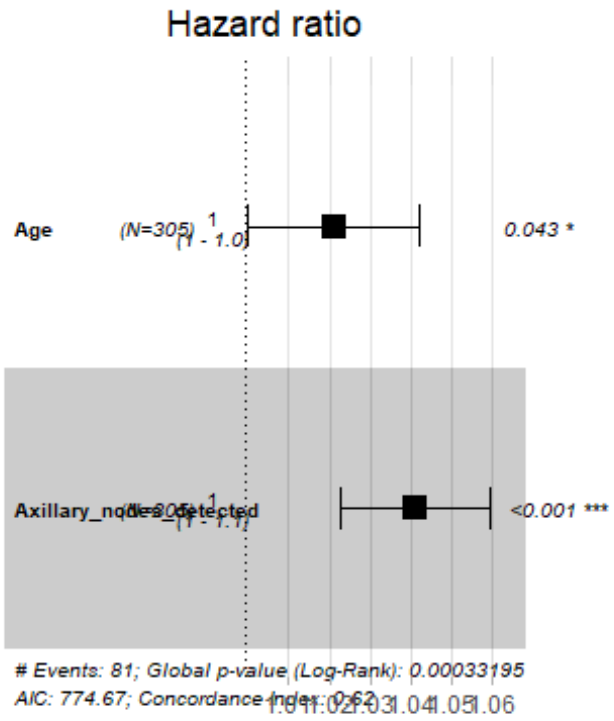
the axillary nodes detected is significant with hazard of death considering other value of the data **the p value of likelihood ratio test** is **decreased** and **highly significant** which improve that **the risk of death increases by number of axillary nodes increases**.

Evaluation of propational hazard asumption:

```
cox.zph(mod1)
```

```
##                          chisq df    p
## Age                     2.6500  1 0.10
## Axillary_nodes_detected 0.0434  1 0.83
## GLOBAL                  3.2165  2 0.20
```

```
ggforest(mod1, data = haberman)
```



In this model all **p-values** of age , axillary nodes detected and global are more than 0.05 so they are **not siginificant**. which means that all factors don't change by time. so this is **an acceptable model**.