Survival analysis of GBSG2 dataset in R

Amira Ibrahim

January 9, 2021

Loading Libraries

```
## Loading required package: survival
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##     geyser
## Loading required package: ggplot2
## Loading required package: ggpubr
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Explore Data :

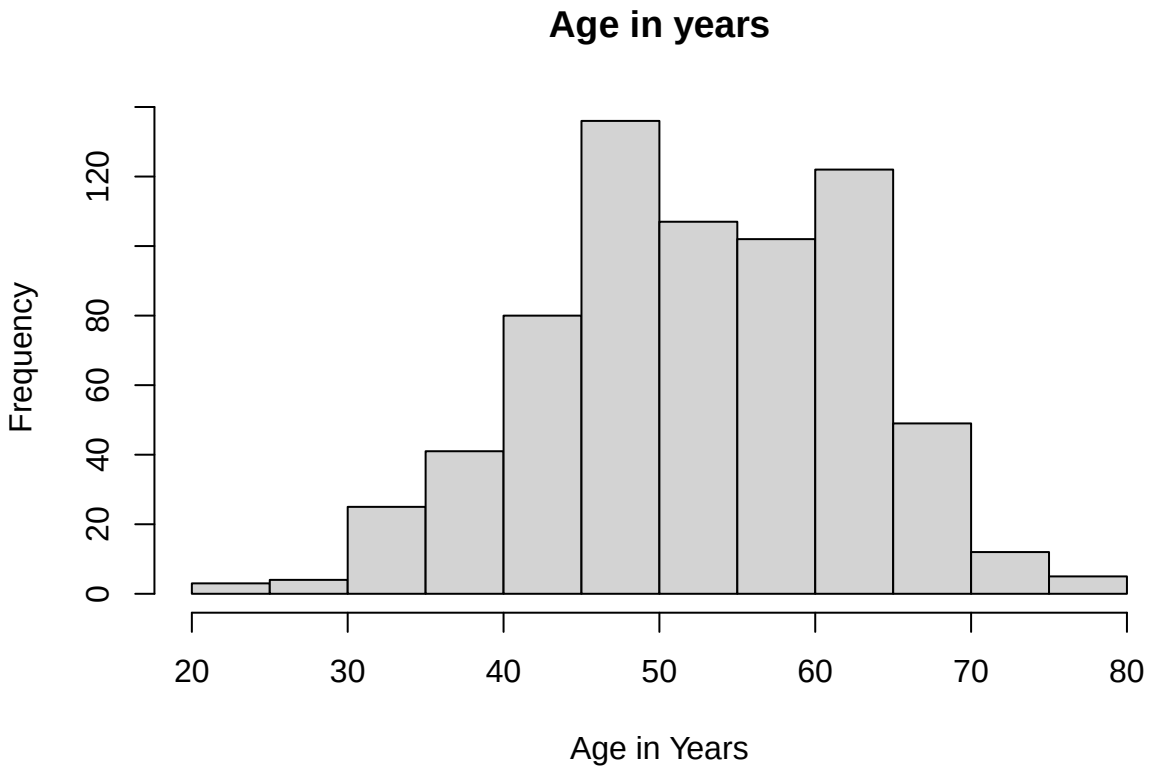
```
##   horTh age menostat tsize tgrade pnodes progrec estrec time cens
## 1    no  70     Post    21     II      3      48     66 1814    1
## 2   yes  56     Post    12     II      7      61     77 2018    1
## 3   yes  58     Post    35     II      9      52    271  712    1
## 4   yes  59     Post    17     II      4      60     29 1807    1
## 5    no  73     Post    35     II      1      26     65  772    1
## 6    no  32      Pre    57    III     24       0     13  448    1
## Rows: 686
## Columns: 10
## $ horTh    <fct> no, yes, yes, yes, no, no, yes, no, no, no, yes, yes, yes,...
## $ age      <int> 70, 56, 58, 59, 73, 32, 59, 65, 80, 66, 68, 71, 59, 50, 70...
```

```
## $ menostat <fct> Post, Post, Post, Post, Post, Pre, Post, Post, Post, Post,...
## $ tsize    <int> 21, 12, 35, 17, 35, 57, 8, 16, 39, 18, 40, 21, 58, 27, 22,...
## $ tgrade   <ord> II, II, II, II, II, III, II, II, II, II, II, II, II, III, ...
## $ pnodes   <int> 3, 7, 9, 4, 1, 24, 2, 1, 30, 7, 9, 9, 1, 1, 3, 1, 4, 1, 1,...
## $ progrec  <int> 48, 61, 52, 60, 26, 0, 181, 192, 0, 0, 16, 0, 154, 16, 113...
## $ estrec   <int> 66, 77, 271, 29, 65, 13, 0, 25, 59, 3, 20, 0, 101, 12, 139...
## $ time     <int> 1814, 2018, 712, 1807, 772, 448, 2172, 2161, 471, 2014, 57...
## $ cens     <int> 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1...
```

summarize data :

```
##  horTh          age        menostat       tsize        tgrade
##  no :440   Min.   :21.00   Pre :290   Min.   :  3.00   I  : 81
##  yes:246   1st Qu.:46.00   Post:396   1st Qu.: 20.00   II :444
##            Median :53.00              Median : 25.00   III:161
##            Mean   :53.05              Mean   : 29.33
##            3rd Qu.:61.00              3rd Qu.: 35.00
##            Max.   :80.00              Max.   :120.00
##      pnodes         progrec           estrec             time
##  Min.   : 1.00   Min.   :   0.0   Min.   :   0.00   Min.   :   8.0
##  1st Qu.: 1.00   1st Qu.:   7.0   1st Qu.:   8.00   1st Qu.: 567.8
##  Median : 3.00   Median :  32.5   Median :  36.00   Median :1084.0
##  Mean   : 5.01   Mean   : 110.0   Mean   :  96.25   Mean   :1124.5
##  3rd Qu.: 7.00   3rd Qu.: 131.8   3rd Qu.: 114.00   3rd Qu.:1684.8
##  Max.   :51.00   Max.   :2380.0   Max.   :1144.00   Max.   :2659.0
##       cens
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4359
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Age :



Hormonal Theraby :

```
## horTh
##  no yes Sum
## 440 246 686
```

```
## horTh
##   no  yes
## 64.1 35.9
```

35.9 % of patients are taking hormonal theraby .

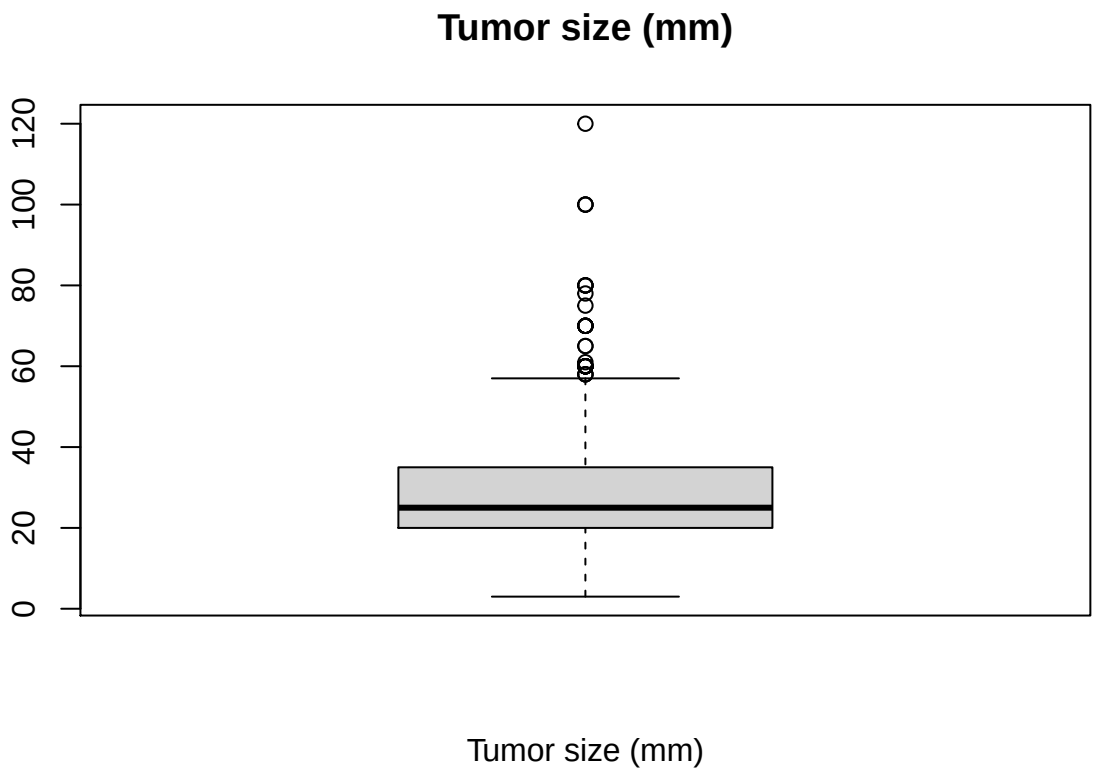
Menopausal status :

```
## menostat
##  Pre Post  Sum
##  290  396  686
```

```
## menostat
##  Pre Post
## 42.3 57.7
```

57.7 % of patients are considered to be postmenopausal.

Tumor size :



Tumor grade :

```
## tgrade
##   I  II III Sum
##  81 444 161 686
```

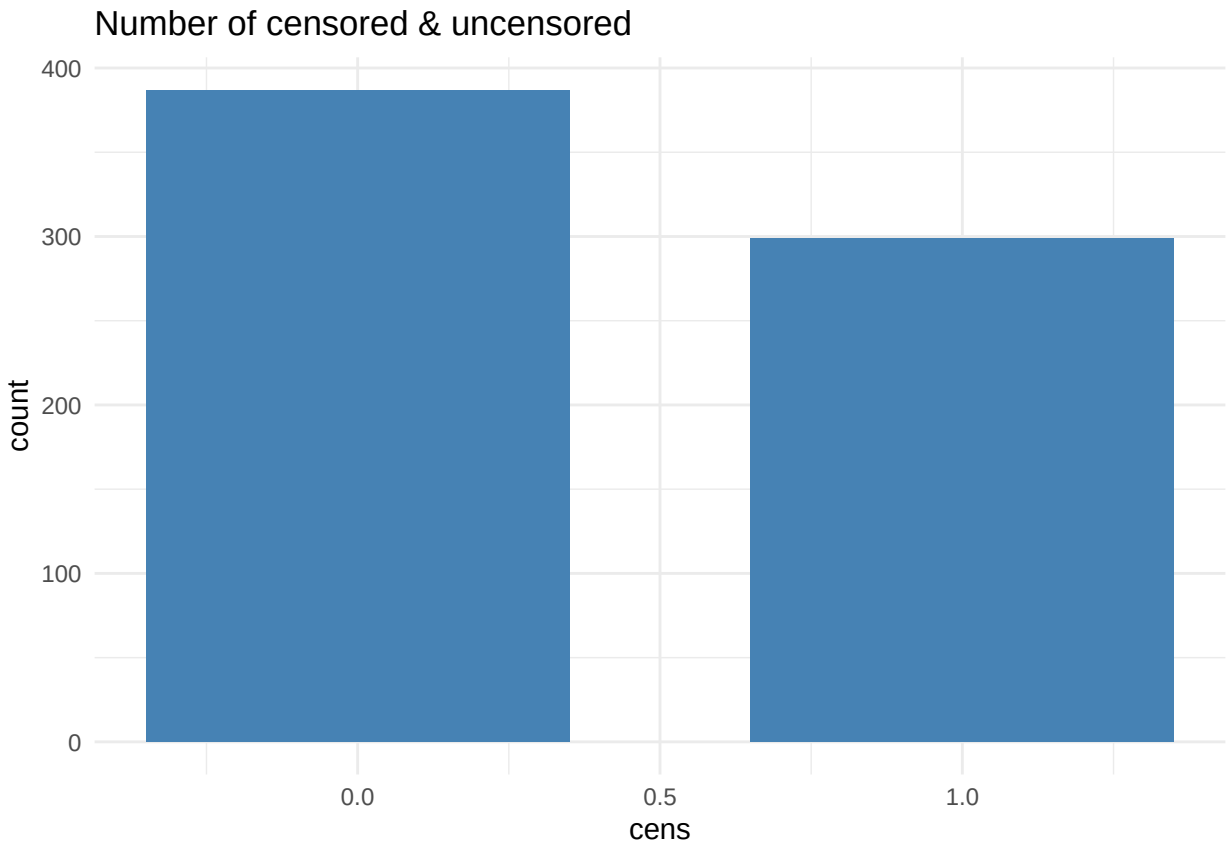
```
## tgrade
##    I   II  III
## 11.8 64.7 23.5
```

11.8% of patients have grade I tumor
64.7 % of patients have grade II tumor
23.5 % of patients have grade III tumor

Count censored and uncensored data

```
##
##   0   1
## 387 299
```

Create barplot of censored and uncensored data



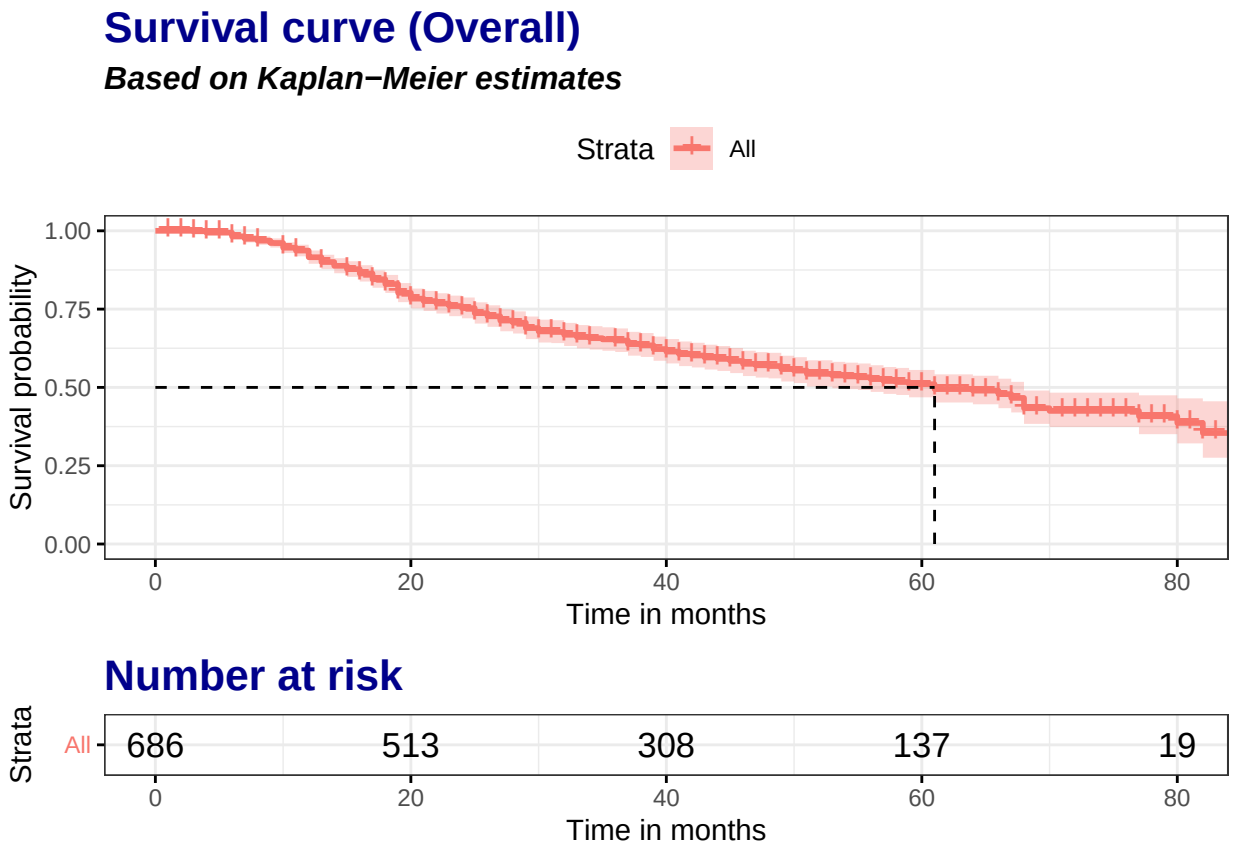
Convert time into months :

Create Surv-Object

Survival distribution of the total sample :

Kaplan-Meier estimate

plot of the Kaplan-Meier estimate



Estimating median survival from a Weibull model :

Compute the median survival from the model

```
##        1
## 56.96077
```

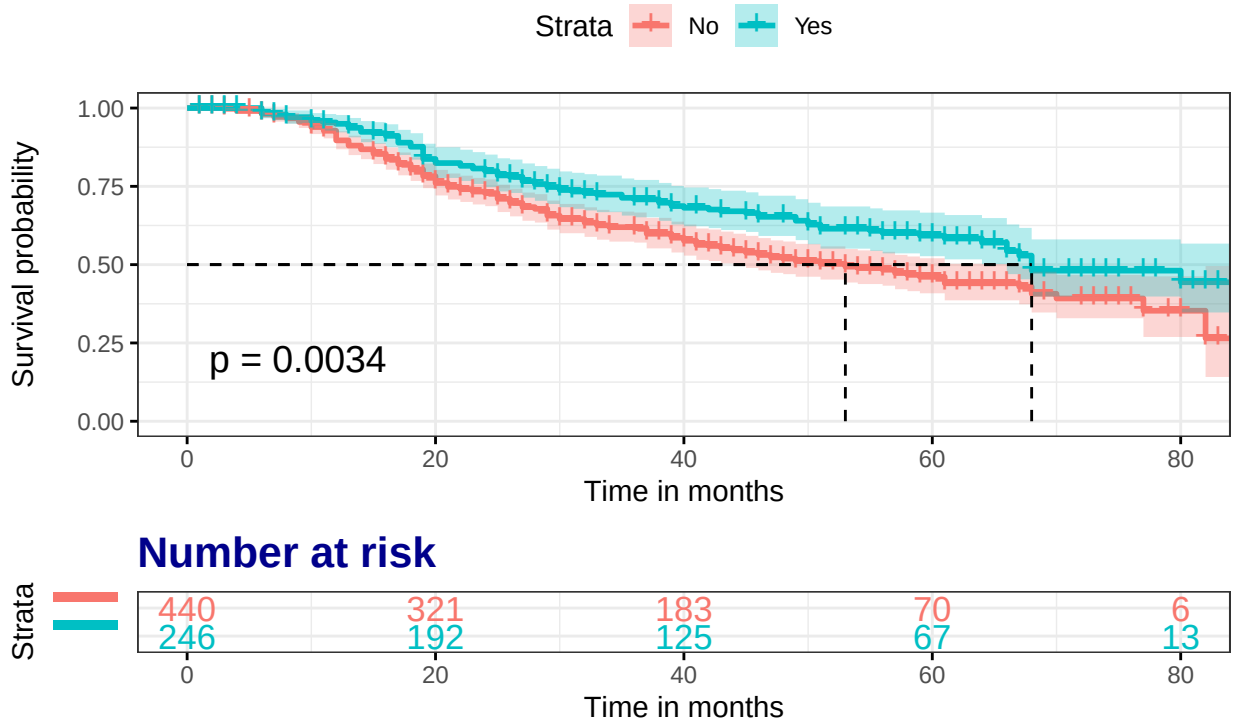
Half the patients live longer than 56.96 months and half die before.

Survival distribution by Hormonal theraby:

Kaplan-Meier estimate

plot of the Kaplan-Meier estimate

```
## Warning: Vectorized input to `element_text()` is not officially supported.
## Results may be unexpected or may change in future versions of ggplot2.
```




```
## Call: survfit(formula = sobj ~ horTh, data = Data)
##
##             n events median 0.95LCL 0.95UCL
## horTh=no  440    205     53      44      67
## horTh=yes 246     94     68      66      NA
```

median survival time for females who took hormonal therapy was 68 months median survival time for females who didn't take hormonal therapywas 53 months

Test for difference between Who take Hormonal theraby or Not :

#(logrank test)

```
## Call:
## survdiff(formula = sobj ~ horTh, data = Data, rho = 0)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## horTh=no  440      205      181      3.32      8.56
## horTh=yes 246       94      118      5.06      8.56
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

p value = 0.003 , There's significant difference between the females Who took hormonal theraby and who did not in their survival times.

Weibull model for imaginary patients(take hormonal therapy with definite tumor size):

Weibull model

Imaginary patients : "imaginary patients": the two levels of horTh and the 25%, 50%, and 75% quantiles of tsize.

```
##   horTh tsize
## 1    no    20
## 2   yes    20
## 3    no    25
## 4   yes    25
## 5    no    35
## 6   yes    35
```

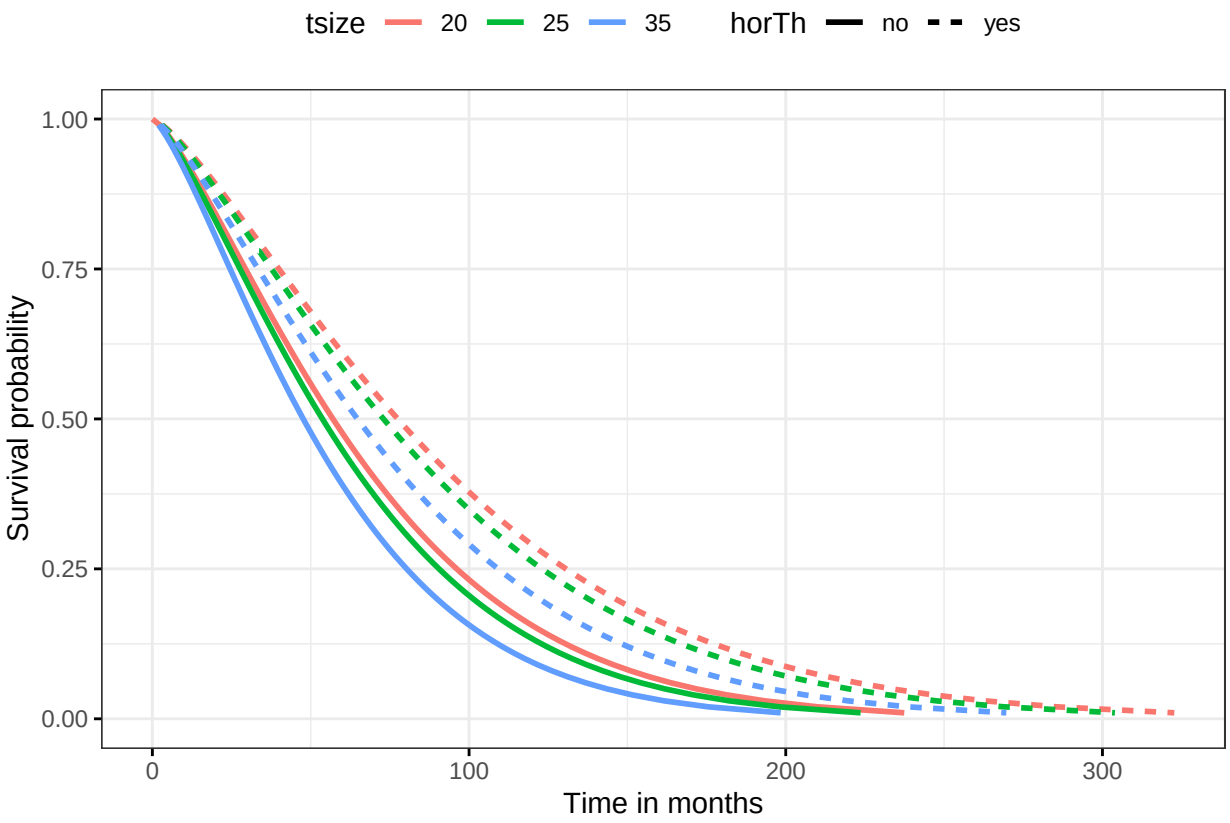
Compute survival curves

combine the information in newdat with t

bring the data.frame to long format

add the correct survival probabilities surv

Plot the survival curves



The visualization shows that patients with smaller tumors tend to survive longer and patients who receive hormonal therapy tend to survive longer.

```
##
## Call:
## survreg(formula = Surv(time_months, cens) ~ horTh + tsize, data = Data)
##                Value Std. Error     z       p
## (Intercept)  4.55872    0.10217 44.62 < 2e-16
## horThyes     0.30684    0.09428  3.25  0.0011
## tsize       -0.01197    0.00267 -4.48 7.5e-06
## Log(scale)  -0.28318    0.04943 -5.73 1.0e-08
##
## Scale= 0.753
##
## Weibull distribution
## Loglik(model)= -1608.5   Loglik(intercept only)= -1622.6
##  Chisq= 28.28 on 2 degrees of freedom, p= 7.2e-07
## Number of Newton-Raphson Iterations: 5
## n= 686
```

the probability of surviving falls by 0.011 per unit increase in the tumor size and increases by 0.307 if taking hormonal therapy.