Proportional Hazards Regression with R: Part Two[*]

The kidney data set has data on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored.

```
> rm(list=ls());  # options(scipen=999)
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival) # Do this every time
> 
> # help(kidney)
> head(kidney)
  id time status age sex disease frail
1  1    8      1  28   1   Other   2.3
2  1   16      1  28   1   Other   2.3
3  2   23      1  48   2      GN   1.9
4  2   13      0  48   2      GN   1.9
5  3   22      1  32   1   Other   1.2
6  3   28      1  32   1   Other   1.2
> summary(kidney)

       id             time            status            age      
 Min.   : 1.0   Min.   :  2.0   Min.   :0.0000   Min.   :10.0  
 1st Qu.:10.0   1st Qu.: 16.0   1st Qu.:1.0000   1st Qu.:34.0  
 Median :19.5   Median : 39.5   Median :1.0000   Median :45.5  
 Mean   :19.5   Mean   :101.6   Mean   :0.7632   Mean   :43.7  
 3rd Qu.:29.0   3rd Qu.:149.8   3rd Qu.:1.0000   3rd Qu.:54.0  
 Max.   :38.0   Max.   :562.0   Max.   :1.0000   Max.   :69.0  
      sex           disease       frail      
 Min.   :1.000   Other:26   Min.   :0.200  
 1st Qu.:1.000   GN   :18   1st Qu.:0.600  
 Median :2.000   AN   :24   Median :1.100  
 Mean   :1.737   PKD  : 8   Mean   :1.184  
 3rd Qu.:2.000              3rd Qu.:1.500  
 Max.   :2.000              Max.   :3.000  
> dim(kidney)
[1] 76  7

> table(kidney$disease)

Other    GN    AN   PKD 
   26    18    24     8 


> contrasts(kidney$disease)
      GN AN PKD
Other  0  0   0
GN     1  0   0
AN     0  1   0
PKD    0  0   1


> with(Kidney,cor(age,frail))
[1] 0.03876767
```

[*] Copyright information is on the last page.

```
> # Make a new data frame with
>     # 1=F, 0=M
>     # age and frailty centered
>
> Kidney = within(kidney,{
+ sex = sex-1 # Indicator for female
+ # Centering age and frailty
+ age = age-mean(age)
+ frail = frail-mean(frail)
+ })
> with(Kidney,cor(age,frail))
[1] 0.03876767

> kmod1 = coxph( Surv(time,status) ~ age + sex + disease + frail, data=Kidney)
> summary(kmod1)
Call:
coxph(formula = Surv(time, status) ~ age + sex + disease + frail,
    data = Kidney)

  n= 76, number of events= 58

                coef exp(coef)  se(coef)      z Pr(>|z|)
age         0.007714  1.007744  0.011907  0.648 0.517055
sex        -2.099844  0.122475  0.392654 -5.348 8.90e-08 ***
diseaseGN   0.130666  1.139587  0.436114  0.300 0.764471
diseaseAN   0.640906  1.898200  0.447886  1.431 0.152442
diseasePKD -2.168515  0.114347  0.648825 -3.342 0.000831 ***
frail       1.791873  6.000682  0.257639  6.955 3.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age           1.0077     0.9923   0.98450    1.0315
sex           0.1225     8.1649   0.05673    0.2644
diseaseGN     1.1396     0.8775   0.48476    2.6790
diseaseAN     1.8982     0.5268   0.78904    4.5665
diseasePKD    0.1143     8.7453   0.03206    0.4079
frail         6.0007     0.1666   3.62158    9.9427

Concordance= 0.822  (se = 0.03 )
Likelihood ratio test= 68.71  on 6 df,   p=8e-13
Wald test            = 60.01  on 6 df,   p=4e-11
Score (logrank) test = 86.24  on 6 df,   p=<2e-16

> # Are se(coef) labelled correctly?
> se = sqrt(diag(vcov(kmod1))); se # Yes
       age        sex  diseaseGN  diseaseAN diseasePKD      frail
0.01190684 0.39265430 0.43611383 0.44788559 0.64882505 0.25763894
>
> # CI for the hazard ratio exp(beta1)
> betahat = coef(kmod1); betahat
        age         sex   diseaseGN   diseaseAN  diseasePKD       frail
 0.00771434 -2.09984449  0.13066624  0.64090624 -2.16851476  1.79187311
> CIbeta1 = c(betahat[1]-1.96*se[1], betahat[1]+1.96*se[1]); CIbeta1
        age         age
-0.01562307  0.03105175
> exp(CIbeta1)
      age       age
0.9844983 1.0315389
> # So summary is giving us confidence intervals for the hazard ratios,
> # not the coefficients.
> summary(kmod1)
Call:
coxph(formula = Surv(time, status) ~ age + sex + disease + frail,
    data = Kidney)

  n= 76, number of events= 58

                coef exp(coef)  se(coef)      z Pr(>|z|)
age         0.007714  1.007744  0.011907  0.648 0.517055
sex        -2.099844  0.122475  0.392654 -5.348 8.90e-08 ***
diseaseGN   0.130666  1.139587  0.436114  0.300 0.764471
diseaseAN   0.640906  1.898200  0.447886  1.431 0.152442
diseasePKD -2.168515  0.114347  0.648825 -3.342 0.000831 ***
frail       1.791873  6.000682  0.257639  6.955 3.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age           1.0077     0.9923   0.98450    1.0315
sex           0.1225     8.1649   0.05673    0.2644
diseaseGN     1.1396     0.8775   0.48476    2.6790
diseaseAN     1.8982     0.5268   0.78904    4.5665
diseasePKD    0.1143     8.7453   0.03206    0.4079
frail         6.0007     0.1666   3.62158    9.9427

Concordance= 0.822  (se = 0.03 )
Likelihood ratio test= 68.71  on 6 df,   p=8e-13
Wald test            = 60.01  on 6 df,   p=4e-11
Score (logrank) test = 86.24  on 6 df,   p=<2e-16

>
> # Estimated hazard of infection is ____ times as great for women as men.
> # Estimated hazard of infection is ____ times as great for disease type AN as it
is for Other.
> # Estimated hazard of infection is ____ times as great for disease type AN as it
is for disease type PKD.
> betahat = coef(kmod1); betahat
       age         sex   diseaseGN   diseaseAN  diseasePKD       frail
 0.00771434 -2.09984449  0.13066624  0.64090624 -2.16851476  1.79187311

> exp(betahat[4]-betahat[5]) # Hazard ratio of AN/PKD
diseaseAN
  16.6003
```

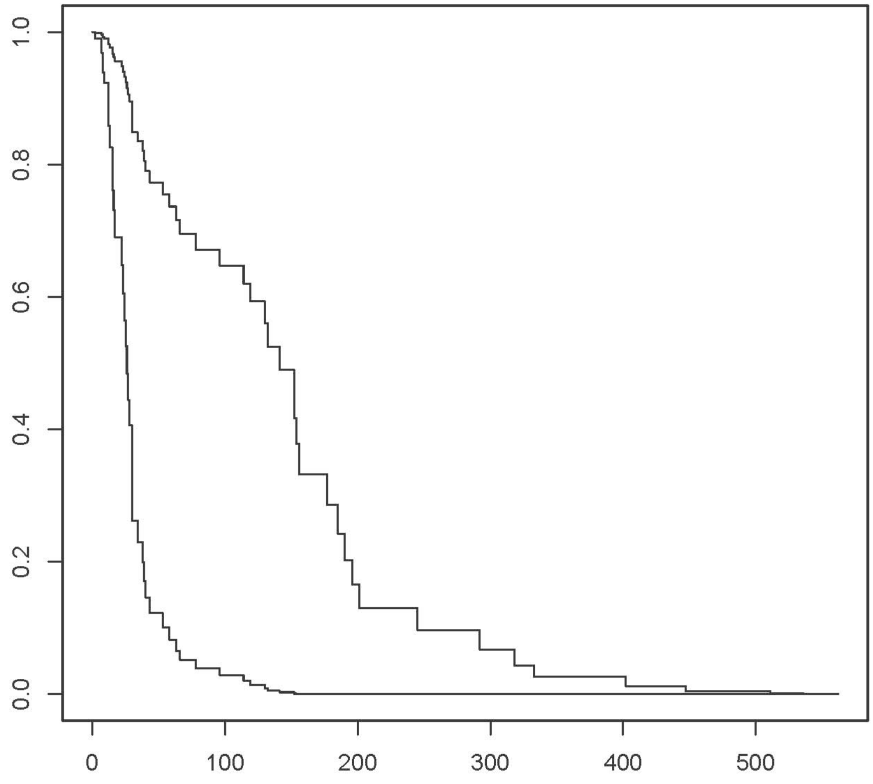
```
> # Test disease type with  a partial likeihood ratio test
> k2 = coxph( Surv(time,status) ~ age + sex + frail, data=Kidney)
> anova(k2,kmod1)

Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ age + sex + frail
 Model 2: ~ age + sex + disease + frail
   loglik Chisq Df Pr(>|Chi|)
1 -167.51
2 -153.55 27.93  3  3.756e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # Wald test: function(L,Tn,Vn,h=0) # H0: L theta = h
> Vn_hat = vcov(kmod1); betahat
        age         sex   diseaseGN   diseaseAN  diseasePKD       frail
 0.00771434 -2.09984449  0.13066624  0.64090624 -2.16851476  1.79187311
> LL = rbind(c(0,0,1,0,0,0),
+            c(0,0,0,1,0,0),
+            c(0,0,0,0,1,0)  )
> round( Wtest(LL,betahat,Vn_hat), 5)
       W       df  p-value
20.67302  3.00000  0.00012



> # Comparing survival functions for males and females
> guy = data.frame(age=0, sex=0, disease="Other", frail=0) # An average guy
> gal = data.frame(age=0, sex=1, disease="Other", frail=0) # An average gal
> sexcomp = rbind(guy,gal); sexcomp

  age sex disease frail
1   0   0   Other     0
2   0   1   Other     0

> rownames(sexcomp) = c("M","F"); sexcomp
  age sex disease frail
M   0   0   Other     0
F   0   1   Other     0
>
> s1 = survfit(kmod1,newdata=sexcomp)
> s1
Call: survfit(formula = kmod1, newdata = sexcomp)

   n events median 0.95LCL 0.95UCL
M 76     58     26      22      NA
F 76     58    141      96     245
```

```
> s1 = survfit(kmod1,newdata=sexcomp)
> s1
Call: survfit(formula = kmod1, newdata = sexcomp)

   n events median 0.95LCL 0.95UCL
M 76     58     26      22      NA
F 76     58    141      96     245


> ls(s1)
 [1] "call"      "conf.int"  "conf.type" "cumhaz"    "logse"     "lower"
 [7] "n"         "n.censor"  "n.event"   "n.risk"    "std.chaz"  "std.err"
[13] "surv"      "time"      "upper"

> summary(s1)
Call: survfit(formula = kmod1, newdata = sexcomp)

 time n.risk n.event survival1 survival2
    2     76       1  9.90e-01  9.99e-01
    7     71       2  9.68e-01  9.96e-01
    8     69       2  9.39e-01  9.92e-01
    9     65       1  9.23e-01  9.90e-01
   12     64       2  8.58e-01  9.81e-01
   13     62       1  8.26e-01  9.77e-01
   15     60       2  7.61e-01  9.67e-01


  185     13       1  9.45e-06  2.42e-01
  190     12       1  2.15e-06  2.02e-01
  196     11       1  4.26e-07  1.66e-01
  201     10       1  5.72e-08  1.30e-01
  245      9       1  5.35e-09  9.70e-02
  292      8       1  2.74e-10  6.74e-02
  318      7       1  7.23e-12  4.32e-02
  333      6       1  1.17e-13  2.61e-02
  402      5       1  2.30e-16  1.22e-02
  447      4       1  6.26e-20  4.45e-03
  511      3       1  8.53e-25  1.13e-03
  536      2       1  1.53e-34  7.22e-05
  562      1       1  3.86e-56  1.63e-07
>
> head(s1$cumhaz)
            [,1]        [,2]
[1,] 0.009766264 0.001196128
[2,] 0.009766264 0.001196128
[3,] 0.009766264 0.001196128
[4,] 0.009766264 0.001196128
[5,] 0.032125563 0.003934594
[6,] 0.062982652 0.007713830

>
> head(s1$surv)
             M         F
[1,] 0.9902813 0.9988046
[2,] 0.9902813 0.9988046
[3,] 0.9902813 0.9988046
[4,] 0.9902813 0.9988046
[5,] 0.9683850 0.9960731
[6,] 0.9389598 0.9923158
```

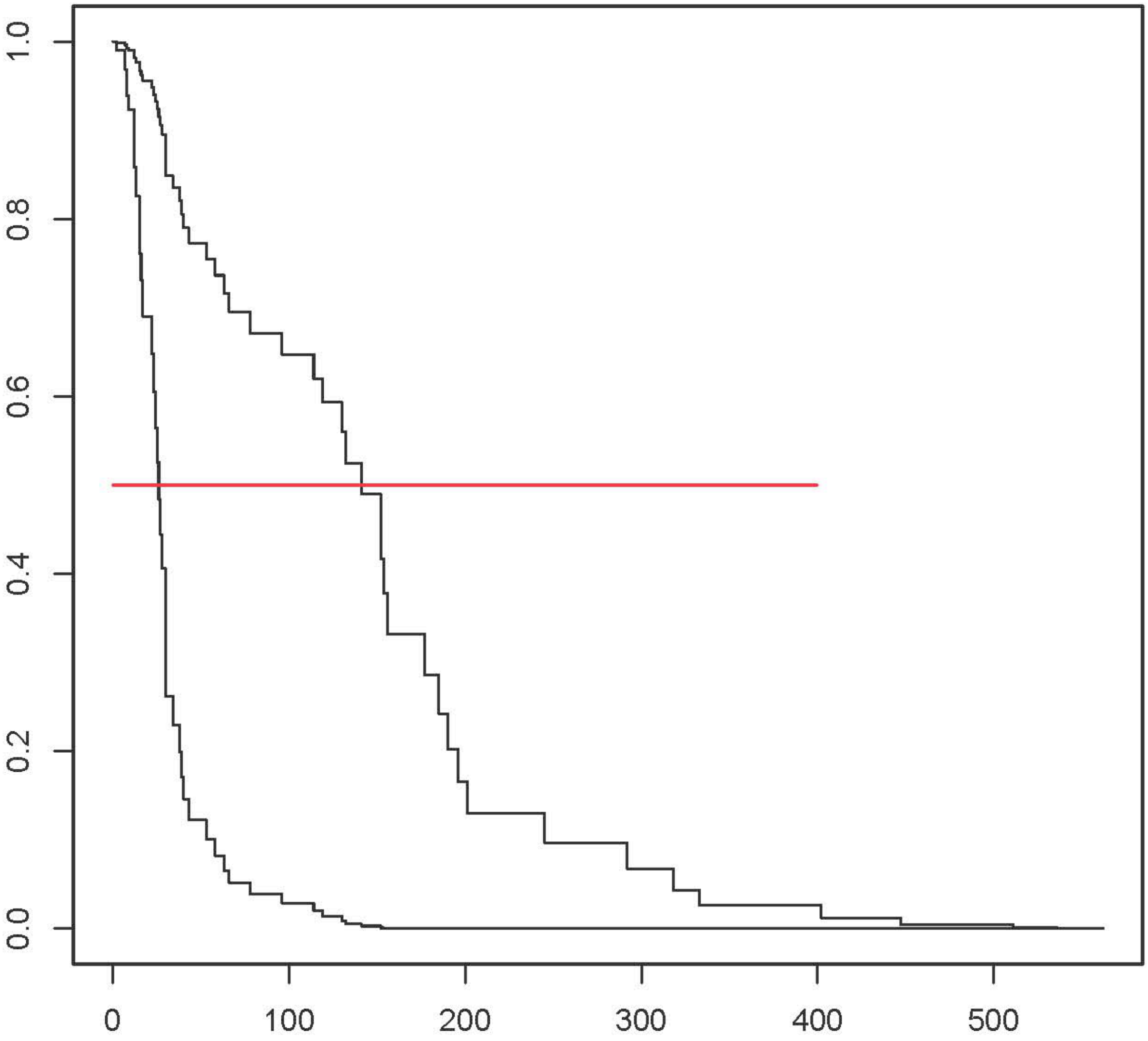
```
> ls(s1)
 [1] "call"        "conf.int"   "conf.type" "cumhaz"     "logse"      "lower"
 [7] "n"           "n.censor"   "n.event"   "n.risk"     "std.chaz"   "std.err"
[13] "surv"        "time"       "upper"
```

$$H(t) = \int_0^t h(y)\, dy \text{ and } S(t) = e^{-H(t)}$$

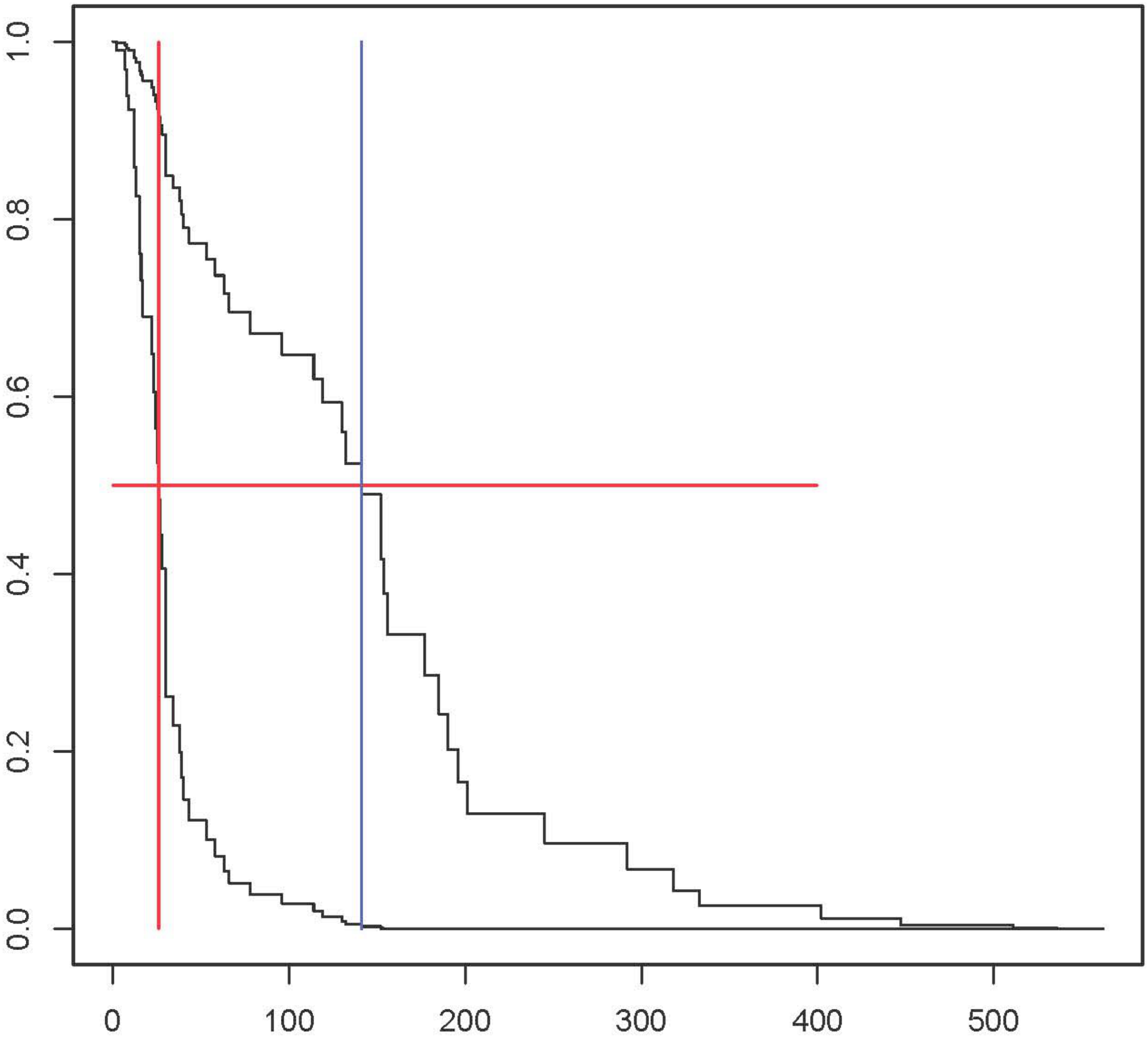
```
> S = s1$surv[1:10,1] # Col 1 is males
> H = s1$cumhaz[1:10,1]
> Q = exp(-H) # Question: Is this the survival function?
> cbind(H,S,Q)
                H         S         Q
 [1,] 0.009766264 0.9902813 0.9902813
 [2,] 0.009766264 0.9902813 0.9902813
 [3,] 0.009766264 0.9902813 0.9902813
 [4,] 0.009766264 0.9902813 0.9902813
 [5,] 0.032125563 0.9683850 0.9683850
 [6,] 0.062982652 0.9389598 0.9389598
 [7,] 0.079866525 0.9232396 0.9232396
 [8,] 0.152667391 0.8584152 0.8584152
 [9,] 0.191322662 0.8258661 0.8258661
[10,] 0.272747550 0.7612849 0.7612849

> plot(s1)
```



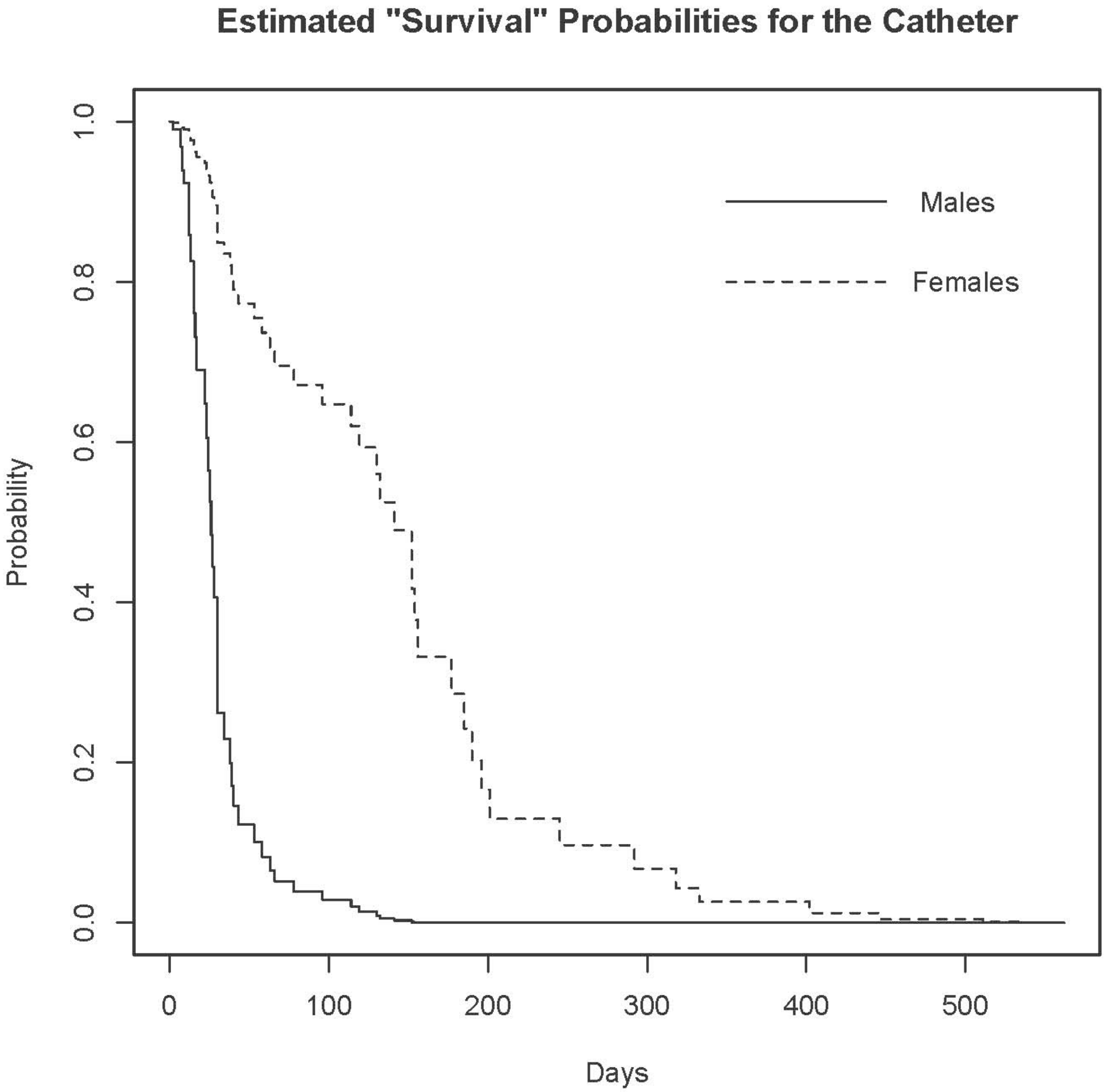
```
> # Try to locate the medians
> xx = c(0,400); yy = c(.5,0.5)
> lines(xx,yy,col="red")
```



```
> # Median for M = 26, F = 141 ?
> # Median for M = 26, F = 141 ?
> xm = c(26,26); ym = c(0,1); lines(xm,ym,col="red")
> xf = c(141,141); yf = c(0,1); lines(xf,yf,col="blue")
```

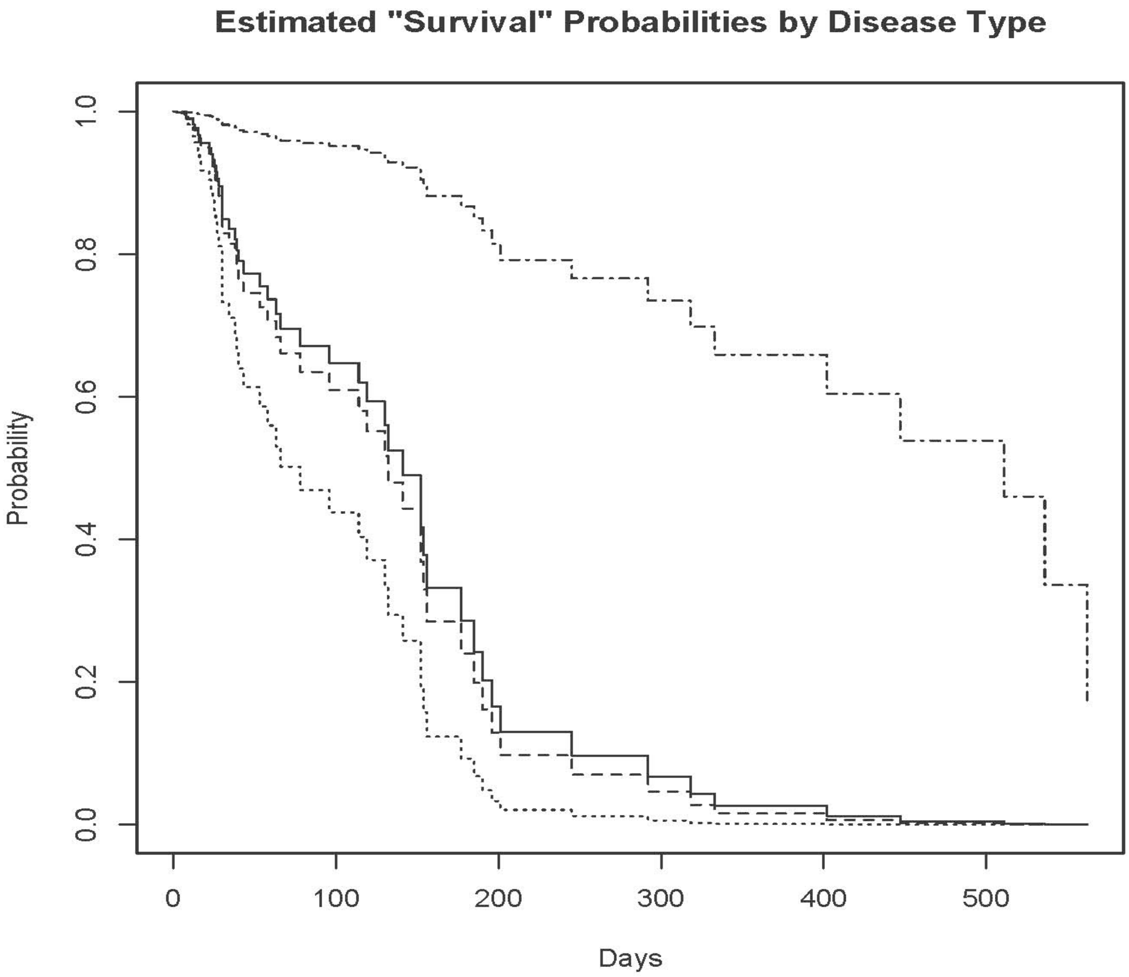


```
> # How about a nicer plot?
> plot(s1,lty = c(1,2),xlab="Days", ylab="Probability")
> title('Estimated "Survival" Probabilities for the Catheter')
>
> xm = c(350,450); ym = c(0.9,0.9); lines(xm,ym,lty=1)
> text(500,0.9,"Males  ")
>
> xf = c(350,450); yf = c(0.8,0.8); lines(xf,yf,lty=2)
> text(500,0.8,"Females")
```



```
> # Compare disease types, just for women
> table(Kidney$disease)

Other    GN    AN   PKD 
   26    18    24     8 
> 
> Other = data.frame(age=0, sex=1, disease="Other", frail=0)
> GN = data.frame(age=0, sex=1, disease="GN", frail=0)
> AN = data.frame(age=0, sex=1, disease="AN", frail=0)
> PKD = data.frame(age=0, sex=1, disease="PKD", frail=0)
> discomp = rbind(Other, GN, AN, PKD)
> rownames(discomp) = c("Other", "GN", "AN", "PKD")
> s2 = survfit(kmod1,newdata=discomp); s2
Call: survfit(formula = kmod1, newdata = discomp)

       n events median 0.95LCL 0.95UCL
Other 76     58    141      96     245
GN    76     58    132      66     318
AN    76     58     78      40     177
PKD   76     58    511     318      NA
> 
> summary(kmod1)
Call:
coxph(formula = Surv(time, status) ~ age + sex + disease + frail, 
    data = Kidney)

  n= 76, number of events= 58 

                coef exp(coef)  se(coef)      z Pr(>|z|)    
age         0.007714  1.007744  0.011907  0.648 0.517055    
sex        -2.099844  0.122475  0.392654 -5.348 8.90e-08 ***
diseaseGN   0.130666  1.139587  0.436114  0.300 0.764471    
diseaseAN   0.640906  1.898200  0.447886  1.431 0.152442    
diseasePKD -2.168515  0.114347  0.648825 -3.342 0.000831 ***
frail       1.791873  6.000682  0.257639  6.955 3.53e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
age           1.0077     0.9923   0.98450    1.0315
sex           0.1225     8.1649   0.05673    0.2644
diseaseGN     1.1396     0.8775   0.48476    2.6790
diseaseAN     1.8982     0.5268   0.78904    4.5665
diseasePKD    0.1143     8.7453   0.03206    0.4079
frail         6.0007     0.1666   3.62158    9.9427

Concordance= 0.822  (se = 0.03 )
Likelihood ratio test= 68.71  on 6 df,   p=8e-13
Wald test            = 60.01  on 6 df,   p=4e-11
Score (logrank) test = 86.24  on 6 df,   p=<2e-16


> 
> # Catheters for patients with PKD stay in longest.
> # How about some pairwise comparisons?
> 
```

```
> s2
Call: survfit(formula = kmod1, newdata = discomp)

       n events median 0.95LCL 0.95UCL
Other 76     58    141      96     245
GN    76     58    132      66     318
AN    76     58     78      40     177
PKD   76     58    511     318      NA

> plot(s2,lty = 1:4,xlab="Days", ylab="Probability")
> title('Estimated "Survival" Probabilities by Disease Type')
```


Estimated "Survival" Probabilities by Disease Type
Probability
Days
