Breast Cancer Survivor

Andrew Nguyen

2023-08-26

Contents

1 Load Libraries and dataset **1**

1.1 Data cleaning: 3

1.2 Fitting cox hazard proportional regression models (survival analysis) 4

1.2.1 Results 5

2 sessionInfo **6**

1 Load Libraries and dataset

I have downloaded a **breast cancer dataset** from Kaggle. It has the following variables:

> The dataset contains information on breast cancer patients, including their Patient_ID, Age, Gender, and expression levels of four proteins (Protein1, Protein2, Protein3, Protein4). The dataset also includes the Breast cancer stage of the patient (Tumor_Stage), Histology (type of cancer), ER, PR, and HER2 status, Surgery_type, Date of Surgery, Date of Last Visit, and Patient Status (Alive/Dead).

Also…

> This information can be used to analyze the relationship between protein expression levels, cancer stage, and patient outcomes. It can also be used to understand the impact of different types of surgeries on patient survival and to identify potential risk factors for breast cancer progression.

From this information, I'm going to assume that the different in time from surgery date and date of last visit is the time of event.

```r
library(tidyverse)# package for data wrangling
library(lubridate) # package for timing
library(survival) # for fitting cox hazard proportional regression models

dat<-read.csv("data/breast_cancer_survival.csv")
dat <- dat %>%
  mutate(across(where(is.character), as.factor))
glimpse(dat)
```

```
## Rows: 334
## Columns: 15
## $ Age                <int> 42, 54, 63, 78, 42, 80, 66, 36, 58, 62, 51, 40, 77,~
## $ Gender             <fct> FEMALE, FEMALE, FEMALE, FEMALE, FEMALE, FEMALE, FEM~
## $ Protein1           <dbl> 0.952560, 0.000000, -0.523030, -0.876180, 0.226110,~
## $ Protein2           <dbl> 2.15000, 1.38020, 1.76400, 0.12943, 1.74910, 2.5797~
## $ Protein3           <dbl> 0.0079716, -0.4980300, -0.3701900, -0.3703800, -0.5~
## $ Protein4           <dbl> -0.048340, -0.507320, 0.010815, 0.132190, -0.390210~
## $ Tumour_Stage       <fct> II, II, II, I, II, III, II, III, II, II, II, II, I,~
## $ Histology          <fct> Infiltrating Ductal Carcinoma, Infiltrating Ductal ~
## $ ER.status          <fct> Positive, Positive, Positive, Positive, Positive, P~
## $ PR.status          <fct> Positive, Positive, Positive, Positive, Positive, P~
## $ HER2.status        <fct> Negative, Negative, Negative, Negative, Positive, N~
## $ Surgery_type       <fct> Other, Other, Lumpectomy, Other, Lumpectomy, Modifi~
## $ Date_of_Surgery    <fct> 20-May-18, 26-Apr-18, 24-Aug-18, 16-Nov-18, 12-Dec-~
## $ Date_of_Last_Visit <fct> 26-Aug-18, 25-Jan-19, 08-Apr-20, 28-Jul-20, 05-Jan-~
## $ Patient_Status     <fct> Alive, Dead, Alive, Alive, Alive, Alive, Alive, Ali~
```

```
summary(dat)
```

```
##       Age           Gender       Protein1            Protein2
##  Min.   :29.00   FEMALE:330   Min.   :-2.340900   Min.   :-0.9787
##  1st Qu.:49.00   MALE  :  4   1st Qu.:-0.358888   1st Qu.: 0.3622
##  Median :58.00                Median : 0.006129   Median : 0.9928
##  Mean   :58.89                Mean   :-0.029991   Mean   : 0.9469
##  3rd Qu.:68.00                3rd Qu.: 0.343598   3rd Qu.: 1.6279
##  Max.   :90.00                Max.   : 1.593600   Max.   : 3.4022
##
##     Protein3          Protein4         Tumour_Stage
##  Min.   :-1.6274   Min.   :-2.025500   I  : 64
##  1st Qu.:-0.5137   1st Qu.:-0.377090   II :189
##  Median :-0.1732   Median : 0.041768   III: 81
##  Mean   :-0.0902   Mean   : 0.009819
##  3rd Qu.: 0.2784   3rd Qu.: 0.425630
##  Max.   : 2.1934   Max.   : 1.629900
##
##                              Histology      ER.status      PR.status
##  Infiltrating Ductal Carcinoma :233   Positive:334   Positive:334
##  Infiltrating Lobular Carcinoma: 89
##  Mucinous Carcinoma            : 12
##
##
##
##
##    HER2.status                       Surgery_type  Date_of_Surgery
##  Negative:305   Lumpectomy                 : 66   06-Dec-18:  5
##  Positive: 29   Modified Radical Mastectomy: 96   06-Nov-18:  5
##                 Other                      :105   16-Nov-18:  5
##                 Simple Mastectomy          : 67   26-Nov-18:  5
##                                                   16-Dec-18:  4
##                                                   17-Oct-18:  4
##                                                   (Other)  :306
##  Date_of_Last_Visit Patient_Status
##           : 17           : 13
##  03-Feb-21:  3      Alive:255
##  09-Aug-19:  3      Dead : 66
##  09-Feb-20:  3
##  13-Feb-21:  3
##  15-Jan-20:  3
##  (Other)  :302
```

```
### gender -> exclude from analysis, only 4 males
#N = 334
# ER.status, PR.status are not informative
#HER2.status is not very informative either

#check how many have survived and not
dat%>%
  group_by(Patient_Status)%>%
  count(Patient_Status)
```

```
## # A tibble: 3 x 2
## # Groups:   Patient_Status [3]
##   Patient_Status     n
##   <fct>          <int>
## 1 ""                13
## 2 "Alive"          255
## 3 "Dead"            66
```

```
dat%>%
  dplyr::group_by(Surgery_type)%>%
  count(Patient_Status)
```

```
## # A tibble: 11 x 3
## # Groups:   Surgery_type [4]
##    Surgery_type                Patient_Status     n
##    <fct>                       <fct>          <int>
##  1 Lumpectomy                  "Alive"           57
##  2 Lumpectomy                  "Dead"             9
##  3 Modified Radical Mastectomy ""                 4
##  4 Modified Radical Mastectomy "Alive"           72
##  5 Modified Radical Mastectomy "Dead"            20
##  6 Other                       ""                 7
##  7 Other                       "Alive"           73
##  8 Other                       "Dead"            25
##  9 Simple Mastectomy           ""                 2
## 10 Simple Mastectomy           "Alive"           53
## 11 Simple Mastectomy           "Dead"            12
```

```
# there are missing values for patient status
```

1.1 Data cleaning:

Data cleaning list to conduct a survival analysis:

1. We need to find out the time between date of surgery and date of last visit.

2. We also need to exclude the missing values for patient status.
3. Exclude the gender,ER.status,PR.status,HER2.status, category from analysis because there are too few males

```
# removing missing patient status values
dat.stat<-dat%>%
  dplyr::filter(Patient_Status!="")
#removed 13 samples

##find out the timing
dat.stat$time<-time_length(interval(dmy(dat.stat$Date_of_Surgery),dmy(dat.stat$Date_of_Last_Visit)),"da
### there are NA's, should be removed
dat.stat<-dat.stat%>%
  dplyr::filter(!is.na(time))%>%
  mutate(ps=as.numeric(ifelse(Patient_Status=="Alive",1,2)))
# alive =1, dead = 0
#dat.stat$Patient_Status<-factor(dat.stat$Patient_Status,levels=c("Dead","Alive"),labels=c("1","2"))
#removes 4 samples
```

1.2 Fitting cox hazard proportional regression models (survival analysis)

Variables of interest:
* Age
* Protein 1-4
* Histology
* Tumour stage
* Surgery type -> probably the most important given that this is the intervention

```
### let's explore protein levels
mod1<-coxph(Surv(time, ps) ~ Protein1+Protein2+Protein3+Protein4+Age+Surgery_type+Tumour_Stage+Histolog

summary(mod1)
```

```
## Call:
## coxph(formula = Surv(time, ps) ~ Protein1 + Protein2 + Protein3 +
##     Protein4 + Age + Surgery_type + Tumour_Stage + Histology,
##     data = dat.stat)
##
##   n= 317, number of events= 62
##
##                                            coef exp(coef)  se(coef)      z
## Protein1                              -0.207075  0.812958  0.248324 -0.834
## Protein2                               0.269374  1.309145  0.159057  1.694
## Protein3                               0.123179  1.131087  0.233552  0.527
## Protein4                               0.568643  1.765869  0.242925  2.341
## Age                                   -0.002671  0.997332  0.011119 -0.240
## Surgery_typeModified Radical Mastectomy  0.488416  1.629733  0.434497  1.124
## Surgery_typeOther                      0.365225  1.440838  0.411068  0.888
## Surgery_typeSimple Mastectomy          0.188490  1.207425  0.470999  0.400
## Tumour_StageII                         0.465304  1.592499  0.383021  1.215
## Tumour_StageIII                        0.830972  2.295549  0.446429  1.861
## HistologyInfiltrating Lobular Carcinoma -0.234382  0.791060  0.317064 -0.739
## HistologyMucinous Carcinoma              0.250518  1.284691  0.634014  0.395
##                                         Pr(>|z|)
## Protein1                                  0.4043
## Protein2                                  0.0903 .
## Protein3                                  0.5979
## Protein4                                  0.0192 *
## Age                                       0.8101
## Surgery_typeModified Radical Mastectomy   0.2610
## Surgery_typeOther                         0.3743
## Surgery_typeSimple Mastectomy             0.6890
## Tumour_StageII                            0.2244
## Tumour_StageIII                           0.0627 .
## HistologyInfiltrating Lobular Carcinoma   0.4598
## HistologyMucinous Carcinoma               0.6927
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                         exp(coef) exp(-coef) lower .95
## Protein1                                   0.8130     1.2301    0.4997
## Protein2                                   1.3091     0.7639    0.9585
## Protein3                                   1.1311     0.8841    0.7156
## Protein4                                   1.7659     0.5663    1.0969
## Age                                        0.9973     1.0027    0.9758
## Surgery_typeModified Radical Mastectomy    1.6297     0.6136    0.6955
## Surgery_typeOther                          1.4408     0.6940    0.6437
## Surgery_typeSimple Mastectomy              1.2074     0.8282    0.4797
## Tumour_StageII                             1.5925     0.6279    0.7517
## Tumour_StageIII                            2.2955     0.4356    0.9569
## HistologyInfiltrating Lobular Carcinoma    0.7911     1.2641    0.4249
## HistologyMucinous Carcinoma                1.2847     0.7784    0.3708
##                                         upper .95
## Protein1                                    1.323
## Protein2                                    1.788
## Protein3                                    1.788
## Protein4                                    2.843
## Age                                         1.019
## Surgery_typeModified Radical Mastectomy     3.819
## Surgery_typeOther                           3.225
## Surgery_typeSimple Mastectomy               3.039
## Tumour_StageII                              3.374
## Tumour_StageIII                             5.507
## HistologyInfiltrating Lobular Carcinoma     1.473
## HistologyMucinous Carcinoma                 4.451
##
## Concordance= 0.624  (se = 0.046 )
## Likelihood ratio test= 15.31  on 12 df,   p=0.2
## Wald test            = 14.48  on 12 df,   p=0.3
## Score (logrank) test = 14.93  on 12 df,   p=0.2
```

1.2.1 Results

No significant effects observed in the hazard rate of breast cancer patients undergoing different types of surgery or patients with different types of histology or tumor status. There may be a trend for tumour status such that patients with stage III tumours have 129% higher hazard death rate than stage III tumours. Patients with greater protein 4 have 76% increase in the hazard death rate per unit of concentration.

2 sessionInfo

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
## 
## Matrix products: default
## 
## 
## locale:
## [1] LC_COLLATE=English_United States.utf8 
## [2] LC_CTYPE=English_United States.utf8   
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.utf8    
## 
## time zone: America/New_York
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] survival_3.5-7  lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0  
##  [5] dplyr_1.1.2     purrr_1.0.2     readr_2.1.4     tidyr_1.3.0    
##  [9] tibble_3.2.1    ggplot2_3.4.3   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] Matrix_1.6-1      gtable_0.3.4      compiler_4.3.1    tidyselect_1.2.0 
##  [5] splines_4.3.1     scales_1.2.1      yaml_2.3.7        fastmap_1.1.1    
##  [9] lattice_0.21-8    R6_2.5.1          generics_0.1.3    knitr_1.43       
## [13] munsell_0.5.0     pillar_1.9.0      tzdb_0.4.0        rlang_1.1.1      
## [17] utf8_1.2.3        stringi_1.7.12    xfun_0.40         timechange_0.2.0 
## [21] cli_3.6.1         withr_2.5.0       magrittr_2.0.3    digest_0.6.33    
## [25] grid_4.3.1        rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.3  
## [29] vctrs_0.6.3       evaluate_0.21     glue_1.6.2        fansi_1.0.4      
## [33] colorspace_2.1-0  rmarkdown_2.24    tools_4.3.1       pkgconfig_2.0.3  
## [37] htmltools_0.5.6
```