A Complete Case Study: the Lung Cancer dataset

Data preparation

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'dplyr' was built under R version 3.6.1

## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(survival)

?lung
```

```
## starting httpd help server ...

##  done
```

```
table(lung$inst)
```

```
##
##  1  2  3  4  5  6  7 10 11 12 13 15 16 21 22 26 32 33
## 36  5 19  4  9 14  8  4 18 23 20  6 16 13 17  6  7  2
```

```
nrow(lung)
```

```
## [1] 228
```

```
d_raw <- as.tibble(lung)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
summary(d_raw)
```

```
##       inst            time            status           age       
##  Min.   : 1.00   Min.   :   5.0   Min.   :1.000   Min.   :39.00  
##  1st Qu.: 3.00   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00  
##  Median :11.00   Median : 255.5   Median :2.000   Median :63.00  
##  Mean   :11.09   Mean   : 305.2   Mean   :1.724   Mean   :62.45  
##  3rd Qu.:16.00   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00  
##  Max.   :33.00   Max.   :1022.0   Max.   :2.000   Max.   :82.00  
##  NA's   :1                                                       
##       sex           ph.ecog          ph.karno        pat.karno     
##  Min.   :1.000   Min.   :0.0000   Min.   : 50.00   Min.   : 30.00  
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00  
##  Median :1.000   Median :1.0000   Median : 80.00   Median : 80.00  
##  Mean   :1.395   Mean   :0.9515   Mean   : 81.94   Mean   : 79.96  
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00  
##  Max.   :2.000   Max.   :3.0000   Max.   :100.00   Max.   :100.00  
##                  NA's   :1        NA's   :1        NA's   :3       
##     meal.cal         wt.loss       
##  Min.   :  96.0   Min.   :-24.000  
##  1st Qu.: 635.0   1st Qu.:  0.000  
##  Median : 975.0   Median :  7.000  
##  Mean   : 928.8   Mean   :  9.832  
##  3rd Qu.:1150.0   3rd Qu.: 15.750  
##  Max.   :2600.0   Max.   : 68.000  
##  NA's   :47       NA's   :14       
```

```
d <- mutate(d_raw,
  event = 0 + (status == 2),
  inst = factor(inst),
  sex = factor(sex, levels = 1:2, labels = c("male", "female"))
)
d
```

```
## # A tibble: 228 x 11
##    inst   time status   age sex    ph.ecog ph.karno pat.karno meal.cal
##    <fct> <dbl>  <dbl> <dbl> <fct>    <dbl>    <dbl>     <dbl>    <dbl>
##  1 3       306      2    74 male         1       90       100     1175
##  2 3       455      2    68 male         0       90        90     1225
##  3 3      1010      1    56 male         0       90        90       NA
##  4 5       210      2    57 male         1       90        60     1150
##  5 1       883      2    60 male         0      100        90       NA
##  6 12     1022      1    74 male         1       50        80      513
##  7 7       310      2    68 fema~        2       70        60      384
##  8 11      361      2    71 fema~        2       60        80      538
##  9 1       218      2    53 male         1       70        80      825
## 10 7       166      2    61 male         2       70        70      271
## # ... with 218 more rows, and 2 more variables: wt.loss <dbl>, event <dbl>
```

```
summary(d)
```

```
##       inst          time            status           age       
##  1      : 36   Min.   :   5.0   Min.   :1.000   Min.   :39.00  
##  12     : 23   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00  
##  13     : 20   Median : 255.5   Median :2.000   Median :63.00  
```

```
##  3       : 19   Mean   : 305.2   Mean   :1.724   Mean   :62.45
##  11      : 18   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00
##  (Other):111   Max.   :1022.0   Max.   :2.000   Max.   :82.00
##  NA's    :  1
##      sex         ph.ecog          ph.karno        pat.karno
##  male  :138   Min.   :0.0000   Min.   : 50.00   Min.   : 30.00
##  female: 90   1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00
##               Median :1.0000   Median : 80.00   Median : 80.00
##               Mean   :0.9515   Mean   : 81.94   Mean   : 79.96
##               3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00
##               Max.   :3.0000   Max.   :100.00   Max.   :100.00
##               NA's   :1        NA's   :1        NA's   :3
##     meal.cal         wt.loss           event
##  Min.   :  96.0   Min.   :-24.000   Min.   :0.0000
##  1st Qu.: 635.0   1st Qu.:  0.000   1st Qu.:0.0000
##  Median : 975.0   Median :  7.000   Median :1.0000
##  Mean   : 928.8   Mean   :  9.832   Mean   :0.7237
##  3rd Qu.:1150.0   3rd Qu.: 15.750   3rd Qu.:1.0000
##  Max.   :2600.0   Max.   : 68.000   Max.   :1.0000
##  NA's   :47       NA's   :14
```

Impute some missing values

```
fit.meal <- lm(meal.cal ~ sex, data = d)
#use this to fill in the missing value
summary(fit.meal)
```

```
##
## Call:
## lm(formula = meal.cal ~ sex, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.54 -252.70   44.46  194.46 1619.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   980.54      37.23  26.335   <2e-16 ***
## sexfemale    -139.84      61.20  -2.285   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 397.5 on 179 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.02835,    Adjusted R-squared:  0.02292
## F-statistic: 5.222 on 1 and 179 DF,  p-value: 0.02348
```

```
d$meal.cal[is.na(d$meal.cal)] <-
  predict(fit.meal, newdata = subset(d, is.na(meal.cal)))
summary(d)
```

```
##       inst          time            status           age
```

```
##  1      : 36   Min.   :   5.0   Min.   :1.000   Min.   :39.00
##  12     : 23   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00
##  13     : 20   Median : 255.5   Median :2.000   Median :63.00
##  3      : 19   Mean   : 305.2   Mean   :1.724   Mean   :62.45
##  11     : 18   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00
##  (Other):111   Max.   :1022.0   Max.   :2.000   Max.   :82.00
##  NA's   :  1
##      sex         ph.ecog          ph.karno        pat.karno
##  male  :138   Min.   :0.0000   Min.   : 50.00   Min.   : 30.00
##  female: 90   1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00
##               Median :1.0000   Median : 80.00   Median : 80.00
##               Mean   :0.9515   Mean   : 81.94   Mean   : 79.96
##               3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00
##               Max.   :3.0000   Max.   :100.00   Max.   :100.00
##               NA's   :1        NA's   :1        NA's   :3
##     meal.cal         wt.loss            event
##  Min.   :  96.0   Min.   :-24.000   Min.   :0.0000
##  1st Qu.: 768.0   1st Qu.:  0.000   1st Qu.:0.0000
##  Median : 977.8   Median :  7.000   Median :1.0000
##  Mean   : 925.3   Mean   :  9.832   Mean   :0.7237
##  3rd Qu.:1075.0   3rd Qu.: 15.750   3rd Qu.:1.0000
##  Max.   :2600.0   Max.   : 68.000   Max.   :1.0000
##                   NA's   :14
```

```
d$wt.loss[is.na(d$wt.loss)] <-
  predict(lm(wt.loss ~ age + sex, data = d), newdata = subset(d, is.na(wt.loss)))
summary(d)
```

```
##       inst          time            status           age
##  1      : 36   Min.   :   5.0   Min.   :1.000   Min.   :39.00
##  12     : 23   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00
##  13     : 20   Median : 255.5   Median :2.000   Median :63.00
##  3      : 19   Mean   : 305.2   Mean   :1.724   Mean   :62.45
##  11     : 18   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00
##  (Other):111   Max.   :1022.0   Max.   :2.000   Max.   :82.00
##  NA's   :  1
##      sex         ph.ecog          ph.karno        pat.karno
##  male  :138   Min.   :0.0000   Min.   : 50.00   Min.   : 30.00
##  female: 90   1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00
##               Median :1.0000   Median : 80.00   Median : 80.00
##               Mean   :0.9515   Mean   : 81.94   Mean   : 79.96
##               3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00
##               Max.   :3.0000   Max.   :100.00   Max.   :100.00
##               NA's   :1        NA's   :1        NA's   :3
##     meal.cal         wt.loss            event
##  Min.   :  96.0   Min.   :-24.000   Min.   :0.0000
##  1st Qu.: 768.0   1st Qu.:  0.000   1st Qu.:0.0000
##  Median : 977.8   Median :  7.764   Median :1.0000
##  Mean   : 925.3   Mean   :  9.854   Mean   :0.7237
##  3rd Qu.:1075.0   3rd Qu.: 15.000   3rd Qu.:1.0000
##  Max.   :2600.0   Max.   : 68.000   Max.   :1.0000
##
```

One sample has no 'institute'??

```r
subset(d, is.na(inst))
```

```
## # A tibble: 1 x 11
##   inst   time status   age sex    ph.ecog ph.karno pat.karno meal.cal
##   <fct> <dbl>  <dbl> <dbl> <fct>    <dbl>    <dbl>     <dbl>    <dbl>
## 1 <NA>    329      2    69 male         2       70        80      713
## # ... with 2 more variables: wt.loss <dbl>, event <dbl>
```

```r
d$y <- with(d, Surv(time / 30.5, event))
d
```

```
## # A tibble: 228 x 12
##    inst   time status   age sex    ph.ecog ph.karno pat.karno meal.cal
##    <fct> <dbl>  <dbl> <dbl> <fct>    <dbl>    <dbl>     <dbl>    <dbl>
##  1 3       306      2    74 male         1       90       100    1175
##  2 3       455      2    68 male         0       90        90    1225
##  3 3      1010      1    56 male         0       90        90     981.
##  4 5       210      2    57 male         1       90        60    1150
##  5 1       883      2    60 male         0      100        90     981.
##  6 12     1022      1    74 male         1       50        80     513
##  7 7       310      2    68 fema~        2       70        60     384
##  8 11      361      2    71 fema~        2       60        80     538
##  9 1       218      2    53 male         1       70        80     825
## 10 7       166      2    61 male         2       70        70     271
## # ... with 218 more rows, and 4 more variables: wt.loss <dbl>,
## #   event <dbl>, y[,"time"] <dbl>, [,"status"] <dbl>
```

```r
head(d$y)
```

```
## [1] 10.032787  14.918033  33.114754+  6.885246  28.950820  33.508197+
```

Exploratory analysis

```r
survfit(y ~ 1, data = d)
```

```
## Call: survfit(formula = y ~ 1, data = d)
##
##       n  events  median 0.95LCL 0.95UCL
##  228.00  165.00   10.16    9.34   11.90
```

```r
survfit(y ~ sex, data = d)
```

```
## Call: survfit(formula = y ~ sex, data = d)
##
##              n events median 0.95LCL 0.95UCL
## sex=male   138    112   8.85    6.95    10.2
## sex=female  90     53  13.97   11.41    18.0
```

```r
str(d)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    228 obs. of  12 variables:
##  $ inst     : Factor w/ 18 levels "1","2","3","4",..: 3 3 3 5 1 10 7 9 1 7 ...
##  $ time     : num  306 455 1010 210 883 ...
##  $ status   : num  2 2 1 2 2 1 2 2 2 2 ...
##  $ age      : num  74 68 56 57 60 74 68 71 53 61 ...
##  $ sex      : Factor w/ 2 levels "male","female": 1 1 1 1 1 1 2 2 1 1 ...
##  $ ph.ecog  : num  1 0 0 1 0 1 2 2 1 2 ...
##  $ ph.karno : num  90 90 90 90 100 50 70 60 70 70 ...
##  $ pat.karno: num  100 90 90 60 90 80 60 80 80 70 ...
##  $ meal.cal : num  1175 1225 981 1150 981 ...
##  $ wt.loss  : num  11.6 15 15 11 0 ...
##  $ event    : num  1 1 0 1 1 0 1 1 1 1 ...
##  $ y        : 'Surv' num [1:228, 1:2] 10.033  14.918  33.115+  6.885  28.951  33.508+ 10.164  11.836
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr  "time" "status"
##   ..- attr(*, "type")= chr "right"
```

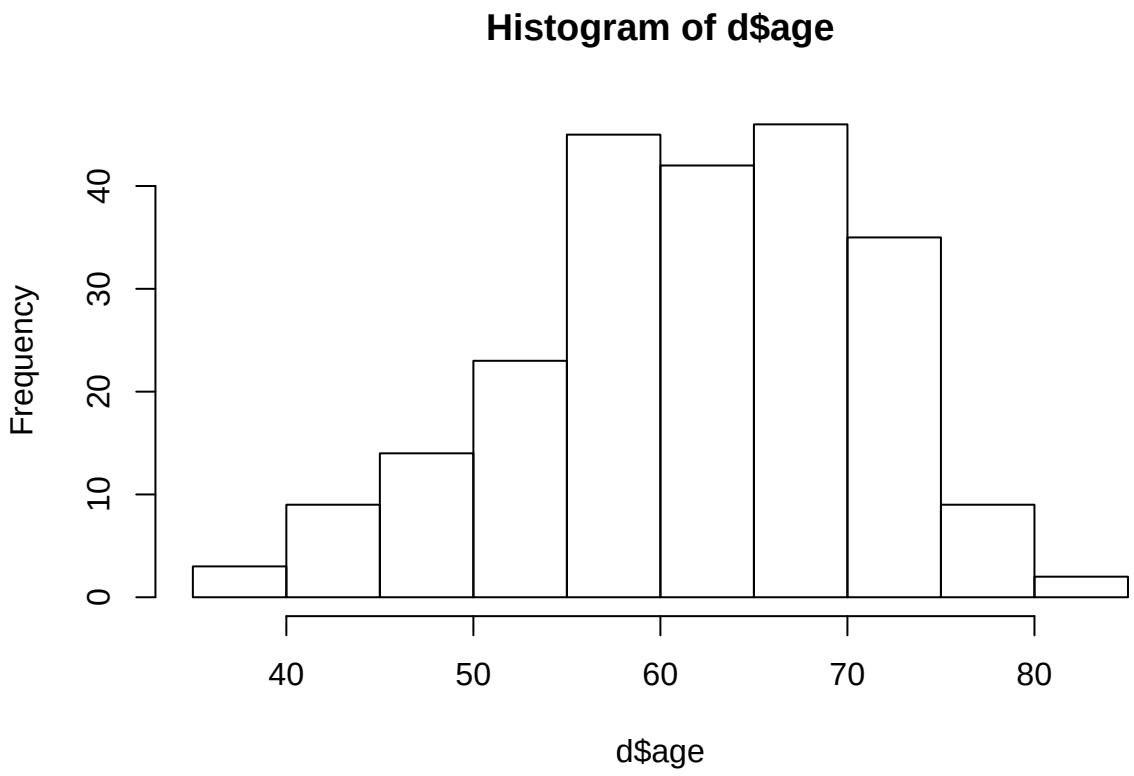
```r
table(d$ph.ecog)
```

```
##
##   0   1   2   3
##  63 113  50   1
```

```r
survfit(y ~ ph.ecog, data = d)
```

```
## Call: survfit(formula = y ~ ph.ecog, data = d)
##
##    1 observation deleted due to missingness
##             n events median 0.95LCL 0.95UCL
## ph.ecog=0  63     37  12.92   11.41   18.82
## ph.ecog=1 113     82  10.03    8.79   14.07
## ph.ecog=2  50     44   6.52    5.11    9.44
## ph.ecog=3   1      1   3.87      NA      NA
```

```r
hist(d$age)
```




```
#categorize age
d$ageCat <- cut(d$age, breaks = c(0, 50, 70, Inf))
table(d$ageCat)
```

```
##
##   (0,50]  (50,70] (70,Inf]
##       26      156       46
```

```
survfit(y ~ ageCat, data = d)
```

```
## Call: survfit(formula = y ~ ageCat, data = d)
##
##                    n events median 0.95LCL 0.95UCL
## ageCat=(0,50]     26     16  10.49    7.84      NA
## ageCat=(50,70]   156    111  11.31    9.44    14.1
## ageCat=(70,Inf]   46     38   9.28    6.59    11.6
```

```
table(d$ph.karno)
```

```
##
##  50  60  70  80  90 100
##   6  19  32  67  74  29
```

```r
table(d$pat.karno)
```

```
##
##  30  40  50  60  70  80  90 100
##   2   2   4  30  41  51  60  35
```

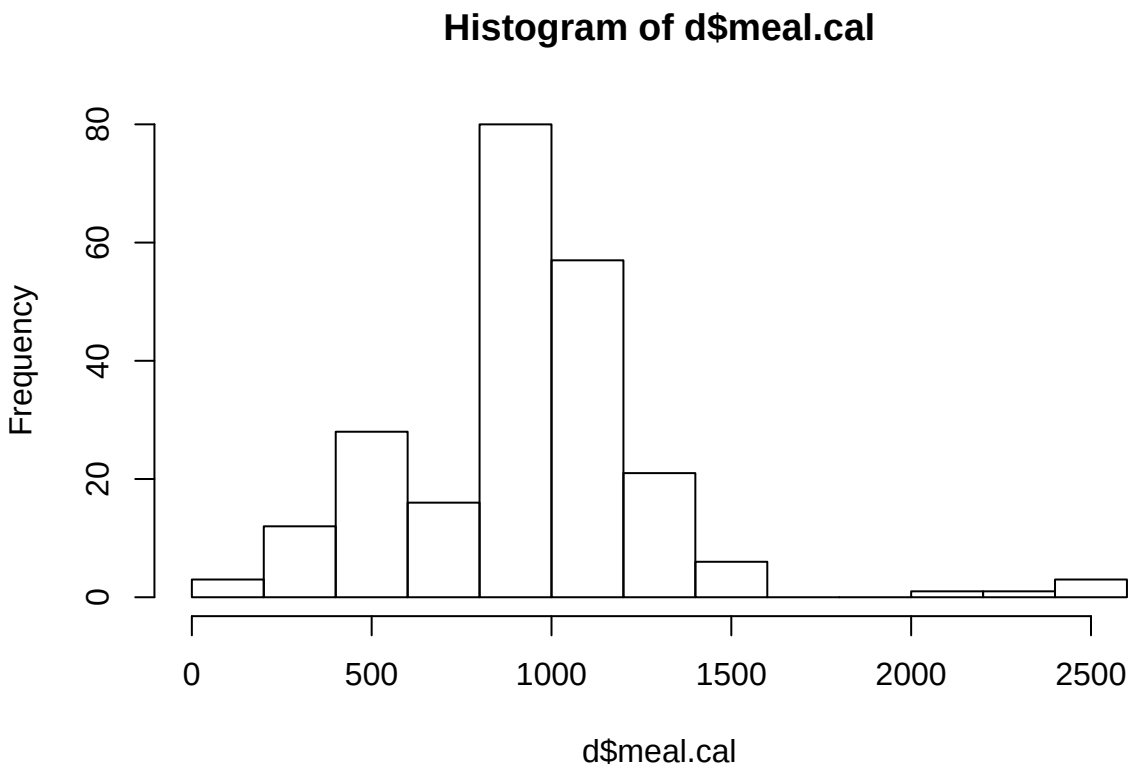
```r
survfit(y ~ ph.karno, data = d)
```

```
## Call: survfit(formula = y ~ ph.karno, data = d)
##
##    1 observation deleted due to missingness
##                n events median 0.95LCL 0.95UCL
## ph.karno=50    6      5   4.90    3.51      NA
## ph.karno=60   19     16   5.90    3.11      NA
## ph.karno=70   32     29   7.15    6.52    10.2
## ph.karno=80   67     47   9.61    7.54    12.8
## ph.karno=90   74     49  12.16   10.03    15.5
## ph.karno=100 29     18  14.03   11.15      NA
```

```r
survfit(y ~ pat.karno, data = d)
```

```
## Call: survfit(formula = y ~ pat.karno, data = d)
##
##    3 observations deleted due to missingness
##                 n events median 0.95LCL 0.95UCL
## pat.karno=30    2      1   5.11   5.115      NA
## pat.karno=40    2      1   3.05   3.049      NA
## pat.karno=50    4      4   3.07   0.393      NA
## pat.karno=60   30     27   6.46   5.344    9.44
## pat.karno=70   41     31   8.75   5.770   17.02
## pat.karno=80   51     39  11.41   7.410   17.05
## pat.karno=90   60     38  13.97   9.377   15.51
## pat.karno=100 35     21  12.16  10.164      NA
```

```r
hist(d$meal.cal)
```




```
survfit(y ~ I(meal.cal < 800), data = d)
```

```
## Call: survfit(formula = y ~ I(meal.cal < 800), data = d)
##
##                             n events median 0.95LCL 0.95UCL
## I(meal.cal < 800)=FALSE 169    123  11.41    9.38    12.9
## I(meal.cal < 800)=TRUE   59     42   9.34    6.82    11.6
```

above we do a survival with group

```
stem(d$wt.loss)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -2 | 4
##   -1 | 65
##   -1 | 310
##   -0 | 875555
##   -0 | 333322222211111
##    0 | 000000000000000000000000000000001111112222222223333333444444444
##    0 | 55555555566666777777888888888889
##    1 | 00000000000011111111111112222333334444
##    1 | 5555555555666777889
```

```
##    2 | 00000000023333344
##    2 | 57777889
##    3 | 0000122334
##    3 | 677789
##    4 |
##    4 | 9
##    5 | 2
##    5 |
##    6 | 0
##    6 | 8
```

```
survfit(y ~ I(wt.loss > 0), data = d)
```

```
## Call: survfit(formula = y ~ I(wt.loss > 0), data = d)
##
##                         n events median 0.95LCL 0.95UCL
## I(wt.loss > 0)=FALSE  61     42  11.90    9.34    15.5
## I(wt.loss > 0)=TRUE  167    123   9.87    8.75    11.6
```

```
stem(d$ph.karno)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    5 | 000000
##    5 |
##    6 | 0000000000000000000
##    6 |
##    7 | 00000000000000000000000000000000
##    7 |
##    8 | 000000000000000000000000000000000000000000000000000000000000000000000000
##    8 |
##    9 | 000000000000000000000000000000000000000000000000000000000000000000000000000
##    9 |
##   10 | 00000000000000000000000000000
```

```
stem(d$meal.cal)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    0 | 037
##    2 | 478911445689
##    4 | 11111146689999111114444899999
##    6 | 1348811333357778
##    8 | 33333333344444444444444444444468888811133333888888888888888888888888
##   10 | 333333333333333333333333468888888888880355588888888888888
##   12 | 333333333333338000000
##   14 | 330000
##   16 |
##   18 |
##   20 |
```

```
##   22 | 05
##   24 | 55
##   26 | 0
```

```
survfit(y ~ I(meal.cal < 800), data = d)
```

```
## Call: survfit(formula = y ~ I(meal.cal < 800), data = d)
##
##                             n events median 0.95LCL 0.95UCL
## I(meal.cal < 800)=FALSE 169    123  11.41    9.38    12.9
## I(meal.cal < 800)=TRUE   59     42   9.34    6.82    11.6
```

```
stem(d$wt.loss)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   -2 | 4
##   -1 | 65
##   -1 | 310
##   -0 | 875555
##   -0 | 333322222211111
##    0 | 000000000000000000000000000000011111122222222233333344444444
##    0 | 555555555566666777777888888888889
##    1 | 000000000000111111111111222233333444 4
##    1 | 5555555555666777889
##    2 | 00000000023333344
##    2 | 57777889
##    3 | 0000122334
##    3 | 677789
##    4 |
##    4 | 9
##    5 | 2
##    5 |
##    6 | 0
##    6 | 8
```

Comparing 2 groups (significant of the difference between group)

```
survdiff(y ~ I(wt.loss > 0), data = d)
```

```
## Call:
## survdiff(formula = y ~ I(wt.loss > 0), data = d)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## I(wt.loss > 0)=FALSE  61       42     44.6    0.1475     0.203
## I(wt.loss > 0)=TRUE  167      123    120.4    0.0546     0.203
##
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.7
```

```r
survdiff(y ~ ageCat, data = d)
```

```
## Call:
## survdiff(formula = y ~ ageCat, data = d)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## ageCat=(0,50]    26       16     20.6      1.02     1.175
## ageCat=(50,70]  156      111    116.7      0.28     0.964
## ageCat=(70,Inf]  46       38     27.7      3.83     4.640
##
##  Chisq= 5.2  on 2 degrees of freedom, p= 0.08
```

```r
survdiff(y ~ sex, data = d)
```

```
## Call:
## survdiff(formula = y ~ sex, data = d)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=male   138      112     91.6      4.55      10.3
## sex=female  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

```r
survdiff(y ~ I(meal.cal < 800), data = d)
```

```
## Call:
## survdiff(formula = y ~ I(meal.cal < 800), data = d)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## I(meal.cal < 800)=FALSE 169      123    127.1     0.130      0.57
## I(meal.cal < 800)=TRUE   59       42     37.9     0.435      0.57
##
##  Chisq= 0.6  on 1 degrees of freedom, p= 0.5
```

```r
#trick them as actegorical
survdiff(y ~ ph.ecog, data = d)
```

```
## Call:
## survdiff(formula = y ~ ph.ecog, data = d)
##
## n=227, 1 observation deleted due to missingness.
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## ph.ecog=0  63       37   54.153    5.4331    8.2119
## ph.ecog=1 113       82   83.528    0.0279    0.0573
## ph.ecog=2  50       44   26.147   12.1893   14.6491
## ph.ecog=3   1        1    0.172    3.9733    4.0040
##
##  Chisq= 22  on 3 degrees of freedom, p= 7e-05
```

```
survdiff(y ~ ph.karno, data = d)
```

```
## Call:
## survdiff(formula = y ~ ph.karno, data = d)
##
## n=227, 1 observation deleted due to missingness.
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## ph.karno=50    6        5     5.71    0.0874    0.0935
## ph.karno=60   19       16     9.64    4.1905    4.5095
## ph.karno=70   32       29    20.53    3.4943    4.0275
## ph.karno=80   67       47    43.30    0.3161    0.4390
## ph.karno=90   74       49    58.85    1.6489    2.5910
## ph.karno=100  29       18    25.97    2.4454    2.9262
##
##  Chisq= 12.3  on 5 degrees of freedom, p= 0.03
```

Lets organize our logrank test results in a more compact manner:

```
test_variable <- function(var_name) {
  e$x <- e[[ var_name ]]
  survdiff(y ~ x, data = e)
}
e <- mutate(d,
  weight_loss = wt.loss > 0,
  meal_calories_low = meal.cal < 800,
  age = ageCat
)

logrank_tests <-
  tibble(variable = c("weight_loss", "age", "sex", "meal_calories_low", "ph.ecog", "ph.karno")) %>%
  mutate(obj = map(variable, test_variable),
         tab = map(obj, broom::glance)) %>%
  unnest(tab)
logrank_tests
```

```
## # A tibble: 6 x 5
##   variable          obj        statistic    df   p.value
##   <chr>             <list>         <dbl> <dbl>     <dbl>
## 1 weight_loss       <survdiff>     0.203     1 0.652
## 2 age               <survdiff>     5.18      2 0.0751
## 3 sex               <survdiff>    10.3       1 0.00131
## 4 meal_calories_low <survdiff>     0.570     1 0.450
## 5 ph.ecog           <survdiff>    22.0       3 0.0000664
## 6 ph.karno          <survdiff>    12.3       5 0.0303
```

Data Modeling and Machine Learning

Models training

here we see higher ph.karno higher risk, but it's the oposite from the data understanding higher should be good

```r
load("./lung.RData")
d1 <- lung
fit <- coxph(y ~ ., data = d1)
summary(fit)
```

```
## Call:
## coxph(formula = y ~ ., data = d1)
##
##   n= 150, number of events= 108
##
##                 coef  exp(coef)   se(coef)      z Pr(>|z|)
## age        0.0098223  1.0098707  0.0117057  0.839 0.401410
## sexfemale -0.6792229  0.5070108  0.2098478 -3.237 0.001209 **
## ph.ecog    0.7358048  2.0871610  0.2217090  3.319 0.000904 ***
## ph.karno   0.0208466  1.0210654  0.0118747  1.756 0.079166 .
## pat.karno -0.0121843  0.9878896  0.0084753 -1.438 0.150539
## meal.cal  -0.0001839  0.9998161  0.0003352 -0.549 0.583249
## wt.loss   -0.0108340  0.9892245  0.0090554 -1.196 0.231532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0099     0.9902    0.9870     1.033
## sexfemale    0.5070     1.9723    0.3360     0.765
## ph.ecog      2.0872     0.4791    1.3516     3.223
## ph.karno     1.0211     0.9794    0.9976     1.045
## pat.karno    0.9879     1.0123    0.9716     1.004
## meal.cal     0.9998     1.0002    0.9992     1.000
## wt.loss      0.9892     1.0109    0.9718     1.007
##
## Concordance= 0.658  (se = 0.028 )
## Likelihood ratio test= 30.56  on 7 df,   p=7e-05
## Wald test            = 30.47  on 7 df,   p=8e-05
## Score (logrank) test = 31.46  on 7 df,   p=5e-05
```

model 1. AIC-STEP

```r
fit.aic <- step(fit)
```

```
## Start:  AIC=881.14
## y ~ age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss
##
##             Df    AIC
## - meal.cal   1 879.45
## - age        1 879.86
## - wt.loss    1 880.62
## <none>         881.14
## - pat.karno  1 881.19
## - ph.karno   1 882.39
## - sex        1 890.14
## - ph.ecog    1 890.32
```

```
##
## Step:  AIC=879.45
## y ~ age + sex + ph.ecog + ph.karno + pat.karno + wt.loss
##
##             Df    AIC
## - age        1 878.42
## - wt.loss    1 878.81
## <none>         879.45
## - pat.karno  1 880.11
## - ph.karno   1 880.59
## - sex        1 888.17
## - ph.ecog    1 888.33
##
## Step:  AIC=878.42
## y ~ sex + ph.ecog + ph.karno + pat.karno + wt.loss
##
##             Df    AIC
## - wt.loss    1 877.84
## <none>         878.42
## - pat.karno  1 878.87
## - ph.karno   1 878.88
## - ph.ecog    1 887.09
## - sex        1 887.37
##
## Step:  AIC=877.84
## y ~ sex + ph.ecog + ph.karno + pat.karno
##
##             Df    AIC
## - pat.karno  1 877.79
## <none>         877.84
## - ph.karno   1 878.08
## - ph.ecog    1 885.40
## - sex        1 886.95
##
## Step:  AIC=877.79
## y ~ sex + ph.ecog + ph.karno
##
##            Df    AIC
## - ph.karno  1 877.35
## <none>        877.79
## - sex       1 887.16
## - ph.ecog   1 888.73
##
## Step:  AIC=877.35
## y ~ sex + ph.ecog
##
##           Df    AIC
## <none>       877.35
## - sex      1 886.18
## - ph.ecog  1 890.03
```

```
summary(fit.aic)
```

```
## Call:
```

```
## coxph(formula = y ~ sex + ph.ecog, data = d1)
##
##   n= 150, number of events= 108
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## sexfemale -0.6558    0.5190   0.2045 -3.207 0.001341 **
## ph.ecog    0.5668    1.7626   0.1477  3.837 0.000125 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sexfemale     0.519     1.9268    0.3476    0.7749
## ph.ecog       1.763     0.5674    1.3195    2.3544
##
## Concordance= 0.651  (se = 0.03 )
## Likelihood ratio test= 24.35  on 2 df,   p=5e-06
## Wald test            = 24.32  on 2 df,   p=5e-06
## Score (logrank) test = 24.76  on 2 df,   p=4e-06
```

```
b.aic <- coef(fit.aic)
```

model 2. manual

```
fit.manual <- coxph(y ~ sex + ph.ecog + pat.karno + wt.loss, data = d1)
b.manual <- coef(fit.manual)
b.manual #from previous study
```

```
##    sexfemale      ph.ecog    pat.karno      wt.loss
## -0.644182430  0.504871795 -0.010404330 -0.009536832
```

model 3. elastic net

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.6.1
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when

## Loaded glmnet 2.0-18
```

```
X <- model.matrix(y ~ ., data = d1)[, -1]
str(X)
```

```
##  num [1:150, 1:7] 55 53 67 71 65 73 53 60 71 72 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:150] "82" "154" "108" "76" ...
##   ..$ : chr [1:7] "age" "sexfemale" "ph.ecog" "ph.karno" ...
```

```
head(X)
```

```
##     age sexfemale ph.ecog ph.karno pat.karno  meal.cal  wt.loss
## 82   55         0       1       70        90 1500.0000 15.00000
## 154  53         1       1       80        60  840.7015  4.00000
## 108  67         0       1       90        90  925.0000 11.33296
## 76   71         1       1       90        90 1075.0000 19.00000
## 64   65         1       1       90        80 1025.0000  0.00000
## 202  73         0       1       60        60 2200.0000  5.00000
```

```
y <- d1$y
fit <- cv.glmnet(X, y, family = "cox")
b.enet.all <- coef(fit, s = "lambda.min")
b.enet <- b.enet.all[b.enet.all != 0]
```

```
## <sparse>[ <logic> ] : .M.sub.i.logical() maybe inefficient
```

```
names(b.enet) <- colnames(X)[as.logical(b.enet.all != 0)]
```

model 4. CCP

```
fits <- plyr::adply(X, 2, function(x) broom::tidy(coxph(y ~ x)))
print(fits)
```

```
##          X1 term      estimate    std.error    statistic     p.value
## 1       age    x  1.805405e-02 0.0110270728  1.637247726 0.101578718
## 2 sexfemale    x -6.187376e-01 0.2038851908 -3.034735240 0.002407469
## 3   ph.ecog    x  5.529944e-01 0.1504871122  3.674695802 0.000238133
## 4  ph.karno    x -1.528937e-02 0.0075909560 -2.014156656 0.043993089
## 5 pat.karno    x -1.926016e-02 0.0065852850 -2.924726881 0.003447587
## 6  meal.cal    x -1.902571e-04 0.0003215785 -0.591635018 0.554095019
## 7   wt.loss    x  4.479606e-05 0.0082587905  0.005424047 0.995672258
##        conf.low     conf.high
## 1 -0.0035586157  0.0396667153
## 2 -1.0183452046 -0.2191299426
```

```
## 3  0.2580450395  0.8479436797
## 4 -0.0301673750 -0.0004113742
## 5 -0.0321670812 -0.0063532386
## 6 -0.0008205393  0.0004400251
## 7 -0.0161421358  0.0162317279
```

```
str(fits)
```

```
## 'data.frame':    7 obs. of  8 variables:
##  $ X1       : Factor w/ 7 levels "age","sexfemale",..: 1 2 3 4 5 6 7
##  $ term     : chr  "x" "x" "x" "x" ...
##  $ estimate : num  0.0181 -0.6187 0.553 -0.0153 -0.0193 ...
##  $ std.error: num  0.01103 0.20389 0.15049 0.00759 0.00659 ...
##  $ statistic: num  1.64 -3.03 3.67 -2.01 -2.92 ...
##  $ p.value  : num  0.101579 0.002407 0.000238 0.043993 0.003448 ...
##  $ conf.low : num  -0.00356 -1.01835 0.25805 -0.03017 -0.03217 ...
##  $ conf.high: num  0.039667 -0.21913 0.847944 -0.000411 -0.006353 ...
```

```
b.CCP <- with(fits, structure(estimate, names = as.character(X1)))
```

```
models_coefficients <- tibble(
  method = c("manual", "aic", "enet", "ccp"),
  coefficients = list(b.manual, b.aic, b.enet, b.CCP)
)
models_coefficients
```

```
## # A tibble: 4 x 2
##   method coefficients
##   <chr>  <list>
## 1 manual <dbl [4]>
## 2 aic    <dbl [2]>
## 3 enet   <dbl [3]>
## 4 ccp    <dbl [7]>
```

Models testing

```
lincom <- function(b, X) rowSums(sweep(X[, names(b), drop = FALSE], 2, b, FUN = "*"))
```

```
load("./lung_newdata.RData")
X.new <- model.matrix(y ~ . - 1, lung_newdata)
y <- lung_newdata$y
```

```
models_performance <- mutate(models_coefficients,
  predictions = map(coefficients, ~ lincom(., X.new)),
  cox_obj = map(predictions, ~ coxph(y ~ I(. / sd(.)))),
  cox_tab = map(cox_obj, broom::tidy)
) %>%
  unnest(cox_tab)
models_performance
```

```
## # A tibble: 4 x 11
##   method coefficients predictions cox_obj term  estimate std.error
##   <chr>  <list>       <list>      <list>  <chr>    <dbl>     <dbl>
## 1 manual <dbl [4]>    <dbl [73]>  <coxph> I(./~    0.340     0.150
## 2 aic    <dbl [2]>    <dbl [73]>  <coxph> I(./~    0.287     0.145
## 3 enet   <dbl [3]>    <dbl [73]>  <coxph> I(./~    0.297     0.144
## 4 ccp    <dbl [7]>    <dbl [73]>  <coxph> I(./~    0.305     0.140
## # ... with 4 more variables: statistic <dbl>, p.value <dbl>,
## #   conf.low <dbl>, conf.high <dbl>
```

```
models_performance <- mutate(models_performance,
  AUC = map_dbl(predictions, ~ survivalROC::survivalROC(y[, 1], y[, 2], ., predict.time = 12, method = 
) %>%
  select(method, estimate, std.error, p.value, AUC)
models_performance
```

```
## # A tibble: 4 x 5
##   method estimate std.error p.value   AUC
##   <chr>     <dbl>     <dbl>   <dbl> <dbl>
## 1 manual    0.340     0.150  0.0234 0.666
## 2 aic       0.287     0.145  0.0488 0.653
## 3 enet      0.297     0.144  0.0395 0.645
## 4 ccp       0.305     0.140  0.0297 0.671
```

Sharing the results outside of R

```
models_coefficients_flat <- mutate(models_coefficients,
  coefficients_tab = map(coefficients, ~ tibble(feature = names(.), coefficient = unname(.)))
) %>%
  unnest(coefficients_tab, .drop = TRUE) %>%
  select(method, feature, coefficient)
models_coefficients_flat
```

```
## # A tibble: 16 x 3
##    method feature   coefficient
##    <chr>  <chr>           <dbl>
##  1 manual sexfemale  -0.644
##  2 manual ph.ecog     0.505
##  3 manual pat.karno  -0.0104
##  4 manual wt.loss    -0.00954
##  5 aic    sexfemale  -0.656
##  6 aic    ph.ecog     0.567
##  7 enet   sexfemale  -0.418
##  8 enet   ph.ecog     0.340
##  9 enet   pat.karno  -0.00427
## 10 ccp    age         0.0181
## 11 ccp    sexfemale  -0.619
## 12 ccp    ph.ecog     0.553
## 13 ccp    ph.karno   -0.0153
## 14 ccp    pat.karno  -0.0193
## 15 ccp    meal.cal   -0.000190
## 16 ccp    wt.loss     0.0000448
```

Write tables on disk:

```
write_csv(models_coefficients_flat, "models_coefficients.csv")
write_csv(models_performance, "models_performance.csv")
```