Logistic regression with right censored (survival) data: Practicals with R

Paul Blanche (June 2024)

We will practice with `R` and the `rotterdam` data of the `survival` package. These data are observational data from the Rotterdam tumour bank. For practicing today, we will pretend that these data have been collected to investigate whether chemotherapy can reduce the 5-year risk of recurrence or death among women treated for breast cancer. Here the time-to-event outcome is recurrence-free survival time, defined as the time from primary surgery to the earlier of disease recurrence or death. The main analysis will aim to estimate the 5-year "causal" risk difference, that is, the risk difference that we expect if we randomize similar patients to chemotherapy or no chemotherapy.

We will further assume that:

- we have collected enough data on potential confounders to believe that the un-measured confounding assumption is reasonable.
- the process to collect and register the data makes the independent censoring assumption within each treatment group plausible.
- following thorough discussions with oncologists, we do not believe that interaction terms are needed in the logistic regression model.

Disclaimer: I know very little about these data and I have no idea whether these assumptions make sense, unfortunately.

Before proceeding to the main analysis, we will perform several supplementary/preliminary analyses, for completeness and to practice more with survival data.

Preliminaries

We first load the data and have a look a the first lines.

```
library(survival)
d <- rotterdam # for convenience
head(d)
```

```
##      pid year age meno  size grade nodes pgr  er hormon chemo rtime recur dtime
## 1393   1 1992  74    1  <=20     3     0  35 291      0     0  1799     0  1799
## 1416   2 1984  79    1 20-50     3     0  36 611      0     0  2828     0  2828
```

```
## 2962   3 1983  44    0  <=20     2     0 138   0      0     0  6012     0  6012
## 1455   4 1985  70    1 20-50     3     0   0  12      0     0  2624     0  2624
## 977    5 1983  75    1  <=20     3     0 260 409      0     0  4915     0  4915
## 617    6 1983  52    0  <=20     3     0 139 303      0     0  5888     0  5888
##      death
## 1393     0
## 1416     0
## 2962     0
## 1455     0
## 977      0
## 617      0
```

We then create a new `status` variable and change the time scale from days to year (for convenience).

```
d$time <- d$rtime/ 365.25
d$status <- d$recur
```

We get summary statistics for all variables.

```
summary(d)
```

```
##       pid              year          age              meno         size
##  Min.   :   1.0   Min.   :1978   Min.   :24.00   Min.   :0.00   <=20 :1387
##  1st Qu.: 753.2   1st Qu.:1986   1st Qu.:45.00   1st Qu.:0.00   20-50:1291
##  Median :1504.5   Median :1988   Median :54.00   Median :1.00   >50  : 304
##  Mean   :1505.0   Mean   :1988   Mean   :55.06   Mean   :0.56
##  3rd Qu.:2254.8   3rd Qu.:1990   3rd Qu.:65.00   3rd Qu.:1.00
##  Max.   :3007.0   Max.   :1993   Max.   :90.00   Max.   :1.00
##      grade           nodes             pgr               er
##  Min.   :2.000   Min.   : 0.000   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:   4.0   1st Qu.:  11.0
##  Median :3.000   Median : 1.000   Median :  41.0   Median :  61.0
##  Mean   :2.734   Mean   : 2.712   Mean   : 161.8   Mean   : 166.6
##  3rd Qu.:3.000   3rd Qu.: 4.000   3rd Qu.: 198.0   3rd Qu.: 202.8
##  Max.   :3.000   Max.   :34.000   Max.   :5004.0   Max.   :3275.0
##      hormon           chemo            rtime            recur
##  Min.   :0.0000   Min.   :0.0000   Min.   :  36.0   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 823.5   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1940.0   Median :1.0000
##  Mean   :0.1137   Mean   :0.1945   Mean   :2097.9   Mean   :0.5091
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:3198.8   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :7043.0   Max.   :1.0000
##      dtime          death             time             status
##  Min.   :  36   Min.   :0.0000   Min.   : 0.09856   Min.   :0.0000
##  1st Qu.:1607   1st Qu.:0.0000   1st Qu.: 2.25462   1st Qu.:0.0000
##  Median :2638   Median :0.0000   Median : 5.31143   Median :1.0000
```

```
##  Mean   :2605   Mean   :0.4266   Mean   : 5.74375   Mean   :0.5091
##  3rd Qu.:3555   3rd Qu.:1.0000   3rd Qu.: 8.75770   3rd Qu.:1.0000
##  Max.   :7043   Max.   :1.0000   Max.   :19.28268   Max.   :1.0000
```

We then create clinically relevant groups by categorizing some quantitative variables. This will be useful to fit a logistic model that does not rely on strong and questionable linearity assumptions.

```
d$yearcat <- cut(d$year,include.lowest=TRUE,
                 breaks=c(1978,1985,1988,1990,1993))
d$agecat <- cut(d$age,include.lowest=TRUE,
                breaks=c(24,35,40,45,50,55,60,65,90))
d$nodescat <- cut(d$nodes,include.lowest=TRUE,
                  breaks=c(0,1,3,5,10,Inf))
d$pgrcat <- cut(d$pgr,include.lowest=TRUE,
                breaks=c(0,20,40,70,100,150,Inf))
d$ercat <- cut(d$er,include.lowest=TRUE,
               breaks=c(0,7,15,40,60,80,100,140,200,Inf))
```

We print simple descriptive statistics for all the created variables.

```
summary(d[,grep("cat",names(d))],maxsum=9)
```

```
##         yearcat        agecat        nodescat          pgrcat
##  [1978,1985]:583   [24,35]:166   [0,1]   :1803   [0,20]   :1202
##  (1985,1988]:974   (35,40]:246   (1,3]   : 397   (20,40]  : 284
##  (1988,1990]:702   (40,45]:371   (3,5]   : 244   (40,70]  : 225
##  (1990,1993]:723   (45,50]:425   (5,10]  : 326   (70,100] : 161
##                    (50,55]:361   (10,Inf]: 212   (100,150]: 209
##                    (55,60]:338                   (150,Inf]: 901
##                    (60,65]:348
##                    (65,90]:727
##
##        ercat
##  [0,7]    :653
##  (7,15]   :210
##  (15,40]  :379
##  (40,60]  :246
##  (60,80]  :158
##  (80,100] :151
##  (100,140]:214
##  (140,200]:220
##  (200,Inf]:751
```

Transform some variables to factor.

```
d$chemo <- factor(d$chemo)
d$grade <- factor(d$grade)
```

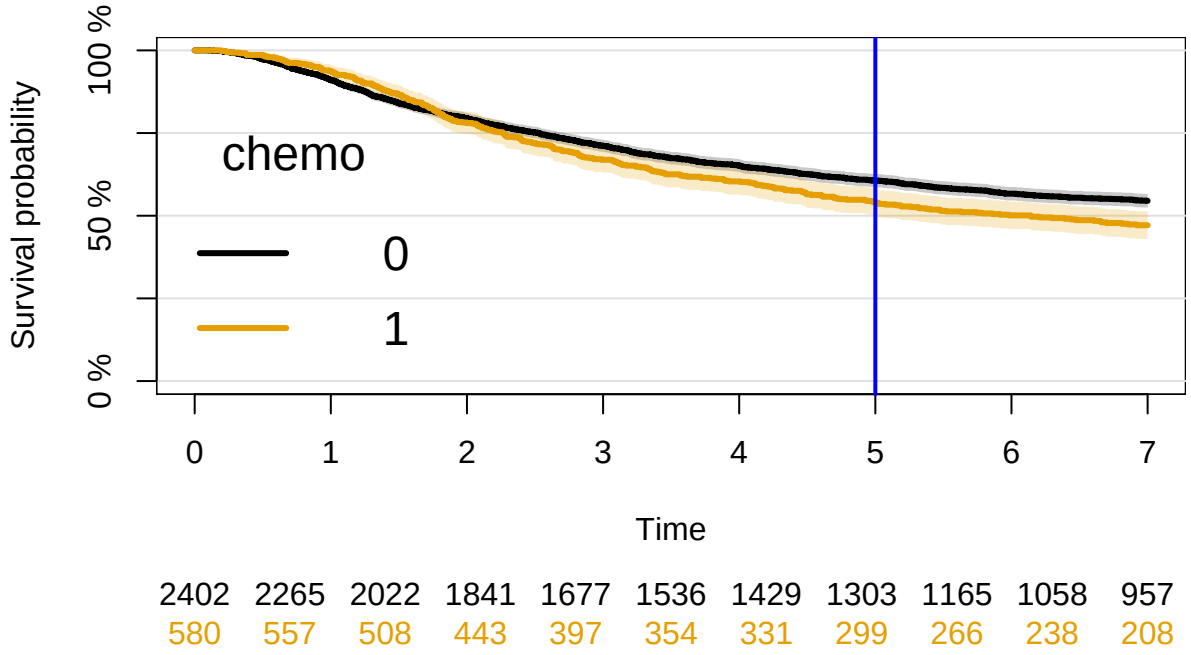
Question 1

Produce a Kaplan-Meier plot showing the estimated progression-free survival functions in each treatment group: with and without chemotherapy. We do that with the `prodlim` package (although the `survival` package could have done the job too). We focus on the results at t=5 years.

```
library(prodlim)
fitKM <- prodlim(Hist(time, status) ~ chemo, data = d)
summary(fitKM,time=5)
```

```
##   chemo time n.risk n.event n.lost  surv se.surv lower upper
## 1     0    5   1279       0      0 0.606  0.0102 0.586 0.626
## 2     1    5    292       0      0 0.541  0.0209 0.500 0.582
```

```
plot(fitKM,xlim=c(0,7),legend.x="bottomleft")
abline(v=5,lwd=2,col="blue")
```



Question 2

Produce a table with descriptive statistic for the following baseline covariates, per treatment group. That is, a usual "Table 1".

- year of inclusion (groups)
- age

- menopausal status
- tumor size
- grade
- number of positive lymph nodes
- progesterone receptors
- estrogen receptors
- hormonal treatment

This can be done via the following code, which computes frequencies and proportions per group for caterogical variables and medians with first and third quartiles for quantitative variables. We will assume that these variables are potential **confounders** that we would like to adjust for. What do you observe? Are the patients similar in the two groups?

```
library(Publish)
Table1 <- univariateTable(chemo~yearcat + Q(age) + meno +
                              size + factor(grade) + Q(nodes)
                          + Q(pgr) + Q(er) + hormon,
                          data=d,
                          compare.groups = FALSE,
                          show.totals = FALSE)
Table1
```

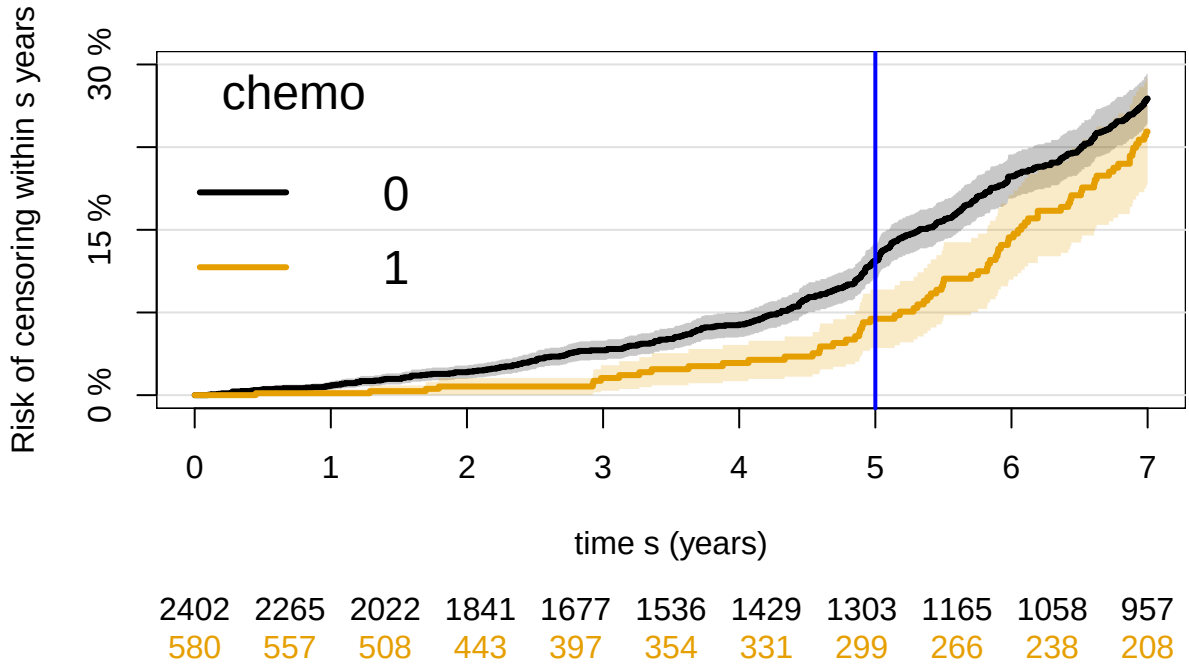
```
##    Variable        Level chemo = 0 (n=2,402) chemo = 1 (n=  580)
## 1   yearcat  [1978,1985]          457 (19.0)          126 (21.7)
## 2            (1985,1988]          795 (33.1)          179 (30.9)
## 3            (1988,1990]          580 (24.1)          122 (21.0)
## 4            (1990,1993]          570 (23.7)          153 (26.4)
## 5       age median [iqr]         58 [48, 67]         45 [40, 49]
## 6      meno             1        1,581 (65.8)           89 (15.3)
## 7                       0          821 (34.2)          491 (84.7)
## 8      size          <=20        1,148 (47.8)          239 (41.2)
## 9                   20-50        1,028 (42.8)          263 (45.3)
## 10                    >50           226 (9.4)           78 (13.4)
## 11    grade             2          640 (26.6)          154 (26.6)
## 12                      3        1,762 (73.4)          426 (73.4)
## 13    nodes median [iqr]             0 [0, 3]            2 [1, 5]
## 14      pgr median [iqr]       38 [  4, 186] 58.5 [  8.0, 234.2]
## 15       er median [iqr]       73 [ 12, 234]     40 [10.0, 98.2]
## 16   hormon             0        2,091 (87.1)          552 (95.2)
## 17                      1          311 (12.9)           28 (4.8)
```

Question 3

Produce a Kaplan-Meier plot showing the estimated censoring cumulative distribution in each treatment group: with and without chemotherapy. We focus on the relevant results

within the first 5 years. What do you observe?

```
fitKMC <- prodlim(Hist(time, status) ~ chemo, data = d,reverse=TRUE)
plot(fitKMC,xlim=c(0,7),
     ylim=c(0,0.3),
     type="cuminc",
     ylab="Risk of censoring within s years",
     xlab="time s (years)")
abline(v=5,lwd=2,col="blue")
```



Question 4

Just for completeness, look at how many patients are observed:

- with a recurrence within 5-years
- lost of follow-up (censored) within 5-years
- recurrence free at 5 years

Do you confirm that many patients are lost of follow-up within 5 years?

```
sum(d$time <=5 & d$status==1)
```

```
## [1] 1181
```

```
sum(d$time <=5 & d$status==0)
```

```
## [1] 230
```

```
sum(d$time >5)
```

```
## [1] 1571
```

Question 5

Fit a logistic regression model for the 5-year risk of recurrence, with chemotherapy as covariates as well as all the other variables listed in the previous **Question 2**. Use the categorical version of each quantitative variable, to facilitate the interpretation and, more importantly, to avoid making strong linearity assumptions. To account for right-censoring, we will use the "outcome weighed estimating equations" approach (oipcw) and compute the censoring weights using a Kaplan-Meier estimator stratified on treatment group. What can we conclude about the chemotherapy, from the fitted model?

```
library(mets)
out.oipcw <- binreg(Event(time, status) ~ chemo +
                        yearcat + agecat + meno +
                        size + grade + nodescat +
                        pgrcat + ercat + hormon,
                    data=d,
                    time=5,
                    cens.model=~strata(chemo))
summary(out.oipcw)
```

```
##
##     n events
##  2982   1181
##
##  2982 clusters
## coeffients:
##                       Estimate    Std.Err       2.5%      97.5% P-value
## (Intercept)        -0.4763831  0.2329776 -0.9330109 -0.0197553  0.0409
## chemo1             -0.4471183  0.1300514 -0.7020143 -0.1922223  0.0006
## yearcat(1985,1988] -0.0027659  0.1211456 -0.2402069  0.2346751  0.9818
## yearcat(1988,1990] -0.2380646  0.1309875 -0.4947954  0.0186662  0.0691
## yearcat(1990,1993] -0.3769903  0.1357711 -0.6430968 -0.1108839  0.0055
## agecat(35,40]      -0.0347107  0.2222912 -0.4703935  0.4009721  0.8759
## agecat(40,45]      -0.5150691  0.2102134 -0.9270798 -0.1030583  0.0143
## agecat(45,50]      -0.4452159  0.2074641 -0.8518381 -0.0385936  0.0319
## agecat(50,55]      -0.8697736  0.2418927 -1.3438745 -0.3956726  0.0003
## agecat(55,60]      -0.9055092  0.2900574 -1.4740112 -0.3370071  0.0018
## agecat(60,65]      -1.2821367  0.2955116 -1.8613288 -0.7029446  0.0000
## agecat(65,90]      -1.2079717  0.2865003 -1.7695021 -0.6464414  0.0000
## meno                0.2273539  0.2016438 -0.1678607  0.6225686  0.2595
## size20-50           0.4238079  0.0927874  0.2419478  0.6056679  0.0000
## size>50             0.6386524  0.1588523  0.3273076  0.9499973  0.0001
```

```
## grade3              0.4681845  0.1008975  0.2704291  0.6659398  0.0000
## nodescat(1,3]       0.9599467  0.1364211  0.6925663  1.2273271  0.0000
## nodescat(3,5]       1.1581213  0.1580345  0.8483793  1.4678632  0.0000
## nodescat(5,10]      1.7649340  0.1513806  1.4682334  2.0616345  0.0000
## nodescat(10,Inf]    2.0742591  0.2025189  1.6773294  2.4711888  0.0000
## pgrcat(20,40]      -0.1686248  0.1607714 -0.4837310  0.1464814  0.2942
## pgrcat(40,70]      -0.1816090  0.1760255 -0.5266127  0.1633947  0.3022
## pgrcat(70,100]     -0.2893758  0.2051257 -0.6914148  0.1126632  0.1583
## pgrcat(100,150]    -0.4212544  0.1935918 -0.8006874 -0.0418213  0.0296
## pgrcat(150,Inf]    -0.5992131  0.1312264 -0.8564120 -0.3420141  0.0000
## ercat(7,15]        -0.1019958  0.1826219 -0.4599282  0.2559366  0.5765
## ercat(15,40]        0.1111146  0.1633860 -0.2091160  0.4313452  0.4965
## ercat(40,60]        0.0662026  0.1880163 -0.3023025  0.4347077  0.7248
## ercat(60,80]       -0.0757457  0.2210714 -0.5090378  0.3575463  0.7319
## ercat(80,100]       0.1760498  0.2172393 -0.2497313  0.6018310  0.4177
## ercat(100,140]      0.0774805  0.2034381 -0.3212509  0.4762120  0.7033
## ercat(140,200]     -0.0173820  0.2009444 -0.4112258  0.3764618  0.9311
## ercat(200,Inf]      0.2208754  0.1592263 -0.0912025  0.5329533  0.1654
## hormon             -0.3585079  0.1598838 -0.6718745 -0.0451413  0.0249
##
## exp(coeffients):
##                    Estimate    2.5%   97.5%
## (Intercept)         0.62103 0.39337  0.9804
## chemo1              0.63947 0.49559  0.8251
## yearcat(1985,1988]  0.99724 0.78647  1.2645
## yearcat(1988,1990]  0.78815 0.60970  1.0188
## yearcat(1990,1993]  0.68592 0.52566  0.8950
## agecat(35,40]       0.96588 0.62476  1.4933
## agecat(40,45]       0.59746 0.39571  0.9021
## agecat(45,50]       0.64069 0.42663  0.9621
## agecat(50,55]       0.41905 0.26083  0.6732
## agecat(55,60]       0.40434 0.22901  0.7139
## agecat(60,65]       0.27744 0.15547  0.4951
## agecat(65,90]       0.29880 0.17042  0.5239
## meno                1.25527 0.84547  1.8637
## size20-50           1.52777 1.27373  1.8325
## size>50             1.89393 1.38723  2.5857
## grade3              1.59709 1.31053  1.9463
## nodescat(1,3]       2.61156 1.99884  3.4121
## nodescat(3,5]       3.18395 2.33586  4.3400
## nodescat(5,10]      5.84119 4.34156  7.8588
## nodescat(10,Inf]    7.95865 5.35125 11.8365
## pgrcat(20,40]       0.84483 0.61648  1.1578
## pgrcat(40,70]       0.83393 0.59060  1.1775
## pgrcat(70,100]      0.74873 0.50087  1.1193
```

```
## pgrcat(100,150]     0.65622 0.44902  0.9590
## pgrcat(150,Inf]     0.54924 0.42468  0.7103
## ercat(7,15]         0.90303 0.63133  1.2917
## ercat(15,40]        1.11752 0.81130  1.5393
## ercat(40,60]        1.06844 0.73911  1.5445
## ercat(60,80]        0.92705 0.60107  1.4298
## ercat(80,100]       1.19250 0.77901  1.8255
## ercat(100,140]      1.08056 0.72524  1.6100
## ercat(140,200]      0.98277 0.66284  1.4571
## ercat(200,Inf]      1.24717 0.91283  1.7040
## hormon              0.69872 0.51075  0.9559
```

Question 6

Use the "weighed estimating equations" approach (ipcw-glm) instead as sensitivity analysis. Is there a substantial difference in the results?

```
out.ipcw.glm <- logitIPCW(Event(time, status) ~ chemo +
                              yearcat + agecat + meno +
                              size + grade + nodescat +
                              pgrcat + ercat + hormon,
                          data=d,
                          time=5,
                          cens.model=~strata(chemo))
summary(out.ipcw.glm)
```

```
##
##     n events
##  2982   1181
##
##  2982 clusters
## coeffients:
##                       Estimate    Std.Err       2.5%      97.5% P-value
## (Intercept)        -0.5896757  0.2345597 -1.0494043 -0.1299471  0.0119
## chemo1             -0.4175124  0.1316862 -0.6756126 -0.1594123  0.0015
## yearcat(1985,1988] -0.0467509  0.1222456 -0.2863478  0.1928460  0.7021
## yearcat(1988,1990] -0.2128520  0.1318017 -0.4711786  0.0454745  0.1063
## yearcat(1990,1993] -0.1485064  0.1373808 -0.4177677  0.1207550  0.2797
## agecat(35,40]      -0.0547314  0.2224373 -0.4907004  0.3812376  0.8056
## agecat(40,45]      -0.5323565  0.2120503 -0.9479674 -0.1167456  0.0121
## agecat(45,50]      -0.4703787  0.2094970 -0.8809852 -0.0597722  0.0248
## agecat(50,55]      -0.8335423  0.2435916 -1.3109731 -0.3561115  0.0006
## agecat(55,60]      -0.9100126  0.2922004 -1.4827149 -0.3373102  0.0018
## agecat(60,65]      -1.2802628  0.2987695 -1.8658403 -0.6946853  0.0000
```

```
## agecat(65,90]      -0.9787684  0.2884564 -1.5441325 -0.4134042  0.0007
## meno                0.2047172  0.2020281 -0.1912505  0.6006849  0.3109
## size20-50           0.4186799  0.0933963  0.2356264  0.6017334  0.0000
## size>50             0.8264430  0.1699809  0.4932866  1.1595995  0.0000
## grade3              0.4504017  0.1014798  0.2515049  0.6492984  0.0000
## nodescat(1,3]       0.9755229  0.1393768  0.7023494  1.2486964  0.0000
## nodescat(3,5]       1.1034291  0.1588341  0.7921200  1.4147383  0.0000
## nodescat(5,10]      1.9327967  0.1608224  1.6175907  2.2480028  0.0000
## nodescat(10,Inf]    2.1270884  0.2100404  1.7154168  2.5387601  0.0000
## pgrcat(20,40]      -0.1305165  0.1633472 -0.4506711  0.1896381  0.4243
## pgrcat(40,70]      -0.2536500  0.1737936 -0.5942791  0.0869791  0.1444
## pgrcat(70,100]     -0.3094388  0.2078382 -0.7167943  0.0979166  0.1365
## pgrcat(100,150]    -0.4337293  0.1982343 -0.8222615 -0.0451972  0.0287
## pgrcat(150,Inf]    -0.5940668  0.1339048 -0.8565155 -0.3316181  0.0000
## ercat(7,15]        -0.0677754  0.1875739 -0.4354135  0.2998627  0.7179
## ercat(15,40]        0.1150027  0.1649508 -0.2082950  0.4383004  0.4857
## ercat(40,60]        0.0988083  0.1896282 -0.2728562  0.4704728  0.6023
## ercat(60,80]       -0.0256277  0.2248799 -0.4663843  0.4151288  0.9093
## ercat(80,100]       0.1844186  0.2184602 -0.2437556  0.6125928  0.3986
## ercat(100,140]      0.1253549  0.2072780 -0.2809025  0.5316124  0.5453
## ercat(140,200]      0.0174126  0.2060851 -0.3865066  0.4213319  0.9327
## ercat(200,Inf]      0.2440421  0.1630151 -0.0754616  0.5635457  0.1344
## hormon             -0.3315344  0.1662067 -0.6572935 -0.0057752  0.0461
##
## exp(coeffients):
##                    Estimate    2.5%   97.5%
## (Intercept)         0.55451 0.35015  0.8781
## chemo1              0.65868 0.50884  0.8526
## yearcat(1985,1988]  0.95433 0.75100  1.2127
## yearcat(1988,1990]  0.80828 0.62427  1.0465
## yearcat(1990,1993]  0.86199 0.65852  1.1283
## agecat(35,40]       0.94674 0.61220  1.4641
## agecat(40,45]       0.58722 0.38753  0.8898
## agecat(45,50]       0.62477 0.41437  0.9420
## agecat(50,55]       0.43451 0.26956  0.7004
## agecat(55,60]       0.40252 0.22702  0.7137
## agecat(60,65]       0.27796 0.15477  0.4992
## agecat(65,90]       0.37577 0.21350  0.6614
## meno                1.22718 0.82593  1.8234
## size20-50           1.51995 1.26570  1.8253
## size>50             2.28518 1.63769  3.1887
## grade3              1.56894 1.28596  1.9142
## nodescat(1,3]       2.65255 2.01849  3.4858
## nodescat(3,5]       3.01449 2.20807  4.1154
## nodescat(5,10]      6.90881 5.04093  9.4688
```

```
## nodescat(10,Inf]    8.39040 5.55899 12.6640
## pgrcat(20,40]       0.87764 0.63720  1.2088
## pgrcat(40,70]       0.77596 0.55196  1.0909
## pgrcat(70,100]      0.73386 0.48832  1.1029
## pgrcat(100,150]     0.64809 0.43944  0.9558
## pgrcat(150,Inf]     0.55208 0.42464  0.7178
## ercat(7,15]         0.93447 0.64700  1.3497
## ercat(15,40]        1.12188 0.81197  1.5501
## ercat(40,60]        1.10385 0.76120  1.6008
## ercat(60,80]        0.97470 0.62727  1.5146
## ercat(80,100]       1.20252 0.78368  1.8452
## ercat(100,140]      1.13355 0.75510  1.7017
## ercat(140,200]      1.01757 0.67943  1.5240
## ercat(200,Inf]      1.27640 0.92732  1.7569
## hormon              0.71782 0.51825  0.9942
```

Question 7

Use standardization after logistic regression to perform the main analysis and estimate the marginal 5-year risk of recurrence for a patient randomized to chemotherapy versus that for of a patient randomized to no chemotherapy. We will use the same logistic regression model as above. What is the risk difference? What can we conclude?

```
ateFit <- binregATE(Event(time, status) ~ chemo +
                        yearcat + agecat + meno +
                        size + grade + nodescat +
                        pgrcat + ercat + hormon,
                    data=d,
                    time=5,
                    treat.model=chemo~1,
                    cens.model=~strata(chemo))
summary(ateFit)
```

```
##
##     n events
##  2982   1181
##
##  2982 clusters
## coeffients:
##                        Estimate    Std.Err       2.5%      97.5% P-value
## (Intercept)         -0.4763653  0.2329862 -0.9330100 -0.0197207  0.0409
## chemo1              -0.4471582  0.1300573 -0.7020659 -0.1922506  0.0006
## yearcat(1985,1988] -0.0027157  0.1211501 -0.2401654  0.2347341  0.9821
## yearcat(1988,1990] -0.2380638  0.1309914 -0.4948023  0.0186747  0.0692
```

```
## yearcat(1990,1993] -0.3769345  0.1357750 -0.6430487 -0.1108203  0.0055
## agecat(35,40]      -0.0347404  0.2223052 -0.4704505  0.4009698  0.8758
## agecat(40,45]      -0.5151493  0.2102257 -0.9271842 -0.1031144  0.0143
## agecat(45,50]      -0.4452827  0.2074767 -0.8519294 -0.0386359  0.0319
## agecat(50,55]      -0.8698223  0.2419036 -1.3439447 -0.3957000  0.0003
## agecat(55,60]      -0.9056431  0.2900704 -1.4741706 -0.3371155  0.0018
## agecat(60,65]      -1.2822907  0.2955249 -1.8615090 -0.7030725  0.0000
## agecat(65,90]      -1.2081233  0.2865130 -1.7696784 -0.6465683  0.0000
## meno                0.2274409  0.2016499 -0.1677857  0.6226675  0.2594
## size20-50           0.4238117  0.0927905  0.2419458  0.6056777  0.0000
## size>50             0.6386931  0.1588642  0.3273250  0.9500612  0.0001
## grade3              0.4682308  0.1009004  0.2704696  0.6659920  0.0000
## nodescat(1,3]       0.9599846  0.1364272  0.6925921  1.2273770  0.0000
## nodescat(3,5]       1.1581614  0.1580424  0.8484039  1.4679189  0.0000
## nodescat(5,10]      1.7649764  0.1513899  1.4682577  2.0616951  0.0000
## nodescat(10,Inf]    2.0743303  0.2025380  1.6773631  2.4712974  0.0000
## pgrcat(20,40]      -0.1686497  0.1607786 -0.4837700  0.1464706  0.2942
## pgrcat(40,70]      -0.1815683  0.1760374 -0.5265952  0.1634587  0.3023
## pgrcat(70,100]     -0.2894069  0.2051369 -0.6914677  0.1126540  0.1583
## pgrcat(100,150]    -0.4212891  0.1935995 -0.8007372 -0.0418411  0.0295
## pgrcat(150,Inf]    -0.5992785  0.1312317 -0.8564878 -0.3420691  0.0000
## ercat(7,15]        -0.1019293  0.1826311 -0.4598797  0.2560212  0.5768
## ercat(15,40]        0.1111791  0.1633941 -0.2090674  0.4314256  0.4962
## ercat(40,60]        0.0662287  0.1880233 -0.3022902  0.4347477  0.7247
## ercat(60,80]       -0.0756762  0.2210834 -0.5089917  0.3576393  0.7321
## ercat(80,100]       0.1761509  0.2172537 -0.2496585  0.6019604  0.4175
## ercat(100,140]      0.0775748  0.2034482 -0.3211764  0.4763259  0.7030
## ercat(140,200]     -0.0172817  0.2009566 -0.4111493  0.3765860  0.9315
## ercat(200,Inf]      0.2209565  0.1592335 -0.0911354  0.5330485  0.1653
## hormon             -0.3585008  0.1598937 -0.6718866 -0.0451150  0.0250
##
## exp(coeffients):
##                    Estimate    2.5%   97.5%
## (Intercept)         0.62104 0.39337  0.9805
## chemo1              0.63944 0.49556  0.8251
## yearcat(1985,1988]  0.99729 0.78650  1.2646
## yearcat(1988,1990]  0.78815 0.60969  1.0189
## yearcat(1990,1993]  0.68596 0.52569  0.8951
## agecat(35,40]       0.96586 0.62472  1.4933
## agecat(40,45]       0.59741 0.39567  0.9020
## agecat(45,50]       0.64064 0.42659  0.9621
## agecat(50,55]       0.41903 0.26081  0.6732
## agecat(55,60]       0.40428 0.22897  0.7138
## agecat(60,65]       0.27740 0.15544  0.4951
## agecat(65,90]       0.29876 0.17039  0.5238
```

```
## meno                1.25538 0.84554  1.8639
## size20-50           1.52777 1.27373  1.8325
## size>50             1.89400 1.38725  2.5859
## grade3              1.59717 1.31058  1.9464
## nodescat(1,3]       2.61166 1.99889  3.4123
## nodescat(3,5]       3.18407 2.33592  4.3402
## nodescat(5,10]      5.84143 4.34166  7.8593
## nodescat(10,Inf]    7.95921 5.35143 11.8378
## pgrcat(20,40]       0.84480 0.61645  1.1577
## pgrcat(40,70]       0.83396 0.59061  1.1776
## pgrcat(70,100]      0.74871 0.50084  1.1192
## pgrcat(100,150]     0.65620 0.44900  0.9590
## pgrcat(150,Inf]     0.54921 0.42465  0.7103
## ercat(7,15]         0.90309 0.63136  1.2918
## ercat(15,40]        1.11760 0.81134  1.5395
## ercat(40,60]        1.06847 0.73912  1.5446
## ercat(60,80]        0.92712 0.60110  1.4299
## ercat(80,100]       1.19262 0.77907  1.8257
## ercat(100,140]      1.08066 0.72530  1.6101
## ercat(140,200]      0.98287 0.66289  1.4573
## ercat(200,Inf]      1.24727 0.91289  1.7041
## hormon              0.69872 0.51074  0.9559
##
## Average Treatment effects (G-formula) :
##             Estimate   Std.Err      2.5%     97.5% P-value
## treat0     0.423955  0.010456  0.403462  0.444449   0e+00
## treat1     0.338627  0.020489  0.298469  0.378785   0e+00
## treat:1-0 -0.085328  0.023614 -0.131612 -0.039045   3e-04
##
## Average Treatment effects (double robust) :
##             Estimate   Std.Err      2.5%     97.5% P-value
## treat0     0.423955  0.010456  0.403462  0.444449   0e+00
## treat1     0.338627  0.020489  0.298469  0.378785   0e+00
## treat:1-0 -0.085328  0.023614 -0.131612 -0.039045   3e-04
```

Question 8

Just for completeness, produce the corresponding unadjusted ("crude") results (risk difference with 95-CI and p-value) and check that they match the plot produced at **Question 1**.

```
# First we extract the estimates & SEs
fitKM.res <- summary(fitKM,time=5)
fitKM0 <- as.matrix(fitKM.res[fitKM.res$chemo==0,c("surv","se.surv")])
fitKM1 <- as.matrix(fitKM.res[fitKM.res$chemo==1,c("surv","se.surv")])
```

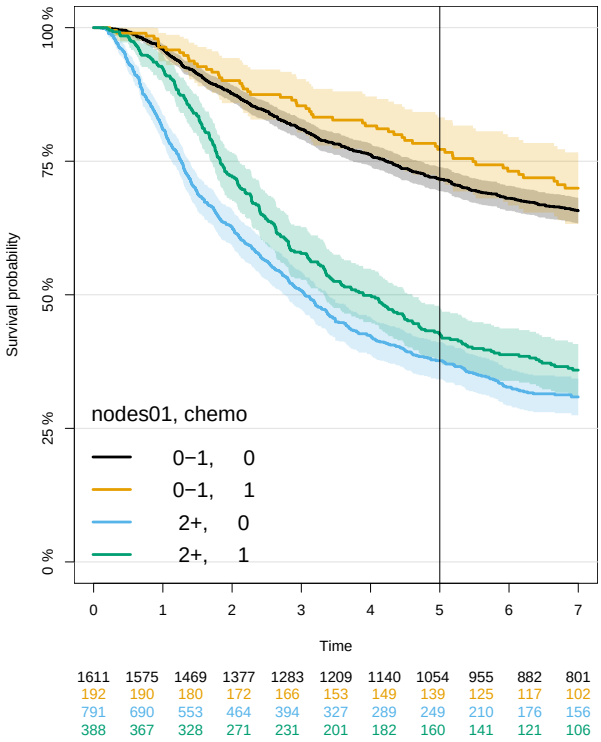
```r
# Second, we compute the risk difference
diffRisk <- (1-fitKM1[1,"surv"]) - (1-fitKM0[1,"surv"])
# Third, we compute the SE of the difference
seDiffRisk <- sqrt(fitKM1[1,"se.surv"]^2 + fitKM0[1,"se.surv"]^2)
# Then we compute the 95% CI
lowerDiffRisk <- diffRisk - qnorm(1-0.05/2)*seDiffRisk
upperDiffRisk <- diffRisk + qnorm(1-0.05/2)*seDiffRisk
# And the p-value (two-sided test)
pvalDiffRisk <- 2*(1-pnorm(abs(diffRisk/seDiffRisk)))
# Put all the results together and print
ResDiffRisk <- c(Diff=unname(diffRisk),
                 lower=unname(lowerDiffRisk),
                 upper=unname(upperDiffRisk),
                 p=unname(pvalDiffRisk))
ResDiffRisk
```

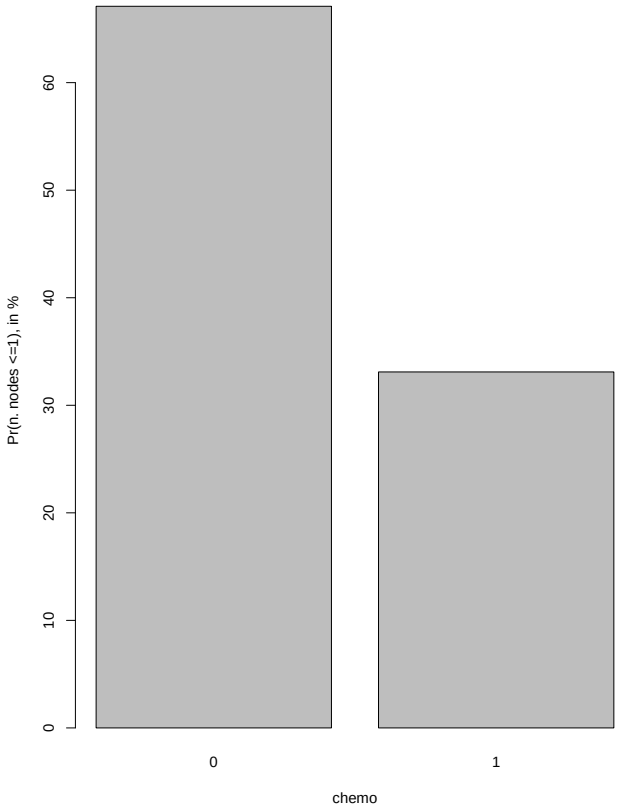
```
##        Diff       lower       upper           p
## 0.065294382 0.019799463 0.110789301 0.004909028
```

Question 9

To better understand the difference between the results of the adjusted and unadjusted analysis, we look again at the baseline Table. We see that there is some important imbalance for the number of positive lymph nodes. We therefore do two things. First, we plot the Kaplan-Meier curves for recurrence-free survival per treatment group within the two subgroups of patients: those with <2 positive lymph nodes and those with ≥ 2. Second, we compute the proportions of patients with < 2 positive nodes in the two treatment groups. What do we observe? Can we provide a tentative explanation for the difference between the adjusted and unadjusted results?

```r
d$nodes01 <- ifelse(d$nodes<2,"0-1","2+")
tabconf <- round(prop.table(table(nodes=d$nodes01,
                                  chemo=d$chemo),margin=2)*100,1)
fitKMnodes <- prodlim(Hist(time, status) ~ nodes01*chemo, data = d)
par(mfrow=c(1,2))
plot(fitKMnodes,xlim=c(0,7),legend.x="bottomleft")
abline(v=5)
barplot(tabconf[1,],ylab="Pr(n. nodes <=1), in %",xlab="chemo")
```


Survival probability
100 %
75 %
50 %
25 %
0 %
nodes01, chemo
0−1, 0
0−1, 1
2+, 0
2+, 1
0
1
2
3
4
5
6
7
Time
1611 1575 1469 1377 1283 1209 1140 1054 955 882 801
192 190 180 172 166 153 149 139 125 117 102
791 690 553 464 394 327 289 249 210 176 156
388 367 328 271 231 201 182 160 141 121 106


Pr(n. nodes <=1), in %
0
10
20
30
40
50
60
0
1
chemo