140.623.01 - Statistical Methods in Public Health III

Assignment 2: Survival in Primary Biliary Cirrhosis

Martin Skarzynski

March 8, 2018

Learning Objectives:

Students who successfully complete this section will be able to: - To evaluate whether the drug DPCA prolongs life in patients. - To identify baseline characteristics of patients which predict longer survival. - Analyze the survival time data (without grouping) by the Kaplan-Meier estimate of the survival function, the log- rank statistic, and Cox proportional hazards model. - Check the estimated model for its consistency with the observed data; in particular, check the proportional hazards assumption using the complementary log-log plot of the estimated survival function. - Summarize the findings for public health readers and document and archive the steps of the statistical analysis by creating a script file in R.

Data Set:

Between January 1974 and May 1984, a double-blinded randomized trial on patients with primary biliary cirrhosis (PBC) of the liver was conducted at the Mayo clinic. A total of 312 patients were randomized to either receive the drug D-penicillin (DPCA) or a placebo. Patients were followed until they died from PBC or until censoring, either because of administrative censoring (withdrawn alive at end of study), death not attributable to PBC, liver transplantation, or loss to follow-up. At baseline, a large number of clinical, biochemical, serological and histologic measurements were recorded on each patient. This data set is a subset of the original data, and includes information on each patient's time to death or censoring, treatment, age, gender, serum bilirubin, and histologic disease stage (1-4). The variables included in this dataset include: - case: unique patient ID number - sex: 0 = male, 1 = female (coded as "Female" and "Male" in the csv file rather than 0/1) - drug: 0 = placebo, 1 = DPCA - bil : serum bilirubin in mg/dl - survyr: time (in years) to death or censoring - death: indicator = 1 if patient died, 0 if censored - ageyr: age in years [continuous variable] - histo: histologic disease stage (1 - 4) [categorical variable] - agecat: age categories, coded as "< 45 yrs", "45 - 55 yrs", and ">= 55 yrs" Also included in the data set for your possible use are the following indicator (dummy) variables:

Age Indicators (indicator versions of agecat): - agegr_2: 1 if patient is 45-55 years old, 0 otherwise - agegr_3: 1 if patient is >= 55 years old, 0 otherwise Histologic Stage Indicators: - hstage2: 1 if patient is in Stage 2, 0 otherwise - hstage3: 1 if patient is in Stage 3, 0 otherwise - hstage4: 1 if patient is in Stage 4, 0 otherwise

The data are stored in the csv data set pbctrial.csv, which may be downloaded from the course website. ## Methods: Use the data set described above and the appropriate statistical analyses to address the specific learning objectives listed on the first page. Hints: The hints shown below are based on a dataset with the name pbcData, read in with the following code. In the following list of commands, if you want to look at differences by other variables than drug, you should change the variable name! Create a new .R file to type/run your commands so that you will have a record of your analysis.

```
library(readr)
pbcData = read_csv("pbctrial.csv")
```

```
## Parsed with column specification:
## cols(
##   case = col_integer(),
##   drug = col_integer(),
##   sex = col_character(),
```

```
##   bil = col_double(),
##   histo = col_integer(),
##   death = col_integer(),
##   survyr = col_double(),
##   `_st` = col_integer(),
##   `_d` = col_integer(),
##   `_t` = col_double(),
##   `_t0` = col_integer(),
##   ageyr = col_double(),
##   agecat = col_character(),
##   agegr_2 = col_integer(),
##   agegr_3 = col_integer(),
##   hstage2 = col_integer(),
##   hstage3 = col_integer(),
##   hstage4 = col_integer()
## )
```

a. Explore the data using descriptive statistics:

- table()
- prop.table()
- summary() etc

```
dim(pbcData)
```

```
## [1] 312  18
```

```
str(pbcData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    312 obs. of  18 variables:
##  $ case   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ drug   : int  1 1 1 1 0 0 0 0 1 0 ...
##  $ sex    : chr  "Female" "Female" "Male" "Female" ...
##  $ bil    : num  14.5 1.1 1.4 1.8 3.4 ...
##  $ histo  : int  4 3 4 4 3 3 3 3 2 4 ...
##  $ death  : int  1 0 1 1 0 1 0 1 1 1 ...
##  $ survyr : num  1.1 12.33 2.77 5.27 4.12 ...
##  $ _st    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ _d     : int  1 0 1 1 0 1 0 1 1 1 ...
##  $ _t     : num  1.1 12.33 2.77 5.27 4.12 ...
##  $ _t0    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ageyr  : num  58.8 56.5 70.1 54.8 38.1 ...
##  $ agecat : chr  ">= 55 yrs" ">= 55 yrs" ">= 55 yrs" "45 - 55 yrs" ...
##  $ agegr_2: int  0 0 0 1 0 0 0 1 0 0 ...
##  $ agegr_3: int  1 1 1 0 0 1 1 0 0 1 ...
##  $ hstage2: int  0 0 0 0 0 0 0 0 1 0 ...
##  $ hstage3: int  0 1 0 0 1 1 1 1 0 0 ...
##  $ hstage4: int  1 0 1 1 0 0 0 0 0 1 ...
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 18
##   .. ..$ case   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ drug   : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ sex    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
```

```
##   .. ..$ bil    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ histo  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ death  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ survyr : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ _st    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ _d     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ _t     : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ _t0    : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ ageyr  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_double" "collector"
##   .. ..$ agecat : list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ agegr_2: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ agegr_3: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ hstage2: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ hstage3: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ hstage4: list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```
summary(pbcData)
```

```
##       case             drug            sex                 bil        
##  Min.   :  1.00   Min.   :0.0000   Length:312         Min.   : 0.300  
##  1st Qu.: 78.75   1st Qu.:0.0000   Class :character   1st Qu.: 0.800  
##  Median :156.50   Median :1.0000   Mode  :character   Median : 1.350  
##  Mean   :156.50   Mean   :0.5064                      Mean   : 3.256  
##  3rd Qu.:234.25   3rd Qu.:1.0000                      3rd Qu.: 3.425  
##  Max.   :312.00   Max.   :1.0000                      Max.   :28.000  
##      histo           death            survyr             _st        
##  Min.   :1.000   Min.   :0.0000   Min.   : 0.1123   Min.   :1  
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.: 3.2630   1st Qu.:1  
##  Median :3.000   Median :0.0000   Median : 5.0397   Median :1  
##  Mean   :3.032   Mean   :0.4006   Mean   : 5.4969   Mean   :1  
##  3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.: 7.3897   3rd Qu.:1  
##  Max.   :4.000   Max.   :1.0000   Max.   :12.4822   Max.   :1  
##        _d               _t               _t0        ageyr      
##  Min.   :0.0000   Min.   : 0.1123   Min.   :0   Min.   :26.30  
##  1st Qu.:0.0000   1st Qu.: 3.2630   1st Qu.:0   1st Qu.:42.27  
##  Median :0.0000   Median : 5.0397   Median :0   Median :49.83  
##  Mean   :0.4006   Mean   : 5.4969   Mean   :0   Mean   :50.05
```

```
##  3rd Qu.:1.0000   3rd Qu.: 7.3897   3rd Qu.:0   3rd Qu.:56.75
##  Max.   :1.0000   Max.   :12.4822   Max.   :0   Max.   :78.49
##     agecat            agegr_2          agegr_3          hstage2
##  Length:312         Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Median :0.0000   Median :0.0000   Median :0.0000
##                     Mean   :0.3237   Mean   :0.3365   Mean   :0.2147
##                     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##                     Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     hstage3          hstage4
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.3846   Mean   :0.3494
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
```

```
library(purrr, help)
map(pbcData, class)
```

```
## $case
## [1] "integer"
##
## $drug
## [1] "integer"
##
## $sex
## [1] "character"
##
## $bil
## [1] "numeric"
##
## $histo
## [1] "integer"
##
## $death
## [1] "integer"
##
## $survyr
## [1] "numeric"
##
## $`_st`
## [1] "integer"
##
## $`_d`
## [1] "integer"
##
## $`_t`
## [1] "numeric"
##
## $`_t0`
## [1] "integer"
##
## $ageyr
## [1] "numeric"
```

```
##
## $agecat
## [1] "character"
##
## $agegr_2
## [1] "integer"
##
## $agegr_3
## [1] "integer"
##
## $hstage2
## [1] "integer"
##
## $hstage3
## [1] "integer"
##
## $hstage4
## [1] "integer"
```

```
pbcData$histo <- as.factor(pbcData$histo)
pbcData$agecat <- as.factor(pbcData$agecat)
map(pbcData, class)
```

```
## $case
## [1] "integer"
##
## $drug
## [1] "integer"
##
## $sex
## [1] "character"
##
## $bil
## [1] "numeric"
##
## $histo
## [1] "factor"
##
## $death
## [1] "integer"
##
## $survyr
## [1] "numeric"
##
## $`_st`
## [1] "integer"
##
## $`_d`
## [1] "integer"
##
## $`_t`
## [1] "numeric"
##
## $`_t0`
## [1] "integer"
```

```
##
## $ageyr
## [1] "numeric"
##
## $agecat
## [1] "factor"
##
## $agegr_2
## [1] "integer"
##
## $agegr_3
## [1] "integer"
##
## $hstage2
## [1] "integer"
##
## $hstage3
## [1] "integer"
##
## $hstage4
## [1] "integer"
```

```
round(prop.table(table(pbcData[c("death", "drug", "sex")])), 3)
```

```
## , , sex = Female
##
##      drug
## death     0     1
##     0 0.279 0.276
##     1 0.167 0.163
##
## , , sex = Male
##
##      drug
## death     0     1
##     0 0.022 0.022
##     1 0.026 0.045
```

b. Define a survival object, defining the time variable (survyr) and the event (death == 1). To do this, you must first install and load the "survival" package:

```
# install.packages("survival")
library(survival)


## only run this the first time

pbcData$SurvObj = with(pbcData, Surv(survyr, death == 1))
```

c. Explore differences in time to death by different baseline variables using graphs and complementary log-log plots.

```
# estimate survival curves for entire sample
km.overall = survfit(SurvObj ~ 1, data = pbcData,
type="kaplan-meier", conf.type="log-log")
km.overall
```

```
## Call: survfit(formula = SurvObj ~ 1, data = pbcData, type = "kaplan-meier",
##     conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
##  312.00  125.00    9.30    8.45   10.52
```

```
summary(km.overall)
```

```
## Call: survfit(formula = SurvObj ~ 1, data = pbcData, type = "kaplan-meier",
##     conf.type = "log-log")
##
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.112    312       1    0.997 0.00320        0.977        1.000
##   0.140    311       1    0.994 0.00452        0.975        0.998
##   0.195    310       1    0.990 0.00552        0.970        0.997
##   0.211    309       1    0.987 0.00637        0.966        0.995
##   0.301    308       1    0.984 0.00711        0.962        0.993
##   0.356    307       1    0.981 0.00778        0.958        0.991
##   0.359    306       1    0.978 0.00838        0.954        0.989
##   0.384    305       1    0.974 0.00895        0.949        0.987
##   0.490    304       1    0.971 0.00948        0.945        0.985
##   0.510    303       1    0.968 0.00997        0.941        0.983
##   0.523    302       1    0.965 0.01044        0.937        0.980
##   0.542    301       1    0.962 0.01089        0.933        0.978
##   0.567    300       1    0.958 0.01131        0.929        0.976
##   0.592    299       1    0.955 0.01172        0.925        0.973
##   0.611    298       1    0.952 0.01211        0.922        0.971
##   0.723    297       2    0.946 0.01285        0.914        0.966
##   0.833    295       1    0.942 0.01320        0.910        0.963
##   0.879    294       1    0.939 0.01354        0.906        0.961
##   0.893    293       1    0.936 0.01387        0.902        0.958
##   0.915    292       1    0.933 0.01418        0.899        0.956
##   0.953    291       1    0.929 0.01449        0.895        0.953
##   1.063    290       1    0.926 0.01479        0.891        0.950
##   1.096    289       1    0.923 0.01509        0.887        0.948
##   1.260    288       1    0.920 0.01537        0.884        0.945
##   1.411    287       1    0.917 0.01565        0.880        0.942
##   1.504    285       1    0.913 0.01592        0.876        0.940
##   1.512    284       1    0.910 0.01619        0.873        0.937
##   1.636    283       1    0.907 0.01644        0.869        0.934
##   1.674    282       1    0.904 0.01670        0.865        0.932
##   1.844    281       1    0.901 0.01695        0.862        0.929
##   1.901    280       1    0.897 0.01719        0.858        0.926
##   1.940    279       1    0.894 0.01742        0.854        0.924
##   2.008    277       1    0.891 0.01766        0.851        0.921
##   2.055    275       1    0.888 0.01789        0.847        0.918
##   2.088    274       1    0.884 0.01811        0.843        0.915
##   2.107    273       1    0.881 0.01833        0.840        0.912
##   2.153    272       1    0.878 0.01855        0.836        0.910
##   2.164    270       1    0.875 0.01877        0.833        0.907
##   2.184    269       1    0.871 0.01898        0.829        0.904
##   2.189    268       1    0.868 0.01918        0.825        0.901
##   2.258    267       1    0.865 0.01938        0.822        0.898
##   2.329    264       1    0.862 0.01958        0.818        0.896
##   2.337    263       1    0.858 0.01978        0.814        0.893
```

```
##   2.353   262       1    0.855 0.01998        0.811        0.890
##   2.438   260       1    0.852 0.02017        0.807        0.887
##   2.477   258       1    0.849 0.02036        0.804        0.884
##   2.548   257       1    0.845 0.02055        0.800        0.881
##   2.584   255       1    0.842 0.02073        0.796        0.878
##   2.660   254       1    0.839 0.02091        0.793        0.875
##   2.668   253       1    0.835 0.02109        0.789        0.872
##   2.685   252       1    0.832 0.02127        0.785        0.869
##   2.737   250       1    0.829 0.02144        0.782        0.866
##   2.740   249       1    0.825 0.02161        0.778        0.863
##   2.773   248       1    0.822 0.02178        0.775        0.860
##   2.841   246       1    0.819 0.02194        0.771        0.857
##   2.951   244       1    0.815 0.02211        0.767        0.854
##   2.959   243       1    0.812 0.02227        0.764        0.851
##   2.967   242       1    0.809 0.02243        0.760        0.848
##   3.156   239       1    0.805 0.02259        0.756        0.845
##   3.192   237       1    0.802 0.02275        0.753        0.842
##   3.205   236       1    0.798 0.02291        0.749        0.839
##   3.263   235       2    0.792 0.02321        0.742        0.833
##   3.321   233       1    0.788 0.02336        0.738        0.830
##   3.334   230       1    0.785 0.02350        0.734        0.827
##   3.384   227       1    0.781 0.02365        0.731        0.824
##   3.553   222       1    0.778 0.02381        0.727        0.820
##   3.699   214       1    0.774 0.02397        0.723        0.817
##   3.715   213       1    0.771 0.02413        0.719        0.814
##   3.726   212       1    0.767 0.02429        0.715        0.811
##   3.871   206       1    0.763 0.02446        0.711        0.807
##   3.910   203       1    0.759 0.02462        0.707        0.804
##   3.929   201       1    0.756 0.02479        0.703        0.800
##   3.956   198       1    0.752 0.02496        0.699        0.797
##   4.074   193       1    0.748 0.02513        0.695        0.793
##   4.088   192       1    0.744 0.02530        0.690        0.790
##   4.208   189       1    0.740 0.02547        0.686        0.786
##   4.318   184       1    0.736 0.02565        0.682        0.783
##   4.540   178       1    0.732 0.02583        0.677        0.779
##   4.608   175       1    0.728 0.02602        0.673        0.775
##   4.630   174       2    0.719 0.02639        0.664        0.767
##   4.770   169       1    0.715 0.02657        0.659        0.764
##   4.893   162       1    0.711 0.02677        0.654        0.760
##   5.005   159       1    0.706 0.02697        0.650        0.755
##   5.060   156       1    0.702 0.02718        0.645        0.751
##   5.274   151       1    0.697 0.02739        0.640        0.747
##   5.630   141       1    0.692 0.02764        0.634        0.743
##   5.701   140       1    0.687 0.02788        0.629        0.738
##   5.726   139       1    0.682 0.02812        0.624        0.734
##   5.767   138       1    0.677 0.02834        0.618        0.729
##   6.093   127       1    0.672 0.02862        0.612        0.725
##   6.181   123       1    0.667 0.02890        0.606        0.720
##   6.268   121       1    0.661 0.02918        0.600        0.715
##   6.293   119       1    0.655 0.02946        0.594        0.710
##   6.537   110       1    0.649 0.02979        0.588        0.704
##   6.575   109       1    0.644 0.03011        0.581        0.699
##   6.627   108       1    0.638 0.03041        0.575        0.694
##   6.756   103       1    0.631 0.03074        0.568        0.688
```

```
##   6.858    100       1    0.625 0.03108        0.561        0.683
##   6.959     96       1    0.619 0.03143        0.554        0.677
##   7.077     88       1    0.612 0.03185        0.546        0.671
##   7.118     87       1    0.604 0.03225        0.538        0.664
##   7.367     80       1    0.597 0.03272        0.530        0.658
##   7.586     76       1    0.589 0.03322        0.521        0.651
##   7.660     74       1    0.581 0.03371        0.512        0.644
##   7.800     71       1    0.573 0.03421        0.503        0.637
##   8.455     60       1    0.563 0.03495        0.492        0.629
##   8.466     59       1    0.554 0.03564        0.481        0.620
##   8.685     53       1    0.543 0.03646        0.469        0.612
##   8.827     52       1    0.533 0.03723        0.457        0.603
##   8.888     50       1    0.522 0.03798        0.445        0.594
##   8.992     48       1    0.511 0.03872        0.433        0.584
##   9.200     45       1    0.500 0.03949        0.420        0.574
##   9.301     43       1    0.488 0.04025        0.407        0.564
##   9.392     41       1    0.476 0.04099        0.394        0.554
##   9.438     40       1    0.465 0.04166        0.381        0.544
##   9.792     37       1    0.452 0.04238        0.368        0.533
##   9.819     34       1    0.439 0.04317        0.353        0.521
##  10.307     30       1    0.424 0.04414        0.337        0.509
##  10.518     27       1    0.408 0.04522        0.319        0.495
##  10.556     25       1    0.392 0.04626        0.302        0.481
##  11.175     17       1    0.369 0.04895        0.274        0.464
##  11.482     13       1    0.341 0.05278        0.240        0.444
```

```
# estimate survival curves for drug group
km.drug = survfit(SurvObj ~ drug, data = pbcData,
type="kaplan-meier", conf.type="log-log")
km.drug
```

```
## Call: survfit(formula = SurvObj ~ drug, data = pbcData, type = "kaplan-meier",
##     conf.type = "log-log")
##
##          n events median 0.95LCL 0.95UCL
## drug=0 154     60   9.39    8.47    10.6
## drug=1 158     65   8.99    6.96    11.5
```

```
summary(km.drug)
```

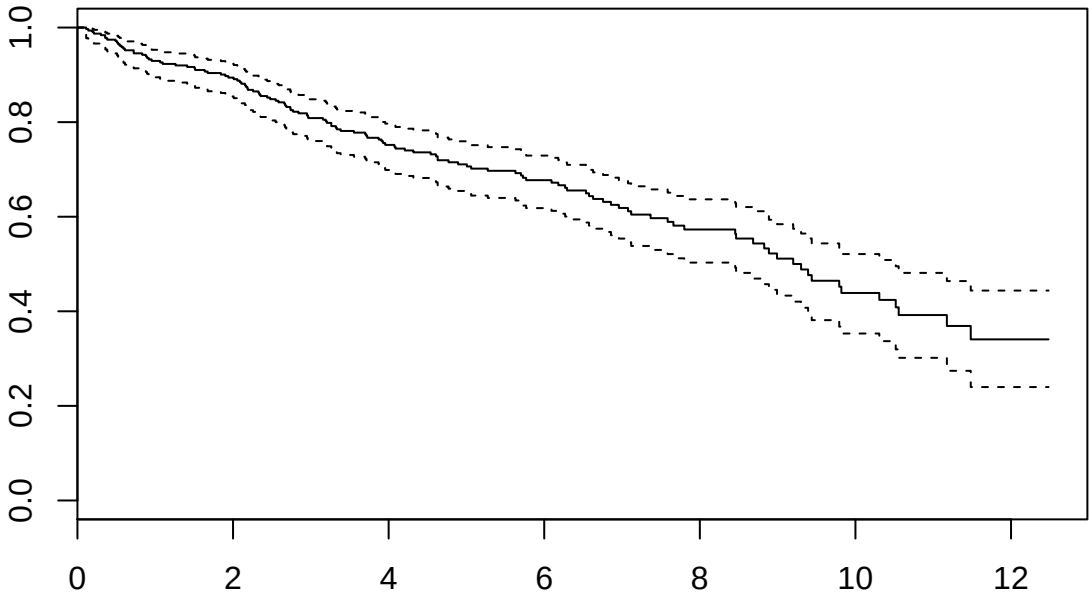
```
## Call: survfit(formula = SurvObj ~ drug, data = pbcData, type = "kaplan-meier",
##     conf.type = "log-log")
##
##                 drug=0
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.140    154       1    0.994 0.00647        0.955        0.999
##   0.211    153       1    0.987 0.00912        0.949        0.997
##   0.301    152       1    0.981 0.01114        0.941        0.994
##   0.356    151       1    0.974 0.01282        0.932        0.990
##   0.510    150       1    0.968 0.01428        0.924        0.986
##   0.523    149       1    0.961 0.01559        0.915        0.982
##   0.567    148       1    0.955 0.01679        0.907        0.978
##   0.592    147       1    0.948 0.01788        0.899        0.974
##   0.723    146       2    0.935 0.01986        0.883        0.965
##   0.833    144       1    0.929 0.02075        0.875        0.960
```

```
##   0.879    143       1    0.922 0.02160        0.867        0.955
##   0.893    142       1    0.916 0.02240        0.859        0.950
##   1.260    141       1    0.909 0.02317        0.851        0.945
##   1.504    140       1    0.903 0.02389        0.844        0.940
##   1.512    139       1    0.896 0.02459        0.836        0.935
##   1.636    138       1    0.890 0.02525        0.828        0.930
##   1.674    137       1    0.883 0.02589        0.821        0.925
##   1.940    136       1    0.877 0.02650        0.813        0.919
##   2.008    135       1    0.870 0.02709        0.806        0.914
##   2.107    134       1    0.864 0.02765        0.799        0.909
##   2.153    133       1    0.857 0.02820        0.791        0.904
##   2.164    131       1    0.851 0.02873        0.784        0.898
##   2.184    130       1    0.844 0.02925        0.776        0.893
##   2.329    128       1    0.837 0.02975        0.769        0.887
##   2.337    127       1    0.831 0.03024        0.762        0.882
##   2.353    126       1    0.824 0.03071        0.754        0.876
##   2.438    125       1    0.818 0.03116        0.747        0.870
##   2.548    124       1    0.811 0.03160        0.740        0.865
##   2.584    123       1    0.804 0.03203        0.732        0.859
##   2.668    122       1    0.798 0.03244        0.725        0.853
##   2.959    118       1    0.791 0.03286        0.718        0.847
##   3.192    115       1    0.784 0.03328        0.710        0.841
##   3.321    114       1    0.777 0.03370        0.703        0.836
##   3.334    111       1    0.770 0.03411        0.695        0.829
##   3.715    103       1    0.763 0.03459        0.687        0.823
##   3.871    101       1    0.755 0.03506        0.678        0.816
##   3.910     98       1    0.748 0.03554        0.670        0.810
##   3.956     95       1    0.740 0.03603        0.661        0.803
##   4.074     93       1    0.732 0.03651        0.652        0.796
##   4.208     91       1    0.724 0.03698        0.644        0.789
##   4.893     79       1    0.715 0.03763        0.633        0.781
##   5.060     76       1    0.705 0.03829        0.623        0.773
##   5.726     69       1    0.695 0.03908        0.611        0.764
##   6.627     56       1    0.683 0.04030        0.596        0.754
##   6.756     53       1    0.670 0.04155        0.581        0.744
##   6.858     51       1    0.657 0.04276        0.566        0.733
##   7.586     40       1    0.640 0.04473        0.545        0.720
##   7.660     38       1    0.623 0.04662        0.525        0.707
##   7.800     35       1    0.605 0.04857        0.503        0.693
##   8.466     32       1    0.587 0.05060        0.481        0.678
##   8.685     29       1    0.566 0.05275        0.457        0.662
##   8.888     28       1    0.546 0.05460        0.433        0.646
##   9.200     26       1    0.525 0.05640        0.409        0.628
##   9.301     24       1    0.503 0.05814        0.385        0.610
##   9.392     22       1    0.480 0.05983        0.360        0.591
##   9.438     21       1    0.457 0.06119        0.335        0.572
##  10.307     15       1    0.427 0.06427        0.300        0.548
##  10.518     13       1    0.394 0.06719        0.264        0.522
##  10.556     12       1    0.361 0.06916        0.230        0.494
##
##                drug=1
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.112    158       1    0.994 0.00631        0.956        0.999
##   0.195    157       1    0.987 0.00889        0.950        0.997
```

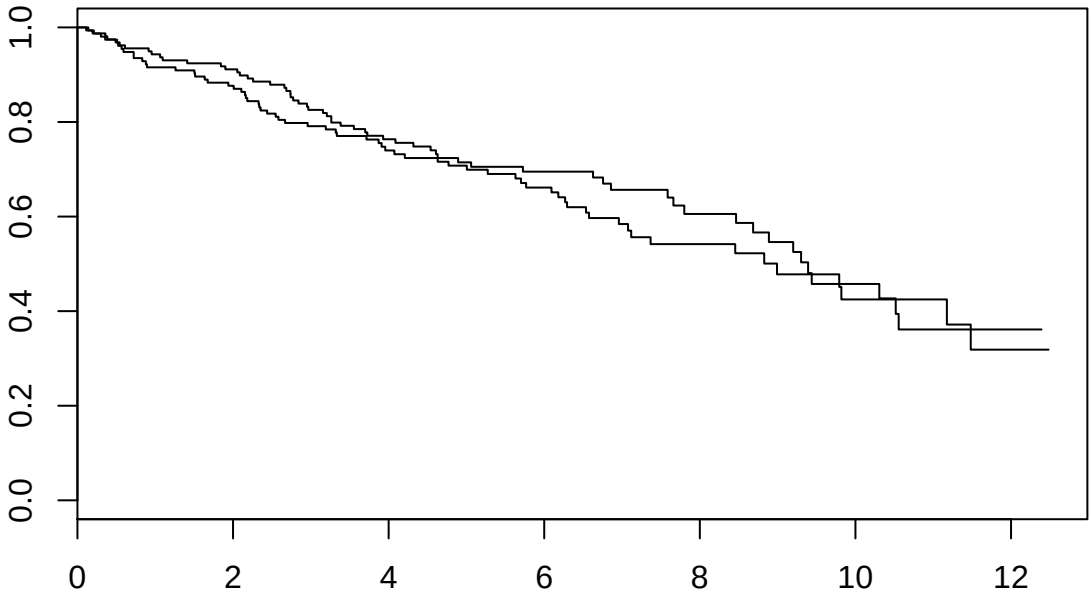
```
##   0.359    156       1    0.981 0.01086        0.942        0.994
##   0.384    155       1    0.975 0.01250        0.934        0.990
##   0.490    154       1    0.968 0.01393        0.926        0.987
##   0.542    153       1    0.962 0.01521        0.917        0.983
##   0.611    152       1    0.956 0.01637        0.909        0.979
##   0.915    151       1    0.949 0.01744        0.901        0.974
##   0.953    150       1    0.943 0.01844        0.893        0.970
##   1.063    149       1    0.937 0.01937        0.886        0.965
##   1.096    148       1    0.930 0.02025        0.878        0.961
##   1.411    147       1    0.924 0.02108        0.870        0.956
##   1.844    145       1    0.918 0.02187        0.862        0.951
##   1.901    144       1    0.911 0.02263        0.855        0.946
##   2.055    141       1    0.905 0.02337        0.847        0.942
##   2.088    140       1    0.898 0.02408        0.839        0.936
##   2.189    139       1    0.892 0.02476        0.832        0.931
##   2.258    138       1    0.885 0.02541        0.824        0.926
##   2.477    134       1    0.879 0.02607        0.817        0.921
##   2.660    132       1    0.872 0.02671        0.809        0.916
##   2.685    131       1    0.866 0.02732        0.801        0.910
##   2.737    130       1    0.859 0.02791        0.794        0.905
##   2.740    129       1    0.852 0.02848        0.786        0.899
##   2.773    128       1    0.846 0.02902        0.778        0.894
##   2.841    127       1    0.839 0.02955        0.771        0.888
##   2.951    126       1    0.832 0.03005        0.763        0.883
##   2.967    125       1    0.826 0.03054        0.756        0.877
##   3.156    124       1    0.819 0.03101        0.749        0.871
##   3.205    122       1    0.812 0.03148        0.741        0.866
##   3.263    121       2    0.799 0.03236        0.726        0.854
##   3.384    117       1    0.792 0.03279        0.719        0.848
##   3.553    114       1    0.785 0.03323        0.711        0.842
##   3.699    111       1    0.778 0.03368        0.703        0.836
##   3.726    110       1    0.771 0.03411        0.695        0.830
##   3.929    105       1    0.764 0.03456        0.687        0.823
##   4.088    100       1    0.756 0.03505        0.679        0.817
##   4.318     97       1    0.748 0.03554        0.670        0.810
##   4.540     93       1    0.740 0.03606        0.661        0.803
##   4.608     92       1    0.732 0.03655        0.653        0.796
##   4.630     91       2    0.716 0.03748        0.635        0.782
##   4.770     87       1    0.708 0.03794        0.626        0.775
##   5.005     82       1    0.699 0.03845        0.616        0.767
##   5.274     78       1    0.690 0.03899        0.607        0.759
##   5.630     72       1    0.681 0.03960        0.596        0.751
##   5.701     71       1    0.671 0.04019        0.585        0.743
##   5.767     70       1    0.661 0.04074        0.575        0.734
##   6.093     65       1    0.651 0.04137        0.564        0.725
##   6.181     63       1    0.641 0.04198        0.552        0.716
##   6.268     61       1    0.630 0.04259        0.541        0.707
##   6.293     60       1    0.620 0.04315        0.529        0.698
##   6.537     54       1    0.608 0.04385        0.517        0.688
##   6.575     53       1    0.597 0.04450        0.504        0.678
##   6.959     47       1    0.584 0.04533        0.490        0.667
##   7.077     42       1    0.570 0.04634        0.474        0.655
##   7.118     41       1    0.556 0.04725        0.459        0.643
##   7.367     38       1    0.542 0.04822        0.443        0.631
```

```
##    8.455     28       1    0.522 0.05023        0.420        0.615
##    8.827     24       1    0.501 0.05264        0.394        0.598
##    8.992     22       1    0.478 0.05495        0.367        0.580
##    9.792     18       1    0.451 0.05795        0.336        0.560
##    9.819     17       1    0.425 0.06032        0.306        0.539
##   11.175      8       1    0.372 0.07247        0.233        0.510
##   11.482      7       1    0.319 0.07922        0.173        0.474
```

```
# plot km curves
plot(km.overall)
```




```
plot(km.drug)
```

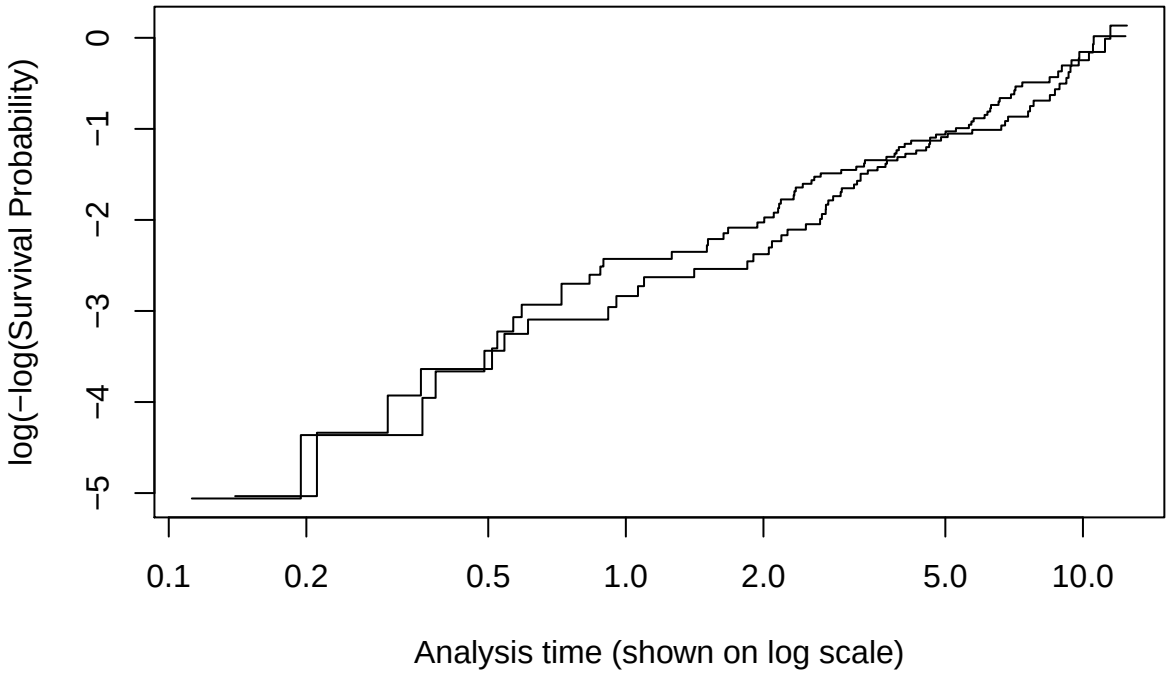



```
# log rank test for equality of survivor functions
survdiff(SurvObj ~ drug, data=pbcData)
```

```
## Call:
```

```
## survdiff(formula = SurvObj ~ drug, data = pbcData)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## drug=0 154       60     61.8    0.0513     0.102
## drug=1 158       65     63.2    0.0502     0.102
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.75
```

```
# complimentary log-log plot
plot(km.drug, fun="cloglog", ylab="log(-log(Survival Probability)",
xlab="Analysis time (shown on log scale)")
```



d. Fit several Cox proportional hazards regression models to the ungrouped survival data:

```
model1 = coxph(SurvObj ~ drug, data = pbcData)
summary(model1)
```

```
## Call:
## coxph(formula = SurvObj ~ drug, data = pbcData)
##
##   n= 312, number of events= 125
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## drug 0.05722   1.05889  0.17916 0.319    0.749
##
##      exp(coef) exp(-coef) lower .95 upper .95
## drug     1.059     0.9444    0.7453     1.504
##
## Concordance= 0.499  (se = 0.025 )
## Rsquare= 0   (max possible= 0.983 )
## Likelihood ratio test= 0.1  on 1 df,   p=0.7494
## Wald test            = 0.1  on 1 df,   p=0.7494
## Score (logrank) test = 0.1  on 1 df,   p=0.7494
```

```r
model2 = coxph(SurvObj ~ sex + bil + histo, data = pbcData)
summary(model2)
```

```
## Call:
## coxph(formula = SurvObj ~ sex + bil + histo, data = pbcData)
##
##   n= 312, number of events= 125
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## sexMale  0.64275   1.90171  0.23926  2.686  0.00722 **
## bil      0.15149   1.16357  0.01424 10.637  < 2e-16 ***
## histo2   1.64339   5.17269  1.03376  1.590  0.11190
## histo3   2.03122   7.62340  1.01631  1.999  0.04565 *
## histo4   2.90689  18.29988  1.01216  2.872  0.00408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sexMale     1.902    0.52584     1.190     3.040
## bil         1.164    0.85943     1.132     1.197
## histo2      5.173    0.19332     0.682    39.233
## histo3      7.623    0.13118     1.040    55.877
## histo4     18.300    0.05465     2.517   133.045
##
## Concordance= 0.812  (se = 0.029 )
## Rsquare= 0.347   (max possible= 0.983 )
## Likelihood ratio test= 133.2  on 5 df,   p=0
## Wald test            = 149.2  on 5 df,   p=0
## Score (logrank) test = 218.8  on 5 df,   p=0
```

e. Save your R script file that documents and archives the steps of your statistical analysis. This file will make your analysis "reproducible."

f. Summarize your findings in a brief report (less than two pages with at most one table and one figure) as if for a biomedical/public health journal. A suggested format is:

- Introduction - a few sentences about the research question(s)
- Data description - simple tabulations describing patient characteristics
- Results from multiple models that address question(s) (e.g., bivariate and multivariable)
- Graphical display that presents evidence in the data relevant to your scientific question.

Introduction

Between January 1974 and May 1984 a double-blinded randomized trial on patients with primary biliary cirrhosis (PBC) of the liver was conducted at the Mayo clinic. A total of 312 patients were randomized to either receive the drug D-penicillimin (DPCA), or a placebo. Patients were followed until they died from PBC, or until censoring, either because of administrative censoring (withdrawn alive at end of study), death not attributable to PBC, liver transplantation, or loss to follow-up. At baseline clinical, biochemical, serological and histologic measurements were recorded on each patient. A sub-study was undertaken to test for increased survival amongst patients on the new treatment, and to investigate the association between survival and patients' age, gender, histologic stage of disease, and serum bilirubin level. The research question that I will try to answer in this report is whether D-penicillin (DPCA), the drug tested in the PBC trial, provided any benefit for the patient population as a whole (n=312) and for sub-groups based on sex, age and disease stage. I hypothesize that the drug effect will not be different between the 3 age categories, but will depend on disease stage. In other wrods, I expect that there will be differences in time to death between the 4 disease

stages, specifically that more advanced disease will be more difficult to treat, which will result in a shorter time to event. I will also assess whether bilirubin is a prognostic marker and whether drug benefit will differ among men versus women.

Data description

There are a total of 312 patients and the median survival time was around 5 years. As for patient characteristics, the representation across age categories and disease stages appears to spread relatively evenly. The `age` and `survyr` variable appear to be normally distributed with a slight leftward skew. Interestingly, bilirubin is skewed highly to the left indicating that there are outliers with high bilirubin values.

Methods

Descriptive statistics were calculated to investigate sample characteristics. Kaplan-Meier estimates of the survivor functions for various sample sub-groupings were calculated. Simple Cox regression models were used to evaluate univariate associations between patient characteristics and survival. Multivariable Cox regression was used to examine the association between survival and multiple patient characteristics simultaneously. Serum bilirubin was the only continuous covariate in the regression models. Age was modeled as a categorical variable, based on tertiles in the sample, to allow for a non-linear relationship between age and the loghazard of death. Both Wald and likelihood ratio methods were used to test for the statistical significance of covariates in the final multiple proportional hazards model. Only predictors achieving statistical significance ($\alpha = .05$) were included in the final multivariable model.

Study Enrollees

The sample consists of 312 patients with primary biliary cirrhosis enrolled from 1974 to 1984 at the Mayo Clinic in Rochester, MN. The sample is majority female (276 patients, 88%) with only 36 male patients (12%). The average patient age at enrollment was 50 years, and the sample age range was from 26 to 78 years. The majority (75%) of the patients were in a later stage of the disease (Histologic Stage 3 or 4) at the time of enrollment. Average serum bilirubin level among participants at time of enrollment was 3.3 mg/dl. At the time of this analysis, 125 patients (40%) had died from causes related to primary biliary cirrhosis. Results Patients in the drug group had 6% greater hazard ("risk") of death than those in the placebo group, but this result was not statistically significant (95% CI, -25% - 50%, p > .05).

Serum bilirubin level, patients age, and histologic stage of disease all had statistically significant ($\alpha = .05$) positive univariate associations with the hazard of death. Males had 62% higher risk of death than females (95% CI 2% - 158%, p = .04). In a multivariable analysis, serum bilirubin level, gender, and histologic stage of disease were found to have statistically significant associations with patient survival. The hazard ratio associated with a 1 mg/dl increase in serum bilirubin level was 1.16 (95% CI 1.13 - 1.19), indicating that a patient's risk of death increases by 16% for each 1 mg/dl increase in serum bilirubin after adjustment for gender, disease stage and age. The hazard ratio of death for males relative to females was 1.70 (95% CI 1.05 - 2.74), indicating that males had a 70% increase in the hazard of death compared to otherwise similar females. Those patients in the highest stage (stage 4) of disease had greater than 14 times the adjusted risk (95% CI 2.01 - 107.40) of dying when compared to patients in the earliest stage (stage 1). Table 1 presents results from both the unadjusted and adjusted sets of analyses. ## Graphical display I plotted cox model fitted values against ageyr and bil variables marking sex and disease stage (`histo`) with color and different symbols, respectively. ## Conclusions DCPA was not found to be statistically significantly associated with increased survival in either univariate or multivariable analyses. As this was a randomized trial with 312 patients, we conclude that DPCA does not appear to be efficacious in the treatment of patients with primary biliary cirrhosis. While primary biliary cirrhosis is a disease that primarily affects females, the prognosis is significantly worse for males. Similarly, the risk of death is much worse for patients in later stages of the disease relative to those in the earlier stages. The results of this research suggest that improved screening

techniques to identify the disease in affected patients early on, coupled with increased outreach to males at risk of developing PBC could result in a better overall prognosis for patients having this disease. It is clear that all of the variables I picked are important in the final model although not all levels of the categorical variables were statistically significant. This work is only the beginning and more precise answers to the research questions discussed in the introduction will require further inspection with models more precisely adapted to each research question.

Results

First, I will produce a few simple summaries of drug response based on `sex`, `agecat` and `histo` variables. First some basic exploratory data analysis will let me know if I am on the right track with the variables I have chosen. If there is no difference between the median survival times of the groups I am interested in, it will be unlikely that I will see anything significant in my model. From this initial analysis it looks like patients in the highest age category that were given placebo fare the worst. These results indicate that elderly patients my stand to benefit the most from taking the drug. Shockingly, men taking the drug appear to have a shorter survival time than with the drug, and do not survive as long as a women in general. Similarly, the drug appeared to have a negative effect on survival in patients with earliest stage of disease (histo = 1). Now I will take a similar approach to the data but using a cox proportional hazards model.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# install.packages("broom")
library(broom)
pbcData %>%
    group_by(sex, drug) %>%
    summarise(med_surv = median(survyr))
```

```
## # A tibble: 4 x 3
## # Groups:   sex [?]
##   sex     drug med_surv
##   <chr>  <int>    <dbl>
## 1 Female     0     5.02
## 2 Female     1     5.33
## 3 Male       0     4.54
## 4 Male       1     3.57
```

```
pbcData %>%
    group_by(agecat, drug) %>%
    summarise(med_surv = median(survyr))
```

```
## # A tibble: 6 x 3
## # Groups:   agecat [?]
##   agecat       drug med_surv
##   <fct>       <int>    <dbl>
## 1 < 45 yrs        0     5.67
```

```
## 2 < 45 yrs        1     5.31
## 3 >= 55 yrs       0     4.00
## 4 >= 55 yrs       1     4.84
## 5 45 - 55 yrs     0     5.87
## 6 45 - 55 yrs     1     5.63
```

```
pbcData %>%
    group_by(histo, drug) %>%
    summarise(med_surv = median(survyr))
```

```
## # A tibble: 8 x 3
## # Groups:   histo [?]
##   histo  drug med_surv
##   <fct> <int>    <dbl>
## 1 1         0    10.4
## 2 1         1     6.89
## 3 2         0     6.30
## 4 2         1     6.86
## 5 3         0     5.27
## 6 3         1     5.46
## 7 4         0     3.38
## 8 4         1     3.57
```

I decided to put all variables of interest into one model rather creating multiple models that address each of the above questions, because the instructions say to have at most one figure and one table. If any of the results are statistically significant, I can explore the question further with a more specific model in the future.

```
cox_all_var = coxph(formula = SurvObj ~ drug + sex + bil + histo + agecat, data = pbcData)
all_var_summary <- summary(cox_all_var)
library(broom, help)
cox_all_var %>%
tidy()
```

```
##                term  estimate  std.error  statistic     p.value
## 1              drug 0.1100643 0.18357667  0.5995547 0.548803056
## 2           sexMale 0.5370231 0.24314396  2.2086632 0.027198075
## 3               bil 0.1509102 0.01411473 10.6916813 0.000000000
## 4            histo2 1.5141214 1.03625787  1.4611434 0.143976080
## 5            histo3 1.8944113 1.01944863  1.8582704 0.063130619
## 6            histo4 2.7098718 1.01535621  2.6688878 0.007610287
## 7   agecat>= 55 yrs 0.5371007 0.24820477  2.1639419 0.030468811
## 8 agecat45 - 55 yrs 0.3930107 0.24653317  1.5941494 0.110902570
##      conf.low conf.high
## 1 -0.24973941 0.4698679
## 2  0.06046971 1.0135765
## 3  0.12324587 0.1785746
## 4 -0.51690671 3.5451495
## 5 -0.10367135 3.8924939
## 6  0.71981021 4.6999334
## 7  0.05062828 1.0235731
## 8 -0.09018542 0.8762068
```

```
coef(cox_all_var) %>%
summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1101  0.3325  0.5371  0.9808  1.6092  2.7099
```

```
df <- data.frame(adj_HR = round(exp(coef(cox_all_var)), 2),
           lower_CI = round(confint(cox_all_var)[,1], 2),
           upper_CI = round(confint(cox_all_var)[,2], 2))
rownames(df) <- rownames(confint(cox_all_var))

#install.packages("captioner")
library(captioner, help)
figs <- captioner(prefix="Figure")
tbls <- captioner(prefix="Table")
library(knitr)
knitr::kable(df, format = "markdown")
```

	adj_HR	lower_CI	upper_CI
drug	1.12	-0.25	0.47
sexMale	1.71	0.06	1.01
bil	1.16	0.12	0.18
histo2	4.55	-0.52	3.55
histo3	6.65	-0.10	3.89
histo4	15.03	0.72	4.70
agecat>= 55 yrs	1.71	0.05	1.02
agecat45 - 55 yrs	1.48	-0.09	0.88

Table 1: Adjusted Hazard Ratio Estimates of Death obtained from Proportional Hazards Regression.

The results of the model Plotting

```
par(mfrow=c(2,2), mar = c(0, 0, 0, 0), oma = c(4, 4, 0.1, 0.1))
palette()
```

```
## [1] "black"   "red"     "green3"  "blue"    "cyan"    "magenta" "yellow"
## [8] "gray"
```

```
# sexplot
km_sex = survfit(SurvObj ~ drug + sex, data = pbcData,
type="kaplan-meier", conf.type="log-log")
plot(km_sex, las = 1,
     xaxt='n', ann=FALSE,
     col = 1:8)
legend("bottomleft",
       legend=names(km_sex$strata),
       col=1:length(km_sex$strata),
       cex = 0.75,
       lty=c(1,1), # gives the legend appropriate symbols (lines)
       lwd=c(2.5,2.5))

# ageplot
km_age = survfit(SurvObj ~ drug + agecat, data = pbcData,
type="kaplan-meier", conf.type="log-log")
plot(km_age,
     xaxt='n', yaxt='n', ann=FALSE,
     col = 1:8, xlab = "Time", ylab = "Survival")
legend("bottomleft",
       legend=names(km_age$strata),
       col=1:length(km_age$strata),
```

```
       cex = 0.65,
       lty=c(1,1), # gives the legend appropriate symbols (lines)
       lwd=c(2.5,2.5))

# histoplot
km_histo = survfit(SurvObj ~ drug + histo, data = pbcData,
type="kaplan-meier", conf.type="log-log")
plot(km_histo, las = 1, col = 1:8)
legend("bottomleft",
       legend=names(km_histo$strata),
       col=1:length(km_histo$strata),
       cex = 0.6,
       lty=c(1,1), # gives the legend appropriate symbols (lines)
       lwd=c(2.5,2.5))



# bilplot
pbcData['bilcat'] <- ifelse(pbcData["bil"][[1]]>median(pbcData["bil"][[1]]), 1, 0)
head(pbcData)
```

```
## # A tibble: 6 x 20
##    case  drug sex       bil histo death survyr `_st`  `_d`  `_t` `_t0`
##   <int> <int> <chr>   <dbl> <fct> <int>  <dbl> <int> <int> <dbl> <int>
## 1     1     1 Female 14.5   4         1   1.10     1     1  1.10     0
## 2     2     1 Female  1.10  3         0  12.3      1     0 12.3      0
## 3     3     1 Male    1.40  4         1   2.77     1     1  2.77     0
## 4     4     1 Female  1.80  4         1   5.27     1     1  5.27     0
## 5     5     0 Female  3.40  3         0   4.12     1     0  4.12     0
## 6     6     0 Female  0.800 3         1   6.86     1     1  6.86     0
## # ... with 9 more variables: ageyr <dbl>, agecat <fct>, agegr_2 <int>,
## #   agegr_3 <int>, hstage2 <int>, hstage3 <int>, hstage4 <int>,
## #   SurvObj <S3: Surv>, bilcat <dbl>
```

```
km_bil = survfit(SurvObj ~ drug + bilcat, data = pbcData,
type="kaplan-meier", conf.type="log-log")
plot(km_bil,
     yaxt='n', ann=FALSE,
     cex.lab = 0.75,
     col = 1:8)
legend("bottomleft",
       legend=names(km_bil$strata),
       col=1:length(km_bil$strata),
       cex = 0.75,
       lty=c(1,1), # gives the legend appropriate symbols (lines)
       lwd=c(2.5,2.5))
mtext("Time (years)", side = 1, outer = TRUE, cex = 1.15, line = 2.2, col = "black")
mtext("Survival", side = 2, outer = TRUE, cex = 1.15, line = 2.2, col = "black")
```

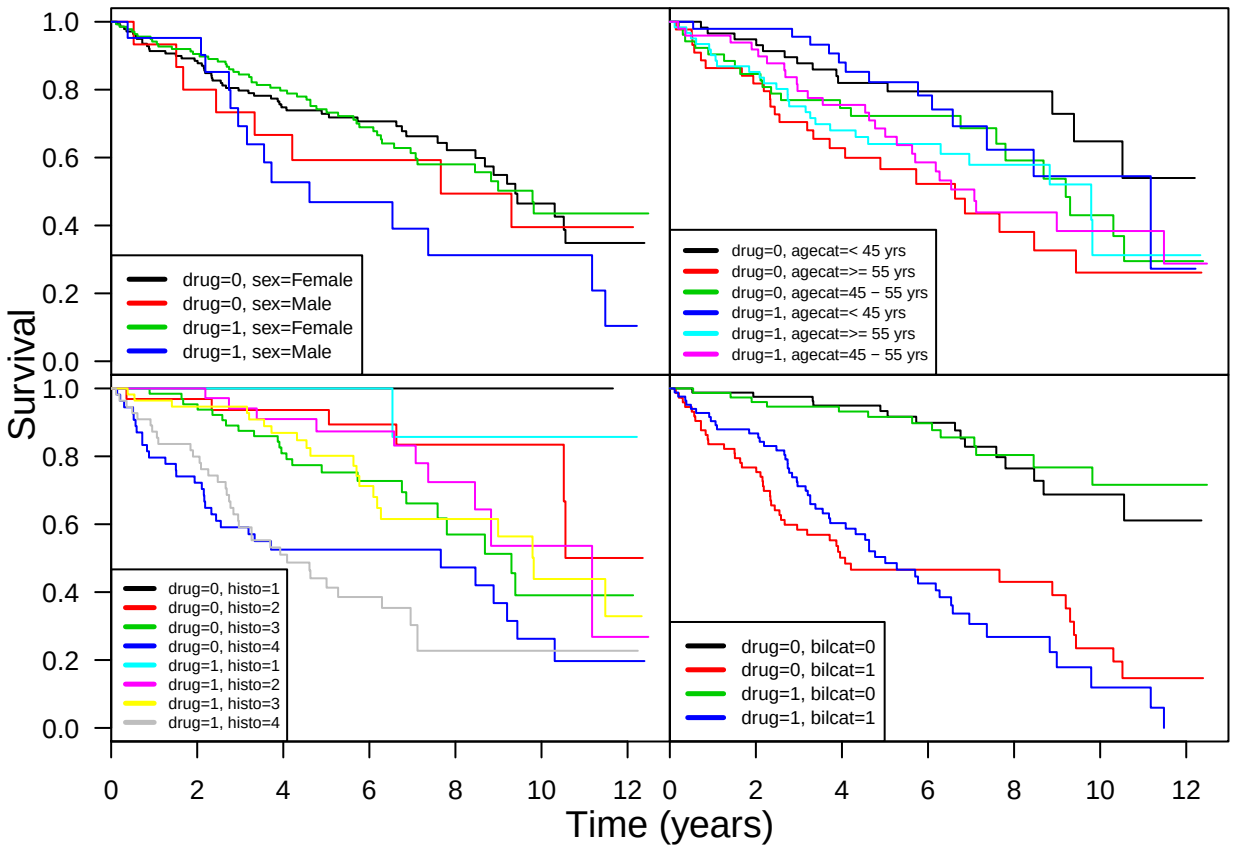


Figure 1: Survival of Primary Biliary Cirrhosis patients treated with D-penicillimin (DPCA) or a placebo.

To make a similar plot with the `bil` variable, I will first create a new categorical (binary) variable called `bilcat`.