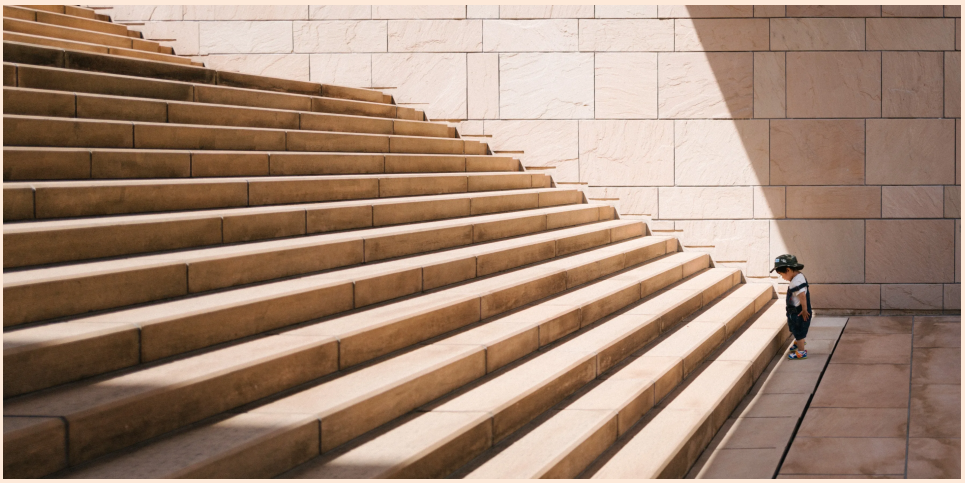
BIOSTATS 640 – Introduction to R
Fall 2023
https://people.umass.edu/biep640w/webpages/demonstrations.html



https://towardsdatascience.com/kaplan-meier-curves-c5768e349479

R Lesson 13
Introduction to Survival Analysis
December 8, 2023
Dataset used
`pbc.Rdata`

Packages Used: **`ggplot2, stargazer, gmodels, survival, survminer`**

1. DPCA Study of Primary Biliary Cirrhosis

`pbc.Rdata`

Source:
Dickson ER, Grambsch PM and Fleming TR (1989) Prognosis in primary biliary-cirrhosis - model for decision making. *Hepatology*, **10**, 1-7.

Introduction:
Bile is a fluid produced in your liver which functions in digesting food and aids in ridding your body of worn-out red blood cells, cholesterol and toxins. Primary biliary cirrhosis is an autoimmune disease in which the body turns against its own cells, in this case bile duct cells. As the bile ducts are increasingly damaged, harmful substances can accumulate. This can lead to irreversible scarring of liver tissue (cirrhosis). Among other things, the sufferer can experience abdominal pain, internal bleeding and, ultimately, liver failure. Primary biliary cirrhosis is also a risk factor for liver cancer.

This illustration utilizes data from a randomized controlled trial of D-penicillamine (DPCA) for the treatment of primary biliary cirrhosis. A total of n=312 consenting subjects were enrolled and randomized to either active treatment or placebo-control (presumably this group received standard care). Time zero is date of diagnosis and initiation of treatment. Study participants were followed to event of end-stage liver disease or censoring. Thus, these are an example of ***"right"*** censored data. Over the approximate 10 years of follow-up, 125 events of death (40%) were observed.

The goal of these analyses was to assess the benefit of randomization to DPCA on survival, overall and after adjustment for selected covariates.

Data dictionary/Codebook

Variable	Label	Type	Codings
years	Time to death, years	numeric	Range: 0.11 – 12.47
status	Event/censoring indicator		1 = Died 0 = Censored
rx	Treatment/Randomization	integer	1 = DPCA 0 = Control
histol	Severity of liver damage at dx	integer	1 = "lowest" .... 4 = "highest"
bilirubin	Serum bilirubin (mg/dl)	numeric	Range: 0.30 – 28.00

2. Prepare Data for Survival Analysis

load R data. inspect structure

```
load(file="pbc.Rdata")
pbc <- as.data.frame(pbc)
str(pbc)

## 'data.frame':    312 obs. of  23 variables:
##  $ number   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ status   : int  1 0 1 1 0 1 0 1 1 1 ...
##  $ rx       : int  0 0 0 0 1 1 1 1 0 1 ...
##  $ sex      : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ asictes  : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ hepatom  : int  1 1 0 1 1 1 1 0 0 0 ...
##  $ spiders  : int  1 1 0 1 1 0 0 0 1 1 ...
##  $ edema    : int  1 0 1 1 0 0 0 0 0 1 ...
##  $ bilirubin: num  14.5 1.1 1.4 1.8 3.4 ...
--- several rows omitted --
##   ..$ yn  : Named int [1:2] 0 1
##   .. ..- attr(*, "names")= chr [1:2] "No" "Yes"
```

prepare data for survival analysis

```
library(tidyverse)
library(survival)

# keep only vars of interest
temp <- pbc %>%
        select(years,status,rx,histol,bilirubin)
temp <- as.data.frame(temp)
str(temp)

## 'data.frame':    312 obs. of  5 variables:
##  $ years    : num  1.1 12.32 2.77 5.27 4.12 ...
##  $ status   : int  1 0 1 1 0 1 0 1 1 1 ...
##  $ rx       : int  0 0 0 0 1 1 1 1 0 1 ...
##  $ histol   : int  4 3 4 4 3 3 3 3 2 4 ...
##  $ bilirubin: num  14.5 1.1 1.4 1.8 3.4 ...

# create labeled factor versions of rx and status
temp$rxf <- factor(temp$rx,
                   levels=c(0,1),
                   labels=c("0=Control", "1=DPCA"))
temp$statusf <- factor(temp$status,
                   levels=c(0,1),
                   labels=c("0=Censored", "1=Died"))


# create survival data object using Surv() in {survival}
# time-to-event variable = years
# censoring variable = status (1=died)
pbc.survival <- with(temp,survival::Surv(years,status))
str(pbc.survival)

##  'Surv' num [1:312, 1:2]  1.095  12.320+  2.771   5.270   4.118+  6.853   5.016+  6.752   6.571   0.140  ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:2] "time" "status"
##  - attr(*, "type")= chr "right"
```

3. Model Free Approaches

descriptives

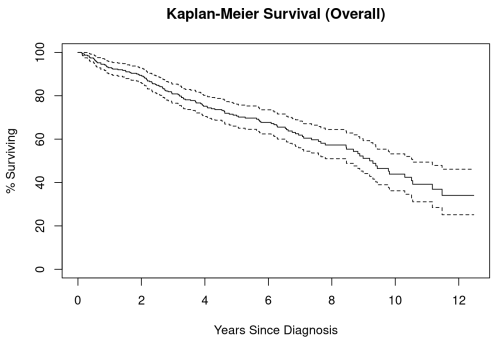
```
library(stargazer)
library(summarytools)

# All variables except factor variables
myvars <- c("rx","status","years","histol","bilirubin")
stargazer(temp[myvars], type="text",median=TRUE,
          title="DPCA Study of Primary Biliary Cirrhosis")

##
## DPCA Study of Primary Biliary Cirrhosis
## ================================================================
## Statistic  N  Mean  St. Dev.  Min  Pctl(25) Median Pctl(75)  Max
## ----------------------------------------------------------------
## rx        312 0.494  0.501     0      0       0       1       1
## status    312 0.401  0.491     0      0       0       1       1
## years     312 5.493  3.075   0.112  3.261   5.036   7.385   12.474
## histol    312 3.032  0.878     1      2       3       4       4
## bilirubin 312 3.256  4.530   0.300  0.800   1.350   3.425   28.000
## ----------------------------------------------------------------

# xtab primary predictor x event of death
ctable(temp$rxf,temp$statusf)

## Cross-Tabulation, Row Proportions
## rxf * statusf
## Data Frame: temp
##
## ----------- --------- ------------- ------------- --------------
##               statusf    0=Censored        1=Died          Total
##         rxf
##   0=Control              93 (58.9%)    65 (41.1%)   158 (100.0%)
##      1=DPCA              94 (61.0%)    60 (39.0%)   154 (100.0%)
##       Total             187 (59.9%)   125 (40.1%)   312 (100.0%)
## ----------- --------- ------------- ------------- --------------
```

Interpretation:
Among 158 randomized to CONTROL, there were 65 deaths (41%)
Among 154 randomized to DPCA, there were 60 deaths (39%)

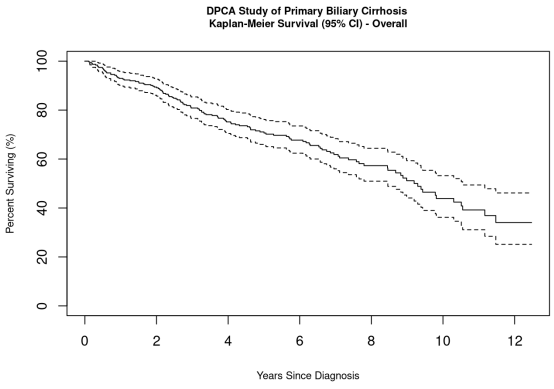
**** Kaplan-Meier Curve Estimation- ALL****

```
library(survival)
library(ggplot2)
library(GGally)

# Kaplan-Meier Curve Estimates
kmall <- survfit(formula = pbc.survival ~ 1,
                 type="kaplan-meier",
                 conf.type="log", conf.int=.95,
                 data=temp)
```

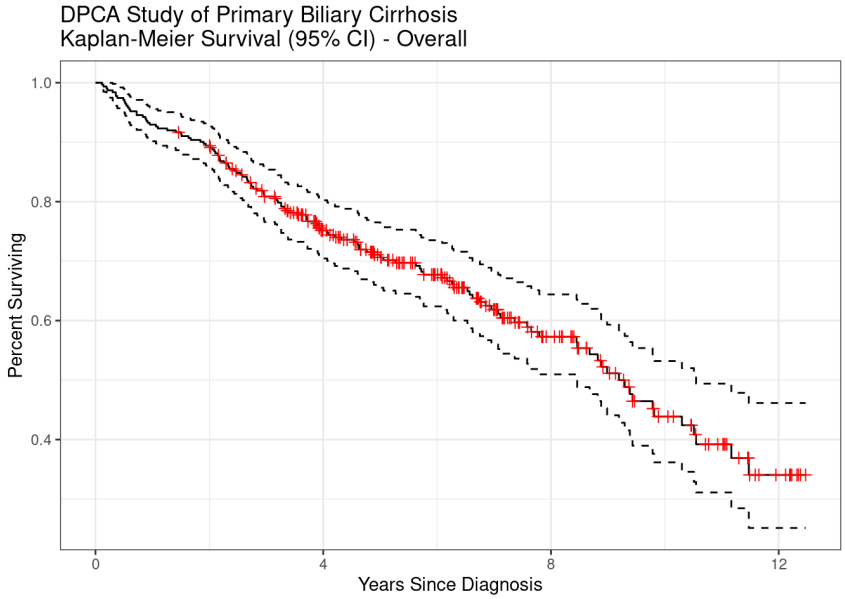
```
summary(kmall)

## Call: survfit(formula = pbc.survival ~ 1, data = temp, type = "kaplan-meier",
##     conf.type = "log", conf.int = 0.95)
##
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.112    312       1    0.997 0.00320        0.991        1.000
##   0.140    311       1    0.994 0.00452        0.985        1.000
##   0.194    310       1    0.990 0.00552        0.980        1.000
##   0.211    309       1    0.987 0.00637        0.975        1.000
##   0.301    308       1    0.984 0.00711        0.970        0.998
--- several rows omitted --
##   9.812     34       1    0.439 0.04317        0.362        0.532
##  10.300     30       1    0.424 0.04414        0.346        0.520
##  10.511     27       1    0.408 0.04522        0.329        0.507
##  10.549     25       1    0.392 0.04626        0.311        0.494
##  11.168     17       1    0.369 0.04895        0.285        0.479
##  11.474     13       1    0.341 0.05278        0.251        0.461

# Kaplan-Meier Curve Plot

# plot() in {base}:  no frills
plot(kmall,xlab="Years Since Diagnosis",
     ylab="% Surviving", yscale=100,
     main="Kaplan-Meier Survival (Overall)")
```




```
# plot() in {base}:  frills
# cex.lab= and cex.main= is used to change size
plot(kmall,
     xlab="Years Since Diagnosis",
     ylab="Percent Surviving (%)", yscale=100, cex.lab=0.75,
     main="DPCA Study of Primary Biliary Cirrhosis \nKaplan-Meier Survival (95% CI) - Overall",
     cex.main=0.75)
```



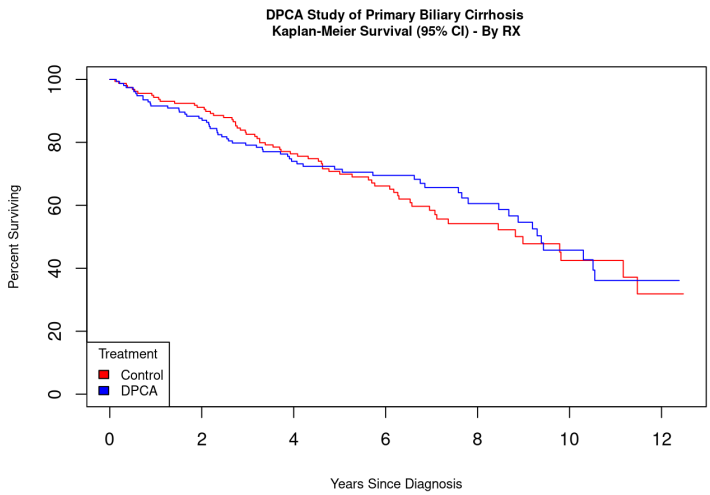
```
# ggsurv( ) in {ggplot2}
ggsurv(kmall) +
  labs(x="Years Since Diagnosis",
       y="Percent Surviving") +
  ggtitle("DPCA Study of Primary Biliary Cirrhosis \nKaplan-Meier Survival (95% CI) - Overall") +
  theme_bw()
```



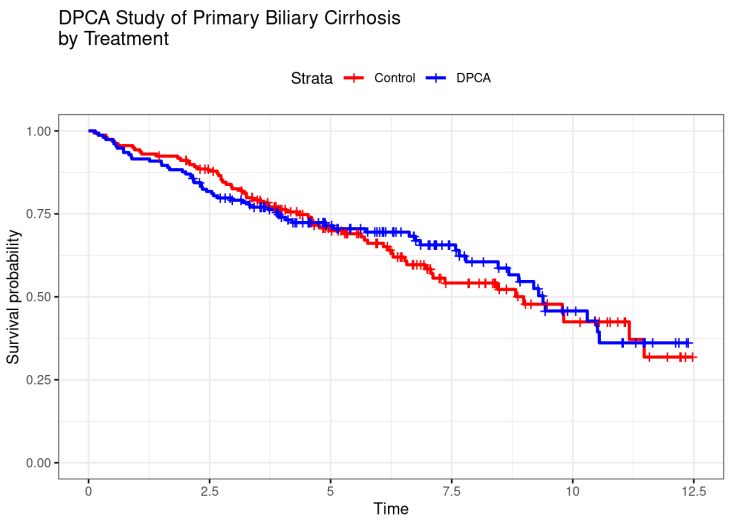
Kaplan-Meier Curve Estimation - by rx group

```
library(survival)
library(survminer)
library(ggplot2)
library(GGally)

# Kaplan-Meier Curve Estimates, by rx group
kmrx <- survfit(formula = pbc.survival ~ rx,
                type="kaplan-meier",
                conf.type="log", conf.int=.95,
                data=temp)
summary(kmrx)

## Call: survfit(formula = pbc.survival ~ rx, data = temp, type = "kaplan-meier",
##     conf.type = "log", conf.int = 0.95)
##
##                 rx=0
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.112    158       1    0.994 0.00631        0.981        1.000
##   0.194    157       1    0.987 0.00889        0.970        1.000
##   0.359    156       1    0.981 0.01086        0.960        1.000
##   0.383    155       1    0.975 0.01250        0.950        0.999
--- several rows omitted --
##  11.474      7       1    0.319 0.07922        0.196        0.519
##
##                 rx=1
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.140    154       1    0.994 0.00647        0.981        1.000
##   0.211    153       1    0.987 0.00912        0.969        1.000
##   0.301    152       1    0.981 0.01114        0.959        1.000
-- several rows omitted --
##  10.511     13       1    0.394 0.06719        0.282        0.551
##  10.549     12       1    0.361 0.06916        0.248        0.526
```

```
# Kaplan-Meier Curve Plot
# plot( ) in {base}: frills
plot(kmrx,
     xlab="Years Since Diagnosis",
     ylab="Percent Surviving",yscale=100,col=c("red","blue"),
     main="DPCA Study of Primary Biliary Cirrhosis \nKaplan-Meier Survival (95% CI) - By RX",
     cex.main=0.75,cex.lab=0.75)                              # NOTE!!  unlike in ggplot, no + here
legend("bottomleft",
       title="Treatment",c("Control","DPCA"),
       fill=c("red","blue"),
       cex=0.75)
```




```
# ggsurvplot() in package {survminer}
ggsurvplot(kmrx,
           #pval = TRUE, conf.int = TRUE,
           #risk.table = TRUE,                                  # Add risk table
           #risk.table.col = "strata",                          # Change risk table color by groups
           #linetype = "strata",                                # Change line type by groups
           #surv.median.line = "hv",                            # Specify median survival
           title="DPCA Study of Primary Biliary Cirrhosis \nby Treatment",
           legend.labs=c("Control", "DPCA"),                    # Legend labels
           ggtheme = theme_bw(),                                # Change ggplot2 theme
           palette=c("red", "blue"))                            # colors
```



Interpretation:
These graphs suggest that the distributions are similar.

log-rank test of equality of survival distributions (NULL: Equality)

```
library(survival)
survival::survdiff(pbc.survival~rx,data=pbc)

## Call:
## survival::survdiff(formula = pbc.survival ~ rx, data = pbc)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=0 158       65     63.2    0.0502     0.102
## rx=1 154       60     61.8    0.0513     0.102
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```

Interpretation:
Do NOT reject the null (log rank test p-value = .70). These data provide NO statistically significant evidence that the survival distributions differ.

4. Cox Proportional Hazards (PH) Model Regression

Cox PH - fit

```
library(survival)
library(stargazer)

# Single predictor = rx.  0/1
fit <- coxph(pbc.survival~rx + bilirubin, data=temp)
fit                                              # show

## Call:
## coxph(formula = pbc.survival ~ rx + bilirubin, data = temp)
##
##               coef exp(coef) se(coef)      z                  p
## rx        -0.20118   0.81776  0.18342 -1.097              0.273      Relative hazard death w treatment = .82 (ns)
## bilirubin  0.15147   1.16354  0.01329 11.400 <0.0000000000000002
##
## Likelihood ratio test=85.86  on 2 df, p=< 0.00000000000000022
## n= 312, number of events= 125

summary(fit)                                     # detailed summary

## Call:
## coxph(formula = pbc.survival ~ rx + bilirubin, data = temp)
##
##   n= 312, number of events= 125
##
##               coef exp(coef) se(coef)      z            Pr(>|z|)
## rx        -0.20118   0.81776  0.18342 -1.097               0.273
## bilirubin  0.15147   1.16354  0.01329 11.400 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rx           0.8178     1.2228    0.5708     1.172
## bilirubin    1.1635     0.8594    1.1336     1.194
##
## Concordance= 0.78  (se = 0.021 )
## Likelihood ratio test= 85.86  on 2 df,   p=<0.0000000000000002
## Wald test            = 130.9  on 2 df,   p=<0.0000000000000002
## Score (logrank) test = 191.2  on 2 df,   p=<0.0000000000000002
```

Cox PH Model - multivariable model development

```
library(survival)                       # if using {stargazer} later, do NOT precede coxph( ) with survival.
library(stargazer)

# Single predictor = rx.  0/1
fit_rx <- coxph(pbc.survival~rx, data=temp)
fit_rx                                                  # Show

## Call:
## coxph(formula = pbc.survival ~ rx, data = temp)
##
##        coef exp(coef) se(coef)      z     p
## rx -0.05722   0.94438  0.17916 -0.319 0.749
##
## Likelihood ratio test=0.1  on 1 df, p=0.7494
## n= 312, number of events= 125

summary(fit_rx)                                         # detailed summary

## Call:
## coxph(formula = pbc.survival ~ rx, data = temp)
##
##   n= 312, number of events= 125
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## rx -0.05722   0.94438  0.17916 -0.319    0.749
##
##    exp(coef) exp(-coef) lower .95 upper .95
## rx    0.9444      1.059    0.6647     1.342
##
## Concordance= 0.499  (se = 0.025 )
## Likelihood ratio test= 0.1  on 1 df,   p=0.7
## Wald test            = 0.1  on 1 df,   p=0.7
## Score (logrank) test = 0.1  on 1 df,   p=0.7
```

Interpretation:
Relative to CONTROLS, patients randomized to DPCA have a lower hazard of death (.94) at all times of follow-up under the assumption of PH. This is not statistically significant however (p-value = .70) and the 95% CI for the hazard ratio HR (.66 - 1.3) includes the "no association" null value of 1.

```
# Single predictor = histol.   Nominal - must declare as factor
fit_histol <- coxph(pbc.survival~factor(histol), data=temp)
fit_histol

## Call:
## coxph(formula = pbc.survival ~ factor(histol), data = temp)
##
##                   coef exp(coef) se(coef)     z      p
## factor(histol)2  1.607     4.988    1.031 1.559 0.1191
## factor(histol)3  2.150     8.581    1.012 2.124 0.0337
## factor(histol)4  3.063    21.387    1.009 3.036 0.0024
##
## Likelihood ratio test=52.74  on 3 df, p=0.00000000002085
## n= 312, number of events= 125

# summary(fit_histol)

# Single predictor = bilirubin.  Continuous
fit_bili <- coxph(pbc.survival~bilirubin, data=temp)
cat(paste(" ", "Fit of Single Predictor - bilirubin"," ",sep="\n"))

##
## Fit of Single Predictor - bilirubin
##
```

```
fit_bili

## Call:
## coxph(formula = pbc.survival ~ bilirubin, data = temp)
##
##              coef exp(coef) se(coef)     z                   p
## bilirubin 0.14892   1.16058  0.01301 11.44 <0.0000000000000002
##
## Likelihood ratio test=84.65  on 1 df, p=< 0.00000000000000022
## n= 312, number of events= 125

# summary(fit_bili)
```

side-by-side comparison of single predictor models

```
library(survival)
library(stargazer)

# fits from previous chunk
stargazer::stargazer(fit_rx, fit_histol,fit_bili,
                     type="text",
                     title="DPCA Study of Primary Biliary Cirrhosis",
                     dep.var.labels=c("y=status (1=died)"),
                     column.labels = c("rx", "histol", "bilirubin"))

##
## DPCA Study of Primary Biliary Cirrhosis
## ================================================================================
##                                       Dependent variable:
##                      ---------------------------------------------------
##                                       y=status (1=died)
##                            rx             histol            bilirubin
##                           (1)              (2)                (3)
## --------------------------------------------------------------------------------
## rx                      -0.057
##                        (0.179)
##
## factor(histol)2                           1.607
##                                          (1.031)
##
## factor(histol)3                          2.150**
##                                          (1.012)
##
## factor(histol)4                          3.063***
##                                          (1.009)
##
## bilirubin                                                    0.149***
##                                                               (0.013)
##
## --------------------------------------------------------------------------------
## Observations              312              312                312
## R2                       0.0003           0.156              0.238
## Max. Possible R2         0.983            0.983              0.983
## Log Likelihood          -639.915         -613.597           -597.641
## Wald Test            0.100 (df = 1) 43.920*** (df = 3) 130.920*** (df = 1)
## LR Test              0.102 (df = 1) 52.738*** (df = 3) 84.651*** (df = 1)
## Score (Logrank) Test 0.102 (df = 1) 53.853*** (df = 3) 190.869*** (df = 1)
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:
All three single-predictor models are statistically significant compared to the null model of zero predictors. But we're not done yet.

multiple predictor model development

```
library(survival)
library(stargazer)

# KEY:  Every model will contain primary predictor of interest = rx

# model 1:  rx
m1 <- coxph(pbc.survival~rx,data=temp)

# model 2:  histol + rx
m2 <- coxph(pbc.survival ~ rx + factor(histol),data=temp)

# model 3: bilirubin + rx
m3 <- coxph(pbc.survival ~ rx + bilirubin, data=temp)

# model 4:  rx + histol + bilirubin
m4 <- coxph(pbc.survival ~ rx + factor(histol) + bilirubin, data=temp)

# Side by side:  betas with 95% CI
stargazer::stargazer(m1, m2, m3, m4,
                     type="text",
                     ci=TRUE,
                     title="Multivariable Logistic Regression Models: Betas (95% CI)",
                     dep.var.labels=c("y=status (1=died)"))
##
## Multivariable Logistic Regression Models: Betas (95% CI)
## ================================================================================================
##                                          Dependent variable:
##                      ------------------------------------------------------------------
##                                           y=status (1=died)
##                           (1)              (2)                (3)                (4)
## ------------------------------------------------------------------------------------------------
## rx                      -0.057           -0.147             -0.201             -0.158
##                     (-0.408, 0.294)  (-0.500, 0.205)    (-0.561, 0.158)    (-0.514, 0.197)
##
## factor(histol)2                          1.629                                 1.526
##                                       (-0.393, 3.650)                       (-0.497, 3.548)
##
## factor(histol)3                         2.177**                               1.923*
##                                       (0.192, 4.162)                        (-0.067, 3.912)
##
## factor(histol)4                         3.093***                              2.797***
##                                       (1.114, 5.072)                        (0.816, 4.778)
##
## bilirubin                                                  0.151***           0.148***
##                                                          (0.125, 0.178)     (0.120, 0.175)
##
## ------------------------------------------------------------------------------------------------
## Observations              312              312                312                312
## R2                      0.0003           0.157              0.241              0.336
## Max. Possible R2         0.983            0.983              0.983              0.983
## Log Likelihood         -639.915         -613.262           -597.038           -576.150
## Wald Test            0.100 (df = 1)  44.530*** (df = 4) 130.940*** (df = 2) 148.440*** (df = 5)
## LR Test              0.102 (df = 1)  53.408*** (df = 4) 85.858*** (df = 2)  127.632*** (df = 5)
## Score (Logrank) Test 0.102 (df = 1)  54.478*** (df = 4) 191.230*** (df = 2) 217.136*** (df = 5)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:

(1) After adjustment for covariates, randomization to DPCA does NOT confer a statistically significant improvement in survival

(2) In adjusted analyses, severe liver damage (histology score=4) and elevated bilirubin ARE statistically significantly associated with poorer survival (earlier event of death).

```
# Side by side:  betas with p-values
stargazer::stargazer(m1, m2, m3, m4,
                     type="text",
                     report=('vc*p'),
                     title="Multivariable Logistic Regression Models: p-values",
                     dep.var.labels=c("y=status (1=died)"))

##
## Multivariable Logistic Regression Models: p-values
## ================================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------------------------
##                                            y=status (1=died)
##                          (1)              (2)                (3)                (4)
## ------------------------------------------------------------------------------------------------
## rx                     -0.057           -0.147             -0.201             -0.158
##                       p = 0.750        p = 0.414          p = 0.273          p = 0.383
##
## factor(histol)2                          1.629                                 1.526
##                                        p = 0.115                             p = 0.140
##
## factor(histol)3                         2.177**                               1.923*
##                                        p = 0.032                             p = 0.059
##
## factor(histol)4                         3.093***                              2.797***
##                                        p = 0.003                             p = 0.006
##
## bilirubin                                                  0.151***           0.148***
##                                                           p = 0.000          p = 0.000
##
## ------------------------------------------------------------------------------------------------
## Observations             312              312                312                312
## R2                      0.0003           0.157              0.241              0.336
## Max. Possible R2         0.983            0.983              0.983              0.983
## Log Likelihood         -639.915         -613.262           -597.038           -576.150
## Wald Test            0.100 (df = 1) 44.530*** (df = 4) 130.940*** (df = 2) 148.440*** (df = 5)
## LR Test              0.102 (df = 1) 53.408*** (df = 4) 85.858*** (df = 2)  127.632*** (df = 5)
## Score (Logrank) Test 0.102 (df = 1) 54.478*** (df = 4) 191.230*** (df = 2) 217.136*** (df = 5)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Interpretation: This table confirms what we just saw.

(1) After adjustment for covariates, randomization to DPCA does NOT confer a statistically significant improvement in survival

(2) In adjusted analyses, severe liver damage (histology score=4) and elevated bilirubin ARE statistically significantly associated with poorer survival (earlier event of death).

5. Regression Diagnostics for Cox Proportional Hazards (PH) Model Regression

test of proportional hazards (NULL: proportional hazards assumption is reasonable)

```
library(survival)

# Model 1:  pbc.survival~rx
cox.zph(m1)

##        chisq df    p
## rx     0.612  1 0.43
## GLOBAL 0.612  1 0.43
```

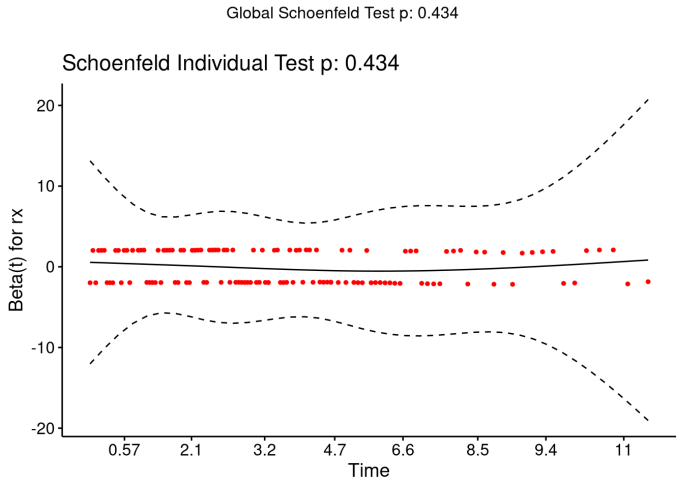
Interpretation:
The null hypothesis of proportional hazards is NOT rejected (p-value = .43)

```
# Model 2:  pbc.survival ~ rx + factor(histol)
cox.zph(m2)

##                chisq df      p
## rx              1.53  1 0.2164
## factor(histol) 11.96  3 0.0075
## GLOBAL         12.56  4 0.0136
```

Interpretation:
(1) For this fitted 2-predictor model, we do not find a departure from proportional hazards for the predictor rx.
(2) However, the assumption of proportional hazards is rejected for the predictor histology.
(3) So, not surprisingly, for the model globally, the assumption of proportional hazards is NOT reasonably met.

graphical assessment of proportional hazards X=time, Y=scaled Schoenfeld residuals (LOOK FOR: band at 0)

```
library(survival)
library(ggplot2)

# Model 1:  pbc.survival~rx
test1.ph = cox.zph(m1)
ggcoxzph(test1.ph)
```




```
# Model 3:  pbc.survival ~ rx + bilirubin
test3.ph = cox.zph(m3)
ggcoxzph(test3.ph)
```