# Biostats - Cox Proportional Hazard

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

# 1) Monoclonal Gammopathy of Undetermined Significance Dataset

**1a) Fit a proportional hazard model and test if age has an association at 5% level.**

```
library(survival)
data(mgus)
mgusdata <- mgus[,c(2,3,7,8,9,10,11)]  # only pull data we need
mgusdata <- mgusdata[complete.cases(mgusdata), ]   # remove any missing data
outmgus <- coxph(Surv(futime, death)~age+sex+alb+creat+hgb, data = mgusdata)
summary(outmgus)
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + sex + alb + creat +
##     hgb, data = mgusdata)
##
##   n= 176, number of events= 165
##
##              coef exp(coef)  se(coef)      z Pr(>|z|)
## age      0.070350  1.072884  0.008555  8.223  < 2e-16 ***
## sexmale  0.204720  1.227181  0.164315  1.246  0.21280
## alb     -0.256087  0.774075  0.201201 -1.273  0.20309
## creat    0.405708  1.500364  0.146719  2.765  0.00569 **
## hgb     -0.107078  0.898455  0.060412 -1.772  0.07632 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        1.0729     0.9321    1.0550     1.091
## sexmale    1.2272     0.8149    0.8893     1.693
## alb        0.7741     1.2919    0.5218     1.148
## creat      1.5004     0.6665    1.1254     2.000
## hgb        0.8985     1.1130    0.7981     1.011
##
## Concordance= 0.71  (se = 0.023 )
## Likelihood ratio test= 97.17  on 5 df,   p=<2e-16
## Wald test            = 92.92  on 5 df,   p=<2e-16
## Score (logrank) test = 99.58  on 5 df,   p=<2e-16
```

$H_0: \beta_{age}=0$ vs. $H_a: \beta_{age}\ne0$.
Test Statistic: z = 0.070350/0.008555 = 8.223 which is very high. The p-value is very low at 2e-16. Thus, we have significant evidence to reject the null hypothesis that age does not have any association with the hazard.

**1b) Estimate relative risk and its 95% CI for the death of a subject with the age of diagnosis 60 compared to the subject with the age of diagnosis 50 while all other covariates remain unchanged.**
The relative risk is \((\frac{exp^{60*.07035}}{exp^{50*.07035}})\) = **2.02.**
So, the relative risk of death for a 60 year old subject at diagnosis is **2 times higher** than that of a 50 year old, holding the other covariates constant.

The 95% CI is 2.02+-(1.96)(0.1728903) = **(1.681,2.359).**

```
library(msm)
sestar = deltamethod(~(exp(60*x1)/exp(50*x1)), 0.070350, outmgus$var[1,1])
sestar
```

```
## [1] 0.1728903
```

**1c) Test if there is any effect of gender, albumin, and hemoglobin at the 5% level.**

```
outH0 <- coxph(Surv(futime, death)~age+creat, data = mgusdata)
anova(outH0, outmgus)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(futime, death)
##  Model 1: ~ age + creat
##  Model 2: ~ age + sex + alb + creat + hgb
##    loglik  Chisq Df P(>|Chi|)
## 1 -671.33
## 2 -667.66 7.3248  3   0.06224 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

\(H_0: \beta_{sex} = \beta_{albumin} = \beta_{hemoglobin}= 0\) versus \(H_a:\) at least one coefficient is nonzero. The chi squared test statistic is 7.3248 resulting in a p-value of 0.06224, which, if using a 5% test, does not provide sufficient evidence to reject the null hypothesis that there is no association of gender, albumin and hemoglobin on time to death.

**1d) obtain the estimate and 95% CI for the 10 year survival probability.**
First Person:

```
out1d1 = survfit(outmgus, newdata=data.frame(age=60, sex="male", alb=3, creat=1, hgb=13.5 ))
index1 = findInterval(3650,out1d1$time)
out1d1$surv[index1]
```

```
## [1] 0.6957969
```

```
c(out1d1$lower[index1],out1d1$upper[index1])
```

```
## [1] 0.6148185 0.7874410
```

Second Person:

```
out1d2 = survfit(outmgus, newdata=data.frame(age=60, sex="male", alb=3, creat=4, hgb=13.5 ))
index2 = findInterval(3650,out1d2$time)
out1d2$surv[index2]
```

```
## [1] 0.2937592
```

```
c(out1d2$lower[index2],out1d2$upper[index2])
```

```
## [1] 0.09939494 0.86819759
```

| Estimate | 95% Confidence Interval |
|---|---|
| **Person 1** | |
| 0.696 | (0.615,0.787) |
| **Person 2** | |
| 0.294 | (0.099,0.868) |

**1e) Repeat d but include 2 factor interactions and choose model with stepwise selection.**

```
library(My.stepwise)
mgusdata2 = model.matrix( ~.^2, data = mgusdata)
mgusdata2 = mgusdata2[,c(-1,-10,-11,-15,-16,-20:-26)]  # drop the intercept and interactions wit
h death and time
mgusdata2 = data.frame(mgusdata2)
my.variable.list = names(mgusdata2)
my.variable.list = my.variable.list[c(-3,-4)]
My.stepwise.coxph(Time="futime", Status="death", variable.list=my.variable.list, data=mgusdata2)
```

```
## # ------------------------------------------------------------------------------------------
-------
## # Initial Model:
## Call:
## coxph(formula = formula, data = data, method = "efron")
##
##   n= 176, number of events= 165
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## age 0.07331   1.07607  0.00846 8.666   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.076     0.9293     1.058     1.094
##
## Concordance= 0.695  (se = 0.022 )
## Likelihood ratio test= 82.28  on 1 df,   p=<2e-16
## Wald test            = 75.1  on 1 df,   p=<2e-16
## Score (logrank) test = 78.81  on 1 df,   p=<2e-16
##
## # ------------------------------------------------------------------------------------------
-------
## ### iter num = 1, Forward Selection by LR Test: + age.creat
## Call:
## coxph(formula = Surv(futime, death) ~ age + age.creat, data = data,
##     method = "efron")
##
##   n= 176, number of events= 165
##
##               coef exp(coef) se(coef)     z Pr(>|z|)
## age       0.066120  1.068355 0.008822 7.494 6.66e-14 ***
## age.creat 0.006638  1.006660 0.001965 3.378 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age           1.068     0.9360     1.050     1.087
## age.creat     1.007     0.9934     1.003     1.011
##
## Concordance= 0.702  (se = 0.022 )
## Likelihood ratio test= 89.94  on 2 df,   p=<2e-16
## Wald test            = 85.41  on 2 df,   p=<2e-16
## Score (logrank) test = 92.65  on 2 df,   p=<2e-16
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Vari
able) or is bigger than 2.5 (Categorical Variable)
##       age age.creat
##  1.315179  1.315179
## # ------------------------------------------------------------------------------------------
-------
## ### iter num = 2, Forward Selection by LR Test: + alb.hgb
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + age.creat + alb.hgb,
##     data = data, method = "efron")
##
##   n= 176, number of events= 165
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age        0.064130  1.066231  0.008731  7.345 2.05e-13 ***
## age.creat  0.006175  1.006194  0.001977  3.123  0.00179 **
## alb.hgb   -0.022023  0.978218  0.009955 -2.212  0.02695 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0662     0.9379    1.0481    1.0846
## age.creat    1.0062     0.9938    1.0023    1.0101
## alb.hgb      0.9782     1.0223    0.9593    0.9975
##
## Concordance= 0.703  (se = 0.023 )
## Likelihood ratio test= 94.73  on 3 df,   p=<2e-16
## Wald test            = 91.15  on 3 df,   p=<2e-16
## Score (logrank) test = 98.79  on 3 df,   p=<2e-16
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Vari
able) or is bigger than 2.5 (Categorical Variable)
##       age age.creat   alb.hgb
##  1.355883  1.482383  1.113181
## # ================================================================================================
=======
## *** Stepwise Final Model (in.lr.test: sle = 0.15; out.lr.test: sls = 0.15; variable selection
restrict in vif = 999):
## Call:
## coxph(formula = Surv(futime, death) ~ age + age.creat + alb.hgb,
##     data = data, method = "efron")
##
##   n= 176, number of events= 165
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age        0.064130  1.066231  0.008731  7.345 2.05e-13 ***
## age.creat  0.006175  1.006194  0.001977  3.123  0.00179 **
## alb.hgb   -0.022023  0.978218  0.009955 -2.212  0.02695 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0662     0.9379    1.0481    1.0846
## age.creat    1.0062     0.9938    1.0023    1.0101
## alb.hgb      0.9782     1.0223    0.9593    0.9975
##
## Concordance= 0.703  (se = 0.023 )
## Likelihood ratio test= 94.73  on 3 df,   p=<2e-16
## Wald test            = 91.15  on 3 df,   p=<2e-16
## Score (logrank) test = 98.79  on 3 df,   p=<2e-16
```

```
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Vari
able) or is bigger than 2.5 (Categorical Variable)
##       age age.creat   alb.hgb
##  1.355883  1.482383  1.113181
```

Predict for person 1:

```
out1e <- coxph(Surv(futime, death)~age+age.creat+alb.hgb, data = mgusdata2)
out1e1 = survfit(out1e, newdata=data.frame(age=60, sex="male", alb=3, creat=1, hgb=13.5, age.cre
at=60,alb.hgb=40.5  ))
index1e = findInterval(3650,out1e1$time)
out1e1$surv[index1e]
```

```
## [1] 0.7101548
```

```
c(out1e1$lower[index1e],out1e1$upper[index1e])
```

```
## [1] 0.6426536 0.7847460
```

Predict for person 2:

```
out1e2 = survfit(out1e, newdata=data.frame(age=60, sex="male", alb=3, creat=4, hgb=13.5, age.cre
at=240,alb.hgb=40.5  ))
index1e2 = findInterval(3650,out1e2$time)
out1e2$surv[index1e2]
```

```
## [1] 0.3534174
```

```
c(out1e2$lower[index1e2],out1e2$upper[index1e2])
```

```
## [1] 0.1677229 0.7447033
```

| Estimate | 95% Confidence Interval |
|---|---|
| **Person 1** | |
| 0.710 | (0.642,0.785) |
| **Person 2** | |
| 0.353 | (0.168,0.745) |

These results are fairly similar to part 1d, except the survival probability estimate for person 2 is higher in part e and has a narrower confidence interval.

# 2) Colon Cancer Dataset

```
library(survival)
data(colon)
colondata = colon[colon$etype==2, c(1,9,10,12,15)] # want to look at etype =2 deaths now (not re
currence of disease)
colondata$time2 = round(colondata$time/91.25,0)
#colondata$time2 = floor(colondata$time/91.25)  # Discussion board has floor or round as valid o
ptions for calculating which quarter
out2a = glm(status ~ as.factor(time2)+nodes+as.factor(extent), family='binomial', data=colondat
a)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(out2a) # doesn't work that well, don't have many events in most time periods in quarter
 26 on, highish standard errors, higher AIC
```

```
##
## Call:
## glm(formula = status ~ as.factor(time2) + nodes + as.factor(extent),
##     family = "binomial", data = colondata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.84997  -0.38266  -0.00006   0.00005   2.74795
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         4.131e+00  9.031e+03   0.000    1.000
## as.factor(time2)1  -5.413e-02  1.049e+04   0.000    1.000
## as.factor(time2)2  -5.989e-02  9.709e+03   0.000    1.000
## as.factor(time2)3  -1.128e-01  9.496e+03   0.000    1.000
## as.factor(time2)4  -1.534e-01  9.474e+03   0.000    1.000
## as.factor(time2)5  -1.709e+01  8.865e+03  -0.002    0.998
## as.factor(time2)6   1.276e+01  9.480e+03   0.001    0.999
## as.factor(time2)7   1.257e+01  9.395e+03   0.001    0.999
## as.factor(time2)8  -6.229e-02  9.330e+03   0.000    1.000
## as.factor(time2)9  -5.576e-02  9.798e+03   0.000    1.000
## as.factor(time2)10 -7.047e-02  9.497e+03   0.000    1.000
## as.factor(time2)11 -3.339e-02  9.601e+03   0.000    1.000
## as.factor(time2)12 -7.501e-02  9.848e+03   0.000    1.000
## as.factor(time2)13 -5.506e-02  9.635e+03   0.000    1.000
## as.factor(time2)14 -1.779e+01  8.865e+03  -0.002    0.998
## as.factor(time2)15 -5.613e-02  1.065e+04   0.000    1.000
## as.factor(time2)16 -1.909e+01  8.865e+03  -0.002    0.998
## as.factor(time2)17 -1.845e+01  8.865e+03  -0.002    0.998
## as.factor(time2)18  1.401e+01  1.049e+04   0.001    0.999
## as.factor(time2)19 -1.854e-01  1.084e+04   0.000    1.000
## as.factor(time2)20 -2.111e+01  8.865e+03  -0.002    0.998
## as.factor(time2)21 -2.120e+01  8.865e+03  -0.002    0.998
## as.factor(time2)22 -2.248e+01  8.865e+03  -0.003    0.998
## as.factor(time2)23 -2.251e+01  8.865e+03  -0.003    0.998
## as.factor(time2)24 -2.313e+01  8.865e+03  -0.003    0.998
## as.factor(time2)25 -2.303e+01  8.865e+03  -0.003    0.998
## as.factor(time2)26 -2.435e+01  8.865e+03  -0.003    0.998
## as.factor(time2)27 -2.371e+01  8.865e+03  -0.003    0.998
## as.factor(time2)28 -2.295e+01  8.865e+03  -0.003    0.998
## as.factor(time2)29 -2.372e+01  8.865e+03  -0.003    0.998
## as.factor(time2)30 -2.363e+01  8.865e+03  -0.003    0.998
## as.factor(time2)31 -2.319e+01  8.865e+03  -0.003    0.998
## as.factor(time2)32 -2.331e+01  8.865e+03  -0.003    0.998
## as.factor(time2)33 -4.116e+01  1.085e+04  -0.004    0.997
## as.factor(time2)34 -4.081e+01  1.311e+04  -0.003    0.998
## as.factor(time2)35 -4.084e+01  1.173e+04  -0.003    0.997
## as.factor(time2)36 -4.121e+01  1.252e+04  -0.003    0.997
## nodes               2.110e-02  4.982e-02   0.424    0.672
## as.factor(extent)2  1.679e+01  1.723e+03   0.010    0.992
## as.factor(extent)3  1.636e+01  1.723e+03   0.009    0.992
## as.factor(extent)4  1.692e+01  1.723e+03   0.010    0.992
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1261.99  on 910  degrees of freedom
## Residual deviance:  330.76  on 870  degrees of freedom
##   (18 observations deleted due to missingness)
## AIC: 412.76
##
## Number of Fisher Scoring iterations: 19
```

```
table(colondata$time2[colondata$time2>=20],colondata$status[colondata$time2>=20])
```

```
##
##       0  1
##   20 20 12
##   21 15  9
##   22 27  4
##   23 46  7
##   24 76  6
##   25 46  4
##   26 42  1
##   27 43  2
##   28 39  4
##   29 25  1
##   30 39  2
##   31 17  1
##   32 17  1
##   33  8  0
##   34  3  0
##   35  5  0
##   36  4  0
```

**Employ Method 2 - Quadratic Function for Time:**

```
colondata$time2cen = colondata$time2-mean(colondata$time2)  # first we need to mean center
out2 = glm(status~time2cen+I(time2cen^2)+nodes+as.factor(extent), family=binomial, data=colondat
a)  # quadratic model
summary(out2)
```

```
##
## Call:
## glm(formula = status ~ time2cen + I(time2cen^2) + nodes + as.factor(extent),
##     family = binomial, data = colondata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8658  -0.3998  -0.1917   0.0262   2.7773
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.61577    1.13900  -0.541  0.58877
## time2cen           -0.55591    0.06273  -8.861  < 2e-16 ***
## I(time2cen^2)       0.01751    0.00589   2.972  0.00296 **
## nodes               0.05249    0.04247   1.236  0.21640
## as.factor(extent)2  1.32822    1.17906   1.127  0.25995
## as.factor(extent)3  0.96512    1.14230   0.845  0.39817
## as.factor(extent)4  1.76092    1.32420   1.330  0.18358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1261.99  on 910  degrees of freedom
## Residual deviance:  374.54  on 904  degrees of freedom
##   (18 observations deleted due to missingness)
## AIC: 388.54
##
## Number of Fisher Scoring iterations: 8
```

**Now predict Survival Probability and Calc 95% CI's given nodes=2 and extent=muscle:**

```
predicttime = c(4,8,12,16,20,24)-mean(colondata$time2)  # need to take into account mean centeri
ng for predict time periods too
eta=predict(out2, newdata=data.frame(time2cen=predicttime,nodes=2, extent=2), se.fit=TRUE)
prob = 1/(1+exp(-eta[[1]]))
round(prob,4)
```

```
##      1      2      3      4      5      6
## 1.0000 0.9998 0.9934 0.8988 0.4799 0.1437
```

```
probcihi = 1/(1+exp(-eta[[1]]-1.96*eta[[2]]))
probcilow = 1/(1+exp(-eta[[1]]+1.96*eta[[2]]))
round(probcilow,4)
```

```
##      1      2      3      4      5      6
## 0.9997 0.9968 0.9688 0.7788 0.3144 0.0798
```

```
round(probcihi,4)
```

```
##      1      2      3      4      5      6
## 1.0000 1.0000 0.9986 0.9573 0.6499 0.2451
```

| Quarter | Estimate | 95% Confidence Interval |
|---|---|---|
| 4 | 1.0000 | (0.9997,1.0000) |
| 8 | 0.9998 | (0.9968,1.0000) |
| 12 | 0.9934 | (0.9688,0.9986) |
| 16 | 0.8988 | (0.7788,0.9573) |
| 20 | 0.4799 | (0.3144,0.6499) |
| 24 | 0.1437 | (0.0798,0.2451) |

**Report Summary:**
Given that we have interval censored data (using quarters instead of days) and many ties for uncensored event data, I first ran the model using method 1 for interval data in the slides using time as a factor variable. Some issues with doing this are that we don't have a lot of events, especially beyond time period 26, so the standard errors were somewhat high resulting in unreasonable estimates, and we also get a large number of model parameters. Also, the AIC is higher at 412.8 when compared to running method 2, using the quadratic function for time instead of using time as a factor variable.

I then reran the analysis using time as a quadratic function. This results in a model with fewer parameters and a lower AIC of 388.5. The estimates for the survival probability and the 95% CI's for each of the listed quarters is shown in the table above (given nodes=2 and extent = muscle).

Notes:
1) I used etype==2 as the initial data filter criteria as we are looking at survival/deaths, rather than simply recurrence of disease.
2) The time variable was mean centered and the round function was used to assign the appropriate quarter (rather than the floor function). If we used floor to assign the quarters, the values in the table above would be slightly to modestly lower than shown.