BIOSTAT 675 – Homework #6

Kyle Kumbier

Problem 1

End-stage renal disease (ESRD; also referred to as 'renal failure') is increasing in many countries worldwide, including the United States and Canada. Due to the shortfall in the available donor organs, donor kidneys are now being transplanted which would in the past have been discarded; the so-called Expanded Criteria Donor (ECD) kidneys. By definition, ECD kidneys are more likely to suffer graft failure (GF), the condition wherein the transplanted kidney stops functioning sufficiently.

A random sample of U.S. transplant recipients was assembled, in order to study the effects on the mortality hazard of ECD (vs non-ECD) kidneys and graft failure (GF).

Data are contained in the file "kidney-ECD-1.sas7bdat", with fields:

IDNUM: patient ID number
ECD: equals 1 for an ECD kidney, and 0 for non-ECD
time-to-GF: time until graft failure (missing, if GF did not occur)
time-to-death: time until death (missing, if death not observed)
time-to-censor: potential time until censoring (non-missing for all patients)
AGE: age at transplant (years)
SEX
DIABETES: indicator that diabetes was the cause of renal failure
COMORBID: number of comorbid conditions (illnesses, not counting ESRD, existing at the time of transplant)

For each of the following parts, submit your code and output as an appendix.

(a) Fit a model with graft failure (GF) as a time-dependent covariate.

See appendix.

(b) Interpret the hazard ratio for GF.

<u>GF Hazard Ratio Interpretation</u>: Those who experience a graft failure have a 67% increase in death hazard in comparison to those who do not experience a graft failure, with all other covariates held constant.

(c) Compare the ECD hazard ratios from models (HW5.3a) and (HW6.1a). What does this tell you about the nature of the ECD effect?

In the previous homework we found that ECD had a significant effect on mortality when we did not adjust for graft failure; the hazard ratio was 1.13 with a significant p-value less than 0.001.

However, after we adjust for graft failure, we find that ECD does not have a significant effect on mortality; the hazard ratio was 0.99 with a large p-value of 0.73.

This tells us that the ECD effect is likely driven by the GF effect. It appears that those with an ECD kidney are more likely to experience a graft failure, thus, increasing the death hazard.

Appendix

```
# put into time-dependent format
kidney$t1 <- 0
kidney$t2 <- 0
kidney$GF <- 0
kidney$death_new <- 0
kidA <- kidney
kidB <- kidney
kidC <- kidney
for(i in 1:nrow(kidney)){
  if(is.na(kidney$time_to_GF[i]) == F){
    kidA$t1[i] <- 0
    kidB$t1[i] <- kidney$time_to_GF[i]
    kidA$t2[i] <- kidney$time_to_GF[i]
    kidB$t2[i] <- kidney$time_to_event[i]
    kidA$GF[i] <- 0
    kidB$GF[i] <- 1
    kidA$death_new[i] <- 0
    kidB$death_new[i] <- kidney$death[i]
  } else{
    kidC$t1[i] <- 0
    kidC$t2[i] <- kidney$time_to_event[i]
    kidC$GF[i] <- 0
    kidC$death_new[i] <- kidney$death[i]
  }
}
kidA <- subset(kidA,is.na(kidney$time_to_GF) == F)
kidB <- subset(kidB,is.na(kidney$time_to_GF) == F)
kidC <- subset(kidC,is.na(kidney$time_to_GF) == T)
kidney_TD <- rbind(kidA,kidB,kidC)
kidney_TD <- arrange(kidney_TD, idnum)

# now fit the model
model02 <- coxph(data = kidney_TD,
                 formula = Surv(t1, t2, death_new) ~ age +
                   male + diabetes + comorbid + ECD + GF)
summary(model02)
```

```
> summary(model02)
Call:
coxph(formula = Surv(t1, t2, death_new) ~ age + male + diabetes +
    comorbid + ECD + GF, data = kidney_TD)

  n= 10504, number of events= 5675

              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.021034  1.021257  0.001351 15.575  < 2e-16 ***
male      0.149355  1.161085  0.026813  5.570 2.54e-08 ***
diabetes  0.490044  1.632388  0.032192 15.223  < 2e-16 ***
comorbid  0.135163  1.144724  0.013275 10.182  < 2e-16 ***
ECD      -0.010456  0.989599  0.030294 -0.345     0.73
GF        0.515169  1.673921  0.033496 15.380  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 2

Asthma remains one of the most common chronic childhood illnesses, and a leading cause of hospital admissions. The file, *asthma_1.sas7bdat* contains data obtained from Alberta Health, a provincial health organization in Canada. Through linkages to several health administrative databases, a random sample of children born between January 1, 1995 and December 31, 1999 were retrospectively followed until their first physician visit for asthma or the end of the observation period (December 31, 1999), whichever occurred first.

The file contains the following fields: IDNUM (patient ID number), DT_BIRTH (date of birth), DT_ASTHMA (date of first physician visit reporting asthma), BWT (birth weight; kg), SEX, URBAN (indicator for living in an urban (as opposed to rural) area), RESP_DIST (indicator for experiencing respiratory distress at birth).

(a) Fit a model which assumes proportionality for all covariates. Code BWT as a continuous covariate. Which factors appear to significantly affect asthma incidence?

```
> summary(model03)
Call:
coxph(formula = Surv(time_to_event, asth) ~ bwt + male + urban + 
    resp_dist, data = asthma)

  n= 1567, number of events= 679 

              coef exp(coef) se(coef)      z Pr(>|z|)    
bwt       -0.09128   0.91276  0.04469 -2.042   0.0411 *  
male       0.53352   1.70491  0.08302  6.426 1.31e-10 ***
urban     -0.05967   0.94208  0.08768 -0.681   0.4961    
resp_dist  0.66090   1.93654  0.11111  5.948 2.71e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
bwt          0.9128     1.0956    0.8362    0.9963
male         1.7049     0.5865    1.4489    2.0062
urban        0.9421     1.0615    0.7933    1.1187
resp_dist    1.9365     0.5164    1.5576    2.4077

Concordance= 0.593  (se = 0.012 )
Rsquare= 0.043   (max possible= 0.997 )
Likelihood ratio test= 68.82  on 4 df,   p=4e-14
Wald test            = 72.24  on 4 df,   p=8e-15
Score (logrank) test = 73.79  on 4 df,   p=4e-15
```

It appears BWT, SEX, and RESP_DIST significantly affect asthma incidence, using a 0.05 level of significance under the proportionality assumption.

(b) Repeat (a), but code BWT using an indicator for *low birth weight* (defined as weighing ≤ 2.5 kg). Compare the parameter estimates with those from (a) and comment on the similarities and/or differences.

```
> summary(model04)
Call:
coxph(formula = Surv(time_to_event, asth) ~ lbw + male + urban + 
    resp_dist, data = asthma)

  n= 1567, number of events= 679 
```

```
               coef exp(coef) se(coef)      z Pr(>|z|)
lbw        0.25327   1.28824  0.09760  2.595  0.00946 **
male       0.48515   1.62441  0.07996  6.067 1.30e-09 ***
urban     -0.05306   0.94832  0.08764 -0.605  0.54488
resp_dist  0.64348   1.90309  0.11158  5.767 8.07e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
lbw          1.2882     0.7763    1.0639     1.560
male         1.6244     0.6156    1.3888     1.900
urban        0.9483     1.0545    0.7986     1.126
resp_dist    1.9031     0.5255    1.5293     2.368

Concordance= 0.593  (se = 0.012 )
Rsquare= 0.044   (max possible= 0.997 )
Likelihood ratio test= 71.05  on 4 df,   p=1e-14
Wald test            = 74.35  on 4 df,   p=3e-15
Score (logrank) test = 75.88  on 4 df,   p=1e-15
```

The results of the Wald tests remained the same for each covariate. The parameter estimates for SEX, URBAN, and RESP_DIST don't change all that much, but changing BWT from a continuous variable to a categorical variable had a noticeable effect. The parameter estimate changed from -0.091 to 0.253, and it becomes a much more significant predictor as the p-value decreases from 0.041 to 0.009.

(c) Suppose, for part (c) only, the RESP_DIST was of no interest, except as an adjustment covariate. Suppose also that you have no knowledge (and no desire to learn) about the nature of the non-proportionality. Fit an appropriate model and briefly defend your choice.

```
> summary(model05)
Call:
coxph(formula = Surv(time_to_event, asth) ~ lbw + male + urban + 
    strata(resp_dist), data = asthma)

  n= 1567, number of events= 679 

          coef exp(coef) se(coef)      z Pr(>|z|)    
lbw    0.24803   1.28150  0.09774  2.538   0.0112 *  
male   0.47692   1.61110  0.07994  5.966 2.43e-09 ***
urban -0.05092   0.95035  0.08764 -0.581   0.5612    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
lbw      1.2815     0.7803    1.0581     1.552
male     1.6111     0.6207    1.3775     1.884
urban    0.9504     1.0522    0.8004     1.128

Concordance= 0.566  (se = 0.013 )
Rsquare= 0.027   (max possible= 0.996 )
Likelihood ratio test= 43.22  on 3 df,   p=2e-09
Wald test            = 42.31  on 3 df,   p=3e-09
Score (logrank) test = 43  on 3 df,   p=2e-09
```

For this part I fit a model that stratifies by RESP_DIST. This allows us to non-parametrically adjust for RESP_DIST, so lacking knowledge of the nature of the nonproportionality isn't a big deal. When you stratify by a variable, you're unable to estimate its effect, but that's not something we're worried about either.

(d) Fit a model which assumes that the RESP_DIST effect follows a year-specific step function. Interpret the RESP_DIST effect, as estimated from this model.

```
> summary(model06)
Call:
coxph(formula = Surv(tstart, time_to_event, asth) ~ lbw + male + 
    urban + rd1 + rd2 + rd3 + rd45, data = asthma_td)

  n= 3376, number of events= 679 

          coef exp(coef) se(coef)      z Pr(>|z|)    
lbw    0.25078   1.28503  0.09768  2.567   0.0102 *  
male   0.47926   1.61489  0.07994  5.995 2.03e-09 ***
urban -0.05113   0.95016  0.08764 -0.583   0.5596    
rd1    0.91820   2.50478  0.13684  6.710 1.94e-11 ***
rd2    0.25754   1.29374  0.24874  1.035   0.3005    
rd3    0.41868   1.51996  0.34931  1.199   0.2307    
rd45  -1.11922   0.32653  1.00956 -1.109   0.2676    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
lbw      1.2850     0.7782   1.06113     1.556
male     1.6149     0.6192   1.38069     1.889
urban    0.9502     1.0525   0.80019     1.128
rd1      2.5048     0.3992   1.91556     3.275
rd2      1.2937     0.7730   0.79454     2.107
rd3      1.5200     0.6579   0.76647     3.014
rd45     0.3265     3.0625   0.04514     2.362

Concordance= 0.596  (se = 0.012 )
Rsquare= 0.025   (max possible= 0.932 )
Likelihood ratio test= 83.75  on 7 df,   p=2e-15
Wald test            = 90.06  on 7 df,   p=<2e-16
Score (logrank) test = 93.88  on 7 df,   p=<2e-16
```

Note: There was insufficient data to support the interval from 4 to 5 years, so we decided to merge the last two intervals together (denoted 'rd45').

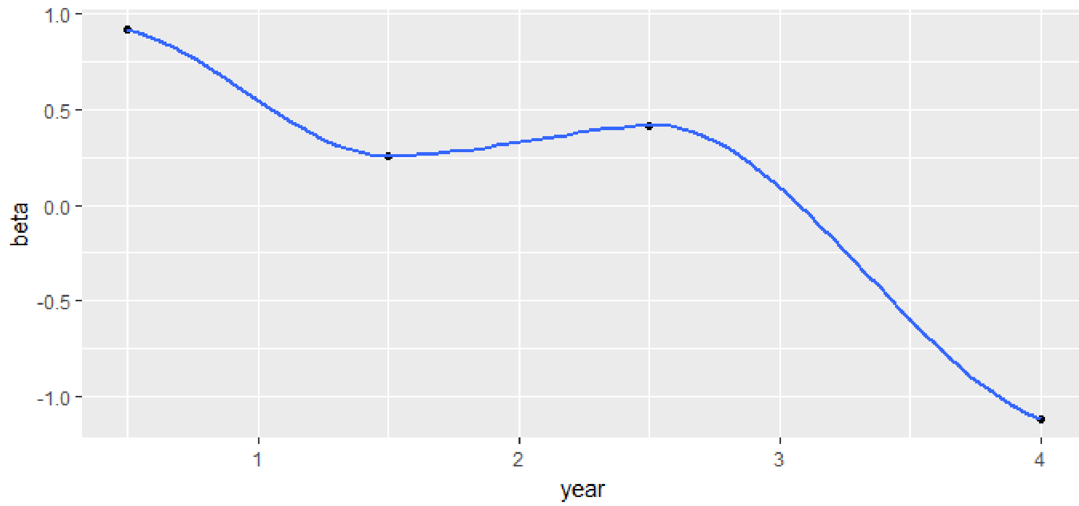
The hazard ratio during the interval from 0 to 1 years was 2.50. Thus, during the first year, children with RESP_DIST have 2.5 times the asthma hazard as children without RESP_DIST, holding all other covariates constant.

The hazard ratio during the interval from 1 to 2 years was 1.29. Thus, during the second year, children with RESP_DIST have a 29% increase in asthma hazard compared to children without RESP_DIST, holding all other covariates constant.

The hazard ratio during the interval from 2 to 3 years was 1.52. Thus, during the third year, children with RESP_DIST have a 52% increase in asthma hazard compared to children without RESP_DIST, holding all other covariates constant.

The hazard ratio during the interval from 3 to 5 years was 0.33. Thus, during the fourth and fifth years, children with RESP_DIST have a 67% decrease in asthma hazard compared to children without RESP_DIST, holding all other covariates constant.

(e) Plot the age-specific RESP_DIST against the year mid-points. Describe the shape of the plot and its implications (if any) for modelling the RESP_DIST effect.



The increase in asthma hazard associated with RESP_DIST seems to be highest within the first year, approximately constant between years 1 and 3, and after year 3 it appears the hazard may actually be lower for those with RESP_DIST. This plot implies that assuming the log hazard ratio decreases linearly in time may be reasonable assumption (which we will actually implement in part (f)).

(f) Fit a model wherein the RESP_DIST regression coefficient is assumed to change linearly with age (scaled to years). Interpret your parameter estimates.

```
> summary(model07)
Call:
coxph(formula = Surv(time_to_event, asth) ~ lbw + male + urban +
    resp_dist + tt(resp_dist), data = asthma, tt = function(x,
    t, ...) x * t/365)

  n= 1567, number of events= 679

                  coef exp(coef) se(coef)      z Pr(>|z|)
lbw            0.24779   1.28119  0.09771  2.536  0.01121 *
male           0.47831   1.61335  0.07991  5.985 2.16e-09 ***
urban         -0.05145   0.94985  0.08763 -0.587  0.55712
resp_dist      1.10252   3.01174  0.16789  6.567 5.14e-11 ***
tt(resp_dist) -0.47199   0.62376  0.14630 -3.226  0.00125 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
lbw              1.2812     0.7805    1.0579    1.5516
male             1.6133     0.6198    1.3794    1.8869
urban            0.9498     1.0528    0.7999    1.1278
resp_dist        3.0117     0.3320    2.1672    4.1853
tt(resp_dist)    0.6238     1.6032    0.4683    0.8309

Concordance= 0.595  (se = 0.233 )
Rsquare= 0.052   (max possible= 0.997 )
Likelihood ratio test= 83.45  on 5 df,   p=<2e-16
Wald test            = 90.39  on 5 df,   p=<2e-16
Score (logrank) test = 93.7  on 5 df,   p=<2e-16
```

$e^{\beta_{LBW}} = 1.28$; Children with a low birth weight have a 28% increase in asthma hazard compared to children with a normal birth weight, holding all other covariates constant.

$e^{\beta_{SEX}} = 1.61$; Males have a 61% increase in asthma hazard compared to females, holding all other covariates constant.

$e^{\beta_{URB}} = 0.95$; Children in urban regions have a 5% decrease in asthma hazard compared to those living in rural regions, holding all other covariates constant.

$e^{\beta_{RD}} = 3.01$; For children just born (age = 0), those with RESP_DIST have 3.01 times the asthma hazard as children without RESP_DIST, holding all other covariates constant.

$e^{\beta_{RD\_T}} = 0.62$; The hazard ratio for RESP_DIST decreases by 38% for each year increase in age.

(g) Based on the model in (f), estimate the age at which children with and without RESP_DIST have equal asthma hazard.

Children with and without RESP_DIST will have equal asthma hazard when $\beta_{RD} + age * \beta_{RD\_T} = 0$.

After substituting and solving for 'age', we find:

$$age = \frac{-\beta_{RD}}{\beta_{RD\_T}} = \frac{-1.10252}{-0.47199} = 2.34$$

The asthma hazard for children with and without RESP_DIST are the same at age 2.34 years, holding all other covariates constant.

Appendix

```
# Create new variables
# Time-to-event
max_days <- as.numeric(as.Date("1999-12-31") - as.Date("1960-01-01"))
asthma$time_to_event <- 0
for(i in 1:nrow(asthma)){
  if(is.na(asthma$dt_asthma[i])){
    asthma$time_to_event[i] <- max_days - asthma$dt_birth[i]
  } else{
    asthma$time_to_event[i] <- asthma$dt_asthma[i] - asthma$dt_birth[i]
  }
}
# Male indicator
asthma <- mutate(asthma, male = 1*(sex == "M"))
# Asthma indicator
asthma <- mutate(asthma, asth = 1*!(is.na(dt_asthma)))

# (a) Fit a model which assumes proportionality for all covariates. Code BWT as a
#     continuous covariate. Which factors appear to significantly affect asthma
#     incidence?
model03 <- coxph(data = asthma,
                 formula = Surv(time_to_event, asth) ~ bwt + male + urban + resp_dist)
summary(model03)
```

```
# (b) Repeat (a), but code BWT using an indicator for low birth weight (defined as
#     weighing <= 2.5 kg). Compare the parameter estimates with those from (a) and
#     comment on the similarities and/or differences.
asthma <- mutate(asthma, lbw = 1*(bwt <= 2.5))
model04 <- coxph(data = asthma,
                 formula = Surv(time_to_event, asth) ~ lbw + male + urban + resp_dist)
summary(model04)

# (c) Suppose, for part (c) only, that RESP_DIST was of no interest, except as an
#     adjustment covariate. Suppose also that you have no knowledge (an no desire
#     to learn) about the nature of theh non-proportionality. Fit an appropriate
#     model, and briefly defend your choice.
model05 <- coxph(data = asthma,
                 formula = Surv(time_to_event, asth) ~ lbw + male + urban +
                   strata(resp_dist))
summary(model05)

# (d) Fit a model which assumes that the RESP_DIST effect follows a year-specific
#     step function. Interpret the RESP_DIST effect, as estimated from this model.
asthma_td <- survSplit(Surv(time_to_event,asth) ~ lbw + male + urban + resp_dist,
                       data = asthma,
                       cut = c(365,730,1095),
                       episode = "tgroup")
asthma_td <- mutate(asthma_td,
                    rd1 = resp_dist*(tgroup == 1),
                    rd2 = resp_dist*(tgroup == 2),
                    rd3 = resp_dist*(tgroup == 3),
                    rd45 = resp_dist*(tgroup == 4))
model06 <- coxph(data = asthma_td,
                 formula = Surv(tstart, time_to_event, asth) ~ lbw + male + urban +
                   rd1 + rd2 + rd3 + rd45)
summary(model06)

# (e) Plot the age-specific RESP_DIST against the year mid-points. Describe the shape
#     of the plot and its implications (if any) for modelling the RESP_DIST effect.
plot_data <- data.frame(c(0.5,1.5,2.5,4),c(0.918,0.258,0.419,-1.119))
colnames(plot_data) <- c("year","beta")
library(ggplot2)
plot01 <- ggplot(data = plot_data, aes(x=year, y=beta)) +
  geom_point() +
  geom_smooth(se = F)
plot01

# (f) Fit a model wherein the RESP_DIST regression coefficient is assumed to change
#     linearly with age (scaled to years). Interpret your parameter estimates.
model07 <- coxph(data = asthma,
                 formula = Surv(time_to_event, asth) ~ lbw + male + urban + resp_dist +
                   tt(resp_dist),
                 tt = function(x,t,...) x*t/365)
summary(model07)
```