Solutions to tutorial 4b

Question 1

Here are the R commands to make the plots for the aml data (the other data sets are similar).

```r
library(survival)
attach(aml)
S=Surv(time,status)
X=(x=="Maintained")+0

# makes the E vs O plot:
phm=coxph(S~X)
d.phm=coxph.detail(phm)
times=c(0,d.phm$t)
h0=c(0,d.phm$hazard)
S0=exp(-cumsum(h0))
beta=phm$coef
x1=0-mean(X)
x2=1-mean(X)
Sx1=S0^exp(beta*x1)
Sx2=S0^exp(beta*x2)
km=survfit(S~X)
plot(km,col=1:2,lty=2,xlab='t',ylab=expression(hat(S)(t)))
lines(times,Sx1,col=1,type='s')
lines(times,Sx2,col=2,type='s')

# makes the log--log plot:
km$label=c(rep(1,km$strata[1]),rep(2,km$strata[2]))
t1=c(0,subset(km$time,km$label==1))
t2=c(0,subset(km$time,km$label==2))
St1=c(0,subset(km$surv,km$label==1))
St2=c(0,subset(km$surv,km$label==2))
plot(t1,-log(-log(St1)),col=1,type='s',xlab='t',ylab=expression(hat(S)(t)),ylim=c(0,4))
```

32
Plots for three data sets look as follows (dashed lines are observed using KM, solid are expected using PHM).

It appears that there is no problem with the proportional hazards assumption for the aml data above.
These **kidney** data look to violate the proportional hazards assumption: initially the two groups are more different than expected (the observed curves are more extreme than the expected ones), while later the two groups are more similar than expected (the dashed lines fall within the solid ones). This is backed up by the log–log plot, which looks anything but parallel.

It is hard to tell from these plots if the **ovarian** data satisfy the proportional hazards assumption. There are noticeable discrepancies between the observed and expected curves within treatments, but this could be due to the sample size (note how few observations there are from the big jumps the curves make).

**Question 2**

Here are the R commands and output to test the proportional hazards assumption.

```r
> library(survival)
>
> #aml
```
> attach(aml)
> S=Surv(time,status)
> X=(x=="Maintained")+0
> phm=coxph(S~X)
> cox.zph(phm,transform='rank',global=FALSE)

    rho  chisq   p
X  -0.00772 0.00105 0.974
> detach(aml)

>
> #exp vs obs
> #kidney
> attach(kidney)
> S=Surv(time,status)
> X=sex
> phm=coxph(S~X)
> cox.zph(phm,transform='rank',global=FALSE)

    rho  chisq   p
X    0.435   10.8 0.00103
> detach(kidney)

>
> #ovarian
> attach(ovarian)
> S=Surv(futime,fustat)
> X=(age<median(age))+0
> phm=coxph(S~X)
> cox.zph(phm,transform='rank',global=FALSE)

    rho  chisq   p
X    0.416    1.70 0.192
> detach(ovarian)

So, we have no reason to reject the PHA for either the aml or ovarian data, but strong reason to reject it for the kidney data. This seems consistent with the graphs (as we would expect!).

**Question 3**

We can load the data using the following commands.
library(survival,MASS)
cols=c('id','age','sex','hrate','sysbp','diabp','bmi','history','afb',
'  shock','chf','av3','miord','mitype','year','c16','c17','c18','c19',
'  c20','t','delta')
attach(read.table('http://www.umass.edu/statdata/statdata/data/whas500.dat',
  col.names=cols))
S=Surv(t,delta)

The continuous variates in the model we developed in the notes are age, hrate, diabp and bmi. Consider age first:

```
> m0=coxph(S~age+sex+hrate+diabp+bmi+shock+chf+mitype)
> m1=coxph(S~I(age^2)+sex+hrate+diabp+bmi+shock+chf+mitype)
> m2=coxph(S~age+I(age^2)+sex+hrate+diabp+bmi+shock+chf+mitype)
> m3=coxph(S~I(log(age))+sex+hrate+diabp+bmi+shock+chf+mitype)
> m4=coxph(S~I(sqrt(age))+sex+hrate+diabp+bmi+shock+chf+mitype)
> m5=coxph(S~I(1/(age))+sex+hrate+diabp+bmi+shock+chf+mitype)
> aic0=2*(length(m0$coeff) - m0$logl[2])
> aic1=2*(length(m1$coeff) - m1$logl[2])
> aic2=2*(length(m2$coeff) - m2$logl[2])
> aic3=2*(length(m3$coeff) - m3$logl[2])
> aic4=2*(length(m4$coeff) - m4$logl[2])
> aic5=2*(length(m5$coeff) - m5$logl[2])
> c(aic0,aic1,aic2,aic3,aic4,aic5)-max(aic0,aic1,aic2,aic3,aic4,aic5)
[1] -2.7170093 -0.2384528 -1.4507079 -2.7368634 -3.0593355 0.0000000
```

Model m5 fits the best, but the square in age is also pretty good. Using m5
as the basis for further transformations, and considering next heart rate, we
have

```
> m0=coxph(S~hrate+sex+I(1/(age))+diabp+bmi+shock+chf+mitype)
> m1=coxph(S~I(hrate^2)+sex+I(1/(age))+diabp+bmi+shock+chf+mitype)
Warning messages:
1: In fitter(X, Y, strats, offset, init, control, weights = weights, :
   Loglik converged before variable 1,2,4,5,6,7,8 ; beta may be infinite.
```
The parameter estimation routine for models m1 and m2 failed to converge (due to identifiability issues, perhaps), so we drop these models and consider the others only. Of these, m5 fitted best, i.e. again the inverse of the coefficient fits better. We consider next diastolic blood pressure:

```r
m0=coxph(S~diabp+sex+I(1/(hrate))+I(1/(age))+bmi+shock+chf+mitype)
> m1=coxph(S~I(diabp^2)+sex+I(1/(hrate))+I(1/(age))+bmi+shock+chf+mitype)
Warning messages:
1: In fitter(X, Y, strats, offset, init, control, weights = weights, :
   Loglik converged before variable 1,2,5,6,7,8,9 ; beta may be infinite.
2: In coxph(S ~ diabp + I(diabp^2) + sex + I(1/(hrate)) + I(1/(age)) + bmi + :
   X matrix deemed to be singular; variable 3 4
Warning messages:
1: In fitter(X, Y, strats, offset, init, control, weights = weights, :
   Loglik converged before variable 1,2,3,6,7,8,9 ; beta may be infinite.
2: In coxph(S ~ diabp + I(diabp^2) + sex + I(1/(hrate)) + I(1/(age)) + :
   X matrix deemed to be singular; variable 3 4
```
X matrix deemed to be singular; variable 4 5

\[
\begin{align*}
\text{m3} &= \text{coxph}(S \sim \text{I(log(diabp))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(age))} + \text{bmi} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m4} &= \text{coxph}(S \sim \text{I(sqrt(diabp))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(age))} + \text{bmi} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m5} &= \text{coxph}(S \sim \text{I(1/(diabp))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(age))} + \text{bmi} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{aic0} &= 2 \times (\text{length(m0$coeff}) - \text{m0$logl[2]}) \\
\text{aic1} &= 2 \times (\text{length(m1$coeff}) - \text{m1$logl[2]}) \\
\text{aic2} &= 2 \times (\text{length(m2$coeff}) - \text{m2$logl[2]}) \\
\text{aic3} &= 2 \times (\text{length(m3$coeff}) - \text{m3$logl[2]}) \\
\text{aic4} &= 2 \times (\text{length(m4$coeff}) - \text{m4$logl[2]}) \\
\text{aic5} &= 2 \times (\text{length(m5$coeff}) - \text{m5$logl[2]}) \\
\text{c(aic0,aic3,aic4,aic5)} \equiv \text{max(aic0,aic3,aic4,aic5)} \\
\end{align*}
\]

\[\begin{array}{l}
\text{[1]} -5.245742 -5.677699 -5.686878 0.000000 \\
\end{array}\]

Again, terms with the square do not lead to model convergence, and the inverse of diastolic blood pressure is more predictive than the linear term.

Consider body mass index now:

\[
\begin{align*}
\text{m0} &= \text{coxph}(S \sim \text{bmi} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m1} &= \text{coxph}(S \sim \text{I(bmi$^2)} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m2} &= \text{coxph}(S \sim \text{bmi} + \text{I(bmi$^2)} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m3} &= \text{coxph}(S \sim \text{I(log(bmi))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m4} &= \text{coxph}(S \sim \text{I(sqrt(bmi))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{m5} &= \text{coxph}(S \sim \text{I(1/(bmi))} + \text{sex} + \text{I(1/(hrate))} + \text{I(1/(diabp))} + \text{I(1/(age))} + \text{shock} + \text{chf} + \text{mitype}) \\
\text{aic0} &= 2 \times (\text{length(m0$coeff}) - \text{m0$logl[2]}) \\
\text{aic1} &= 2 \times (\text{length(m1$coeff}) - \text{m1$logl[2]}) \\
\text{aic2} &= 2 \times (\text{length(m2$coeff}) - \text{m2$logl[2]}) \\
\text{aic3} &= 2 \times (\text{length(m3$coeff}) - \text{m3$logl[2]}) \\
\text{aic4} &= 2 \times (\text{length(m4$coeff}) - \text{m4$logl[2]}) \\
\text{aic5} &= 2 \times (\text{length(m5$coeff}) - \text{m5$logl[2]}) \\
\text{c(aic0,aic1,aic2,aic3,aic4,aic5)} \equiv \text{max(aic0,aic1,aic2,aic3,aic4,aic5)} \\
\end{align*}
\]

\[\begin{array}{l}
\text{[1]} -3.469758 0.000000 -12.354388 -6.651924 -5.133807 -8.997298 \\
\end{array}\]
Here the term with a square in \texttt{bmi} fits best, though note that the quadratic doesn’t.

So, all the continuous covariates would give a better fit if transformed. Note, though, that their interpretation is much harder. We can no longer talk of the hazard ratio for an increase in \texttt{bmi} of one, for example, as this hazard ratio depends on the value of \texttt{bmi} used as a reference point. To compare the effect of different values we might create plots. Here is an example for \texttt{bmi} with commands.

\begin{verbatim}
b=13:45
mean0=exp(b*m0$coeff[1])
mean1=exp((b^2)*m1$coeff[1])
mean2=exp(b*m2$coeff[1]+(b^2)*m2$coeff[2])
mean3=exp(log(b)*m3$coeff[1])
mean4=exp(sqrt(b)*m4$coeff[1])
mean5=exp((1/b)*m5$coeff[1])

#centre around bmi=22
mean0=mean0/mean0[10]
mean1=mean1/mean1[10]
mean2=mean2/mean2[10]
mean3=mean3/mean3[10]
mean4=mean4/mean4[10]
mean5=mean5/mean5[10]

ylm=range(0,2)
xlm=range(b)

plot(0,0,pch='',xlim=xlm,ylim=ylm,xlab='BMI',ylab='Estimated hazard ratio (relative to BMI=22)')
polygon(c(xlm[1],18.5,18.5,xlm[1],xlm[1]),c(ylm[1],ylm[1],ylm[2], ylm[2],ylm[1]),col=gray(0.9),border=NA)
polygon(c(25,30,30,25,25),c(ylm[1],ylm[1],ylm[2],ylm[2],ylm[1]),
   col=gray(0.9),border=NA)
polygon(c(30,xlm[2],xlm[2],30,30),c(ylm[1],ylm[1],ylm[2],ylm[2], ylm[1]),col=gray(0.8),border=NA)
\end{verbatim}
lines(b,mean0,lty=1,type='l')
lines(b,mean1,lty=2)
lines(b,mean2,lty=3)
lines(b,mean3,lty=4)
lines(b,mean4,lty=5)
lines(b,mean5,lty=6)
den=density(bmi); den$y=0.2*den$y/max(den$y)
lines(den,col=gray(0.5)); text(26,0.05,"bmi distribution",col=gray(0.5))
text(0.5*(xlm[1]+18.5),0.3,"underweight")
text(0.5*(25+30),0.3,"overweight")
text(0.5*(30+xlm[2]),0.3,"obese")
legend(27.5,1.8,c('bmi',expression(bmi^2),expression(bmi+bmi^2),
expression(log(bmi)),expression(sqrt(bmi)),expression(1/bmi)),
lty=1:6,bg=gray(0.95))
Note that within the range of the observed BMI, the estimates of the hazard ratio are all very similar. [Note: a better graph would have made use of colour or used fewer lines.]