4250 Model Building and Assessment Primer

This document will go through a brief example on all aspects of fitting and evaluating a Cox proportional hazards model.

Data Set (colon)

Description

These are data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound previously used to treat worm infestations in animals; 5-FU is a moderately toxic (as these things go) chemotherapy agent. There are two records per person, one for recurrence and one for death.

Variables

- id: id
- study: 1 for all patients
- rx: Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU *sex: 1=male* age: in years
- obstruct: obstruction of colon by tumour
- perfor: perforation of colon
- adhere: adherence to nearby organs
- time: days until event or censoring
- status: censoring status
- extent: Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures)
- surg: time from surgery to registration (0=short, 1=long)
- node4: more than 4 positive lymph nodes

The event time is **time** (days until event or censoring), censor status is **status**. Here's what the data set looks like

```
library(survival)
library(survminer) # for ggcoxfunctional
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

```
colonMod <- read.csv("/Users/andres_th14/Downloads/colonModified.csv",header=TRUE)
head(colonMod)
```

```
##        rx    sex age obstruct perfor adhere status extent surg node4 time
## 1 Lev+5FU   male  49        0      0      0      0      3    0     1 2726
## 2 Lev+5FU female  48        0      0      0      0      3    0     1 2527
## 3     Obs female  50        0      0      0      1      3    0     1  499
## 4     Obs   male  57        1      0      0      1      3    0     1   59
## 5     Obs female  56        0      0      0      1      3    0     0  770
## 6     Obs female  60        0      0      0      0      3    0     0 2294
```

Initial Thoughts

possible covariates:

- age
- sex
- rx
- obstruct
- perfor
- adhere
- extent
- surg
- node4

```
# this code will look at one-variable Cox PH models, and report
# the resulting beta coefficients, test statistics, and p-values
testCovariates <- c("age","sex","rx","obstruct","perfor","adhere","extent","surg","node4")
univ_formulas <- sapply(testCovariates,function(x)as.formula(paste('Surv(time,status)~',x)))
univ_models <- lapply(univ_formulas,function(x){coxph(x,data=colonMod)})
# Extract data
univ_results <- lapply(univ_models,function(x){
  x<-summary(x)
  p.value<-signif(x$wald["pvalue"],digits=3)
  wald.test<-signif(x$wald["test"],digits=3)
  beta<-signif(x$coef[1], digits=3);#coeficient beta
  res<-c(beta, wald.test, p.value)
  names(res)<-c("beta", "wald.test","p.value")
  return(res)
#return(exp(cbind(coef(x),confint(x))))
  })
res <- t(as.data.frame(univ_results))
format(as.data.frame(res),scientific=FALSE)
```

```
##               beta wald.test                            p.value
## age       -0.00332      0.95 0.3290000000000000146549439250520
## sex       -0.04660      0.33 0.5679999999999999493738300770930
## rx        -0.47600     27.00 0.0000013499999999999999918495100
## obstruct   0.23900      5.80 0.0160000000000000003330669073880
## perfor     0.07100      0.09 0.7610000000000000097699626167010
## adhere     0.38400     14.30 0.0001559999999999999971758701810
## extent     0.49200     26.40 0.0000002759999999999999832753140
## surg       0.25700      8.34 0.0038700000000000001759703494030
## node4      0.93300    122.00 0.0000000000000000000000000001860
```

So, individually, all covariates appear significantly correlated with survival times, except for age, sex and perfor.

Double Checking Numerical Predictors

We can look at the univariate relationships between numerical covariates and survival times by computing martingale residuals and plotting them. Let's do this for the numerical covariates.

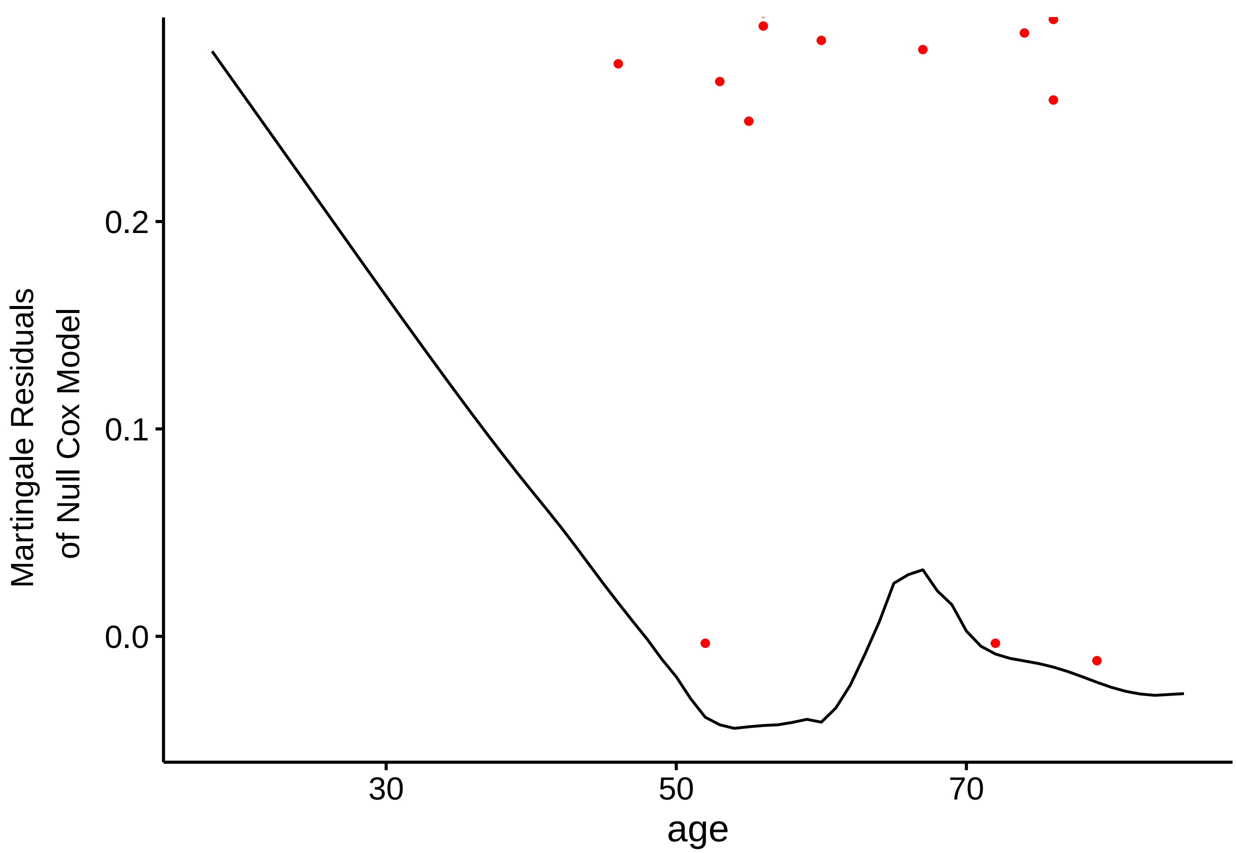
Age

```
fit <- coxph(Surv(time,status)~age,data=colonMod)
fit
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age, data = colonMod)
##
##          coef exp(coef)  se(coef)      z     p
## age -0.003324  0.996682  0.003406 -0.976 0.329
##
## Likelihood ratio test=0.95  on 1 df, p=0.3309
## n= 1200, number of events= 601
```

```
ggcoxfunctional(Surv(time, status) ~ age, data = colonMod)
```

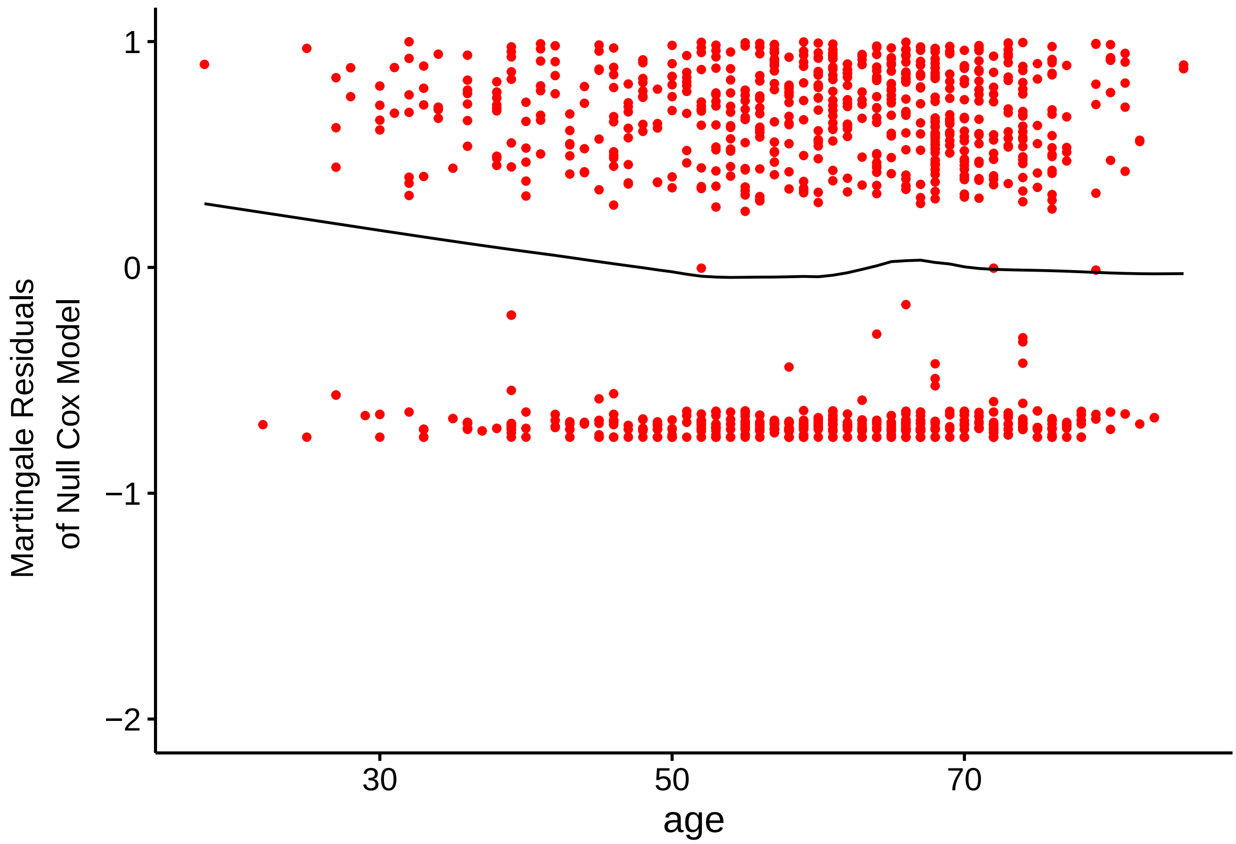
```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```



Age looks roughly linear; I would probably try to put some piecewise function possibly in the cuttoff of the covariate age.

```
ggcoxfunctional(Surv(time,status)~age,data=colonMod,ylim=c(-2,1))
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```




```
# Change Z.range to your desired values
Z.range<- seq(45,55,1) # possible age values
mincutoff <- 0
minaic <- 1e10
for(cutoff in Z.range){
  # Change time, status, age, colon to your data set and variable
fit.temp <- coxph(Surv(time,status)~age*I(age>cutoff),data=colonMod)
aic <- AIC(fit.temp)
cat("cutoff: Z >", cutoff, "; AIC =", aic, "\n")
if(aic < minaic){
  mincutoff <- cutoff
  minaic <- aic
}
}
```

```
## cutoff: Z > 45 ; AIC = 8108.526
```

```
## cutoff: Z > 46 ; AIC = 8108.855
## cutoff: Z > 47 ; AIC = 8108.417
## cutoff: Z > 48 ; AIC = 8108.659
## cutoff: Z > 49 ; AIC = 8108.374
## cutoff: Z > 50 ; AIC = 8108.598
## cutoff: Z > 51 ; AIC = 8109.026
## cutoff: Z > 52 ; AIC = 8109.024
## cutoff: Z > 53 ; AIC = 8109.024
## cutoff: Z > 54 ; AIC = 8109
## cutoff: Z > 55 ; AIC = 8109.075
```

```
cat("optimal cutoff: Z >", mincutoff, "; AIC =", minaic, "\n")
```

```
## optimal cutoff: Z > 49 ; AIC = 8108.374
```

```
fit <- coxph(Surv(time,status)~age,data=colonMod)
fit.test <- coxph(Surv(time,status)~age*I(age>49),data=colonMod)
anova(fit,fit.test)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age
##  Model 2: ~ age * I(age > 49)
##    loglik  Chisq Df P(>|Chi|)
## 1 -4053.7
## 2 -4051.2 4.9627  2   0.08363 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

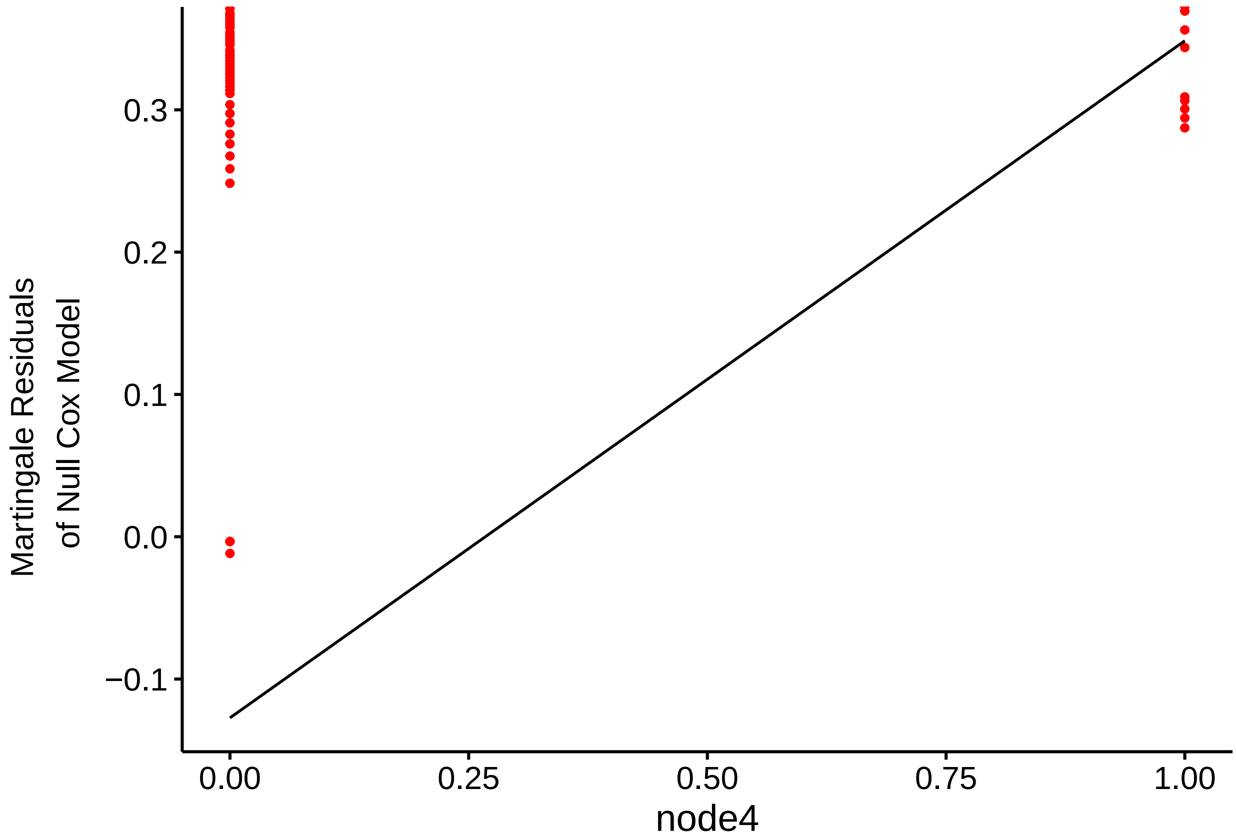
```
fit.test <- coxph(Surv(time,status)~age*I(age>49),data=colonMod)
AIC(fit,fit.test)
```

```
##          df      AIC
## fit       1 8109.336
## fit.test  3 8108.374
```

node4

```
ggcoxfunctional(Surv(time, status) ~ node4, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```

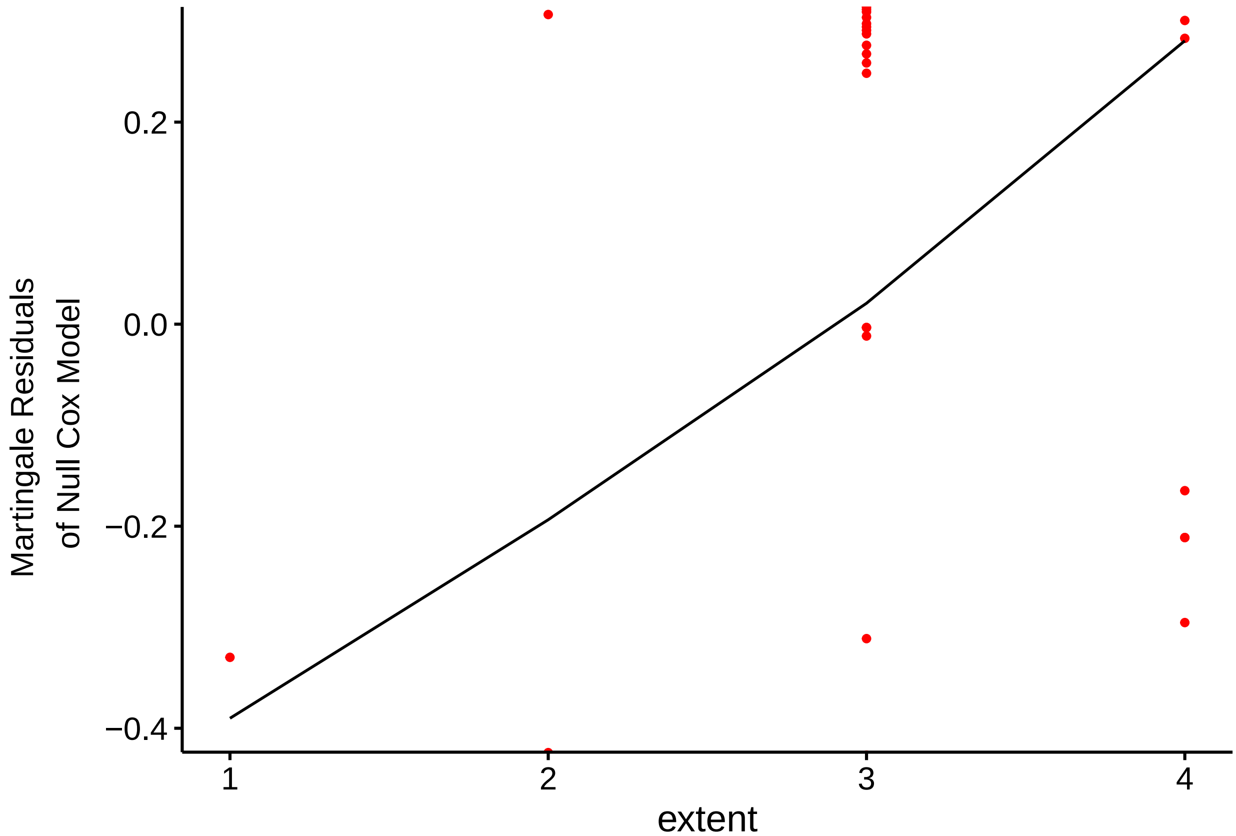


Looks linear.

extent

```
ggcoxfunctional(Surv(time, status) ~ extent, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```

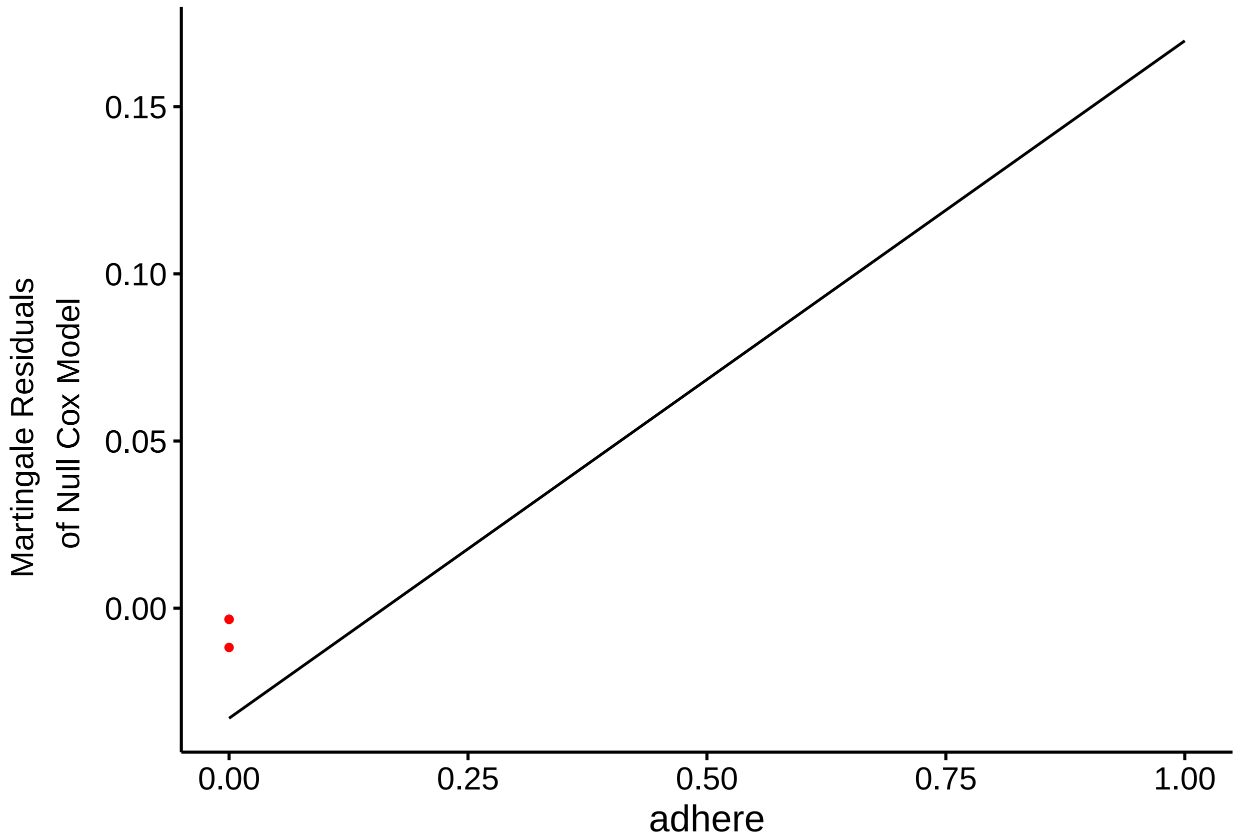


This looks linear enough. Piecewise is possible, but the sample size for extent less than 3 is very small and this is probably not worth investigating.

adhere

```
ggcoxfunctional(Surv(time, status) ~ adhere, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```

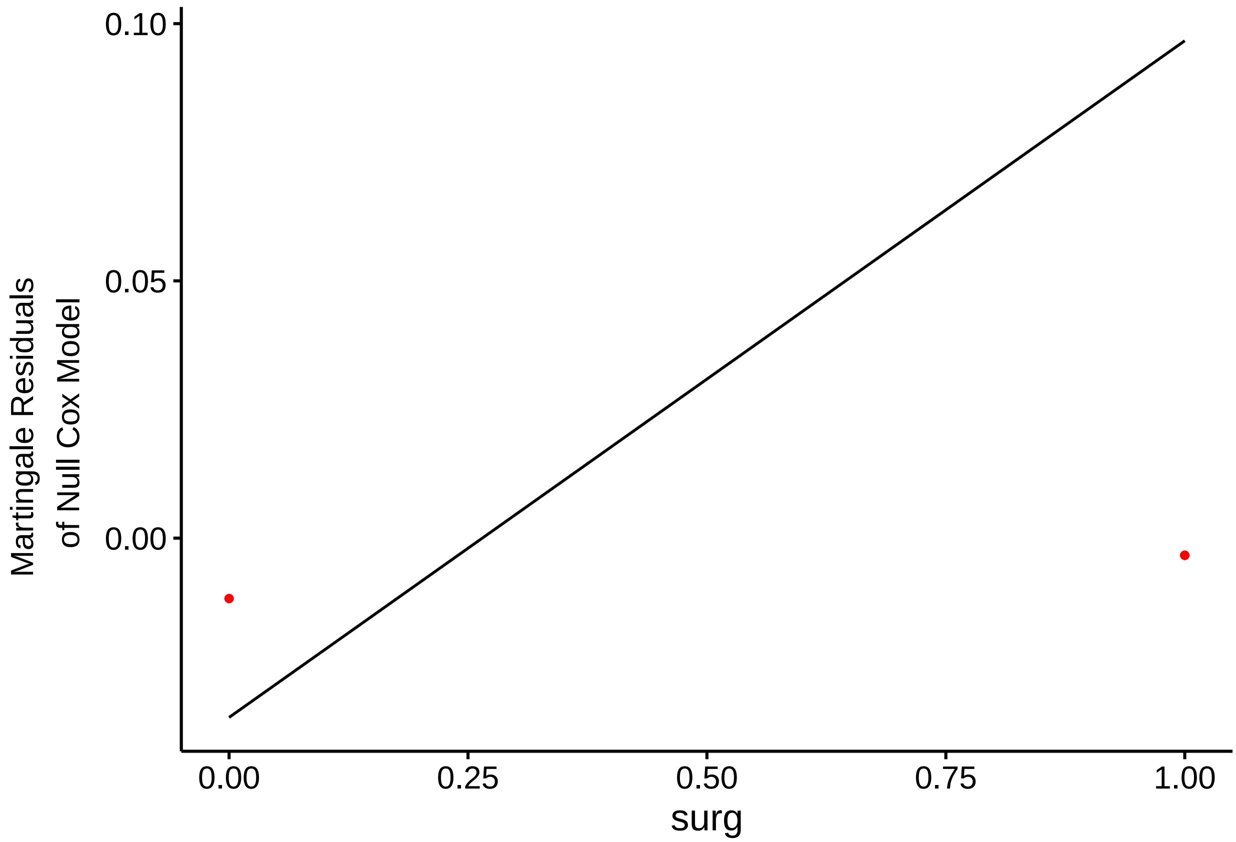


Looks linear.

surg

```
ggcoxfunctional(Surv(time, status) ~ surg, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```

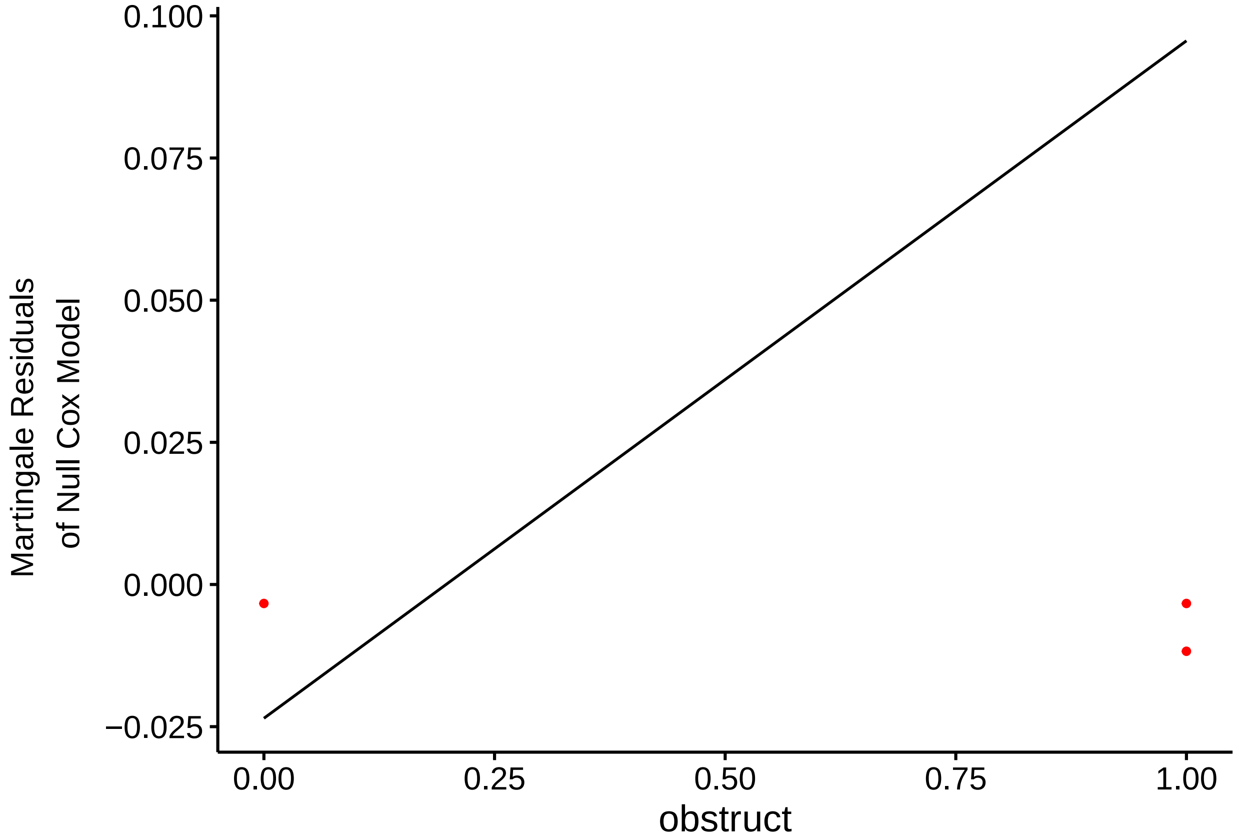


Looks linear.

obstruct

```
ggcoxfunctional(Surv(time, status) ~ obstruct, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```

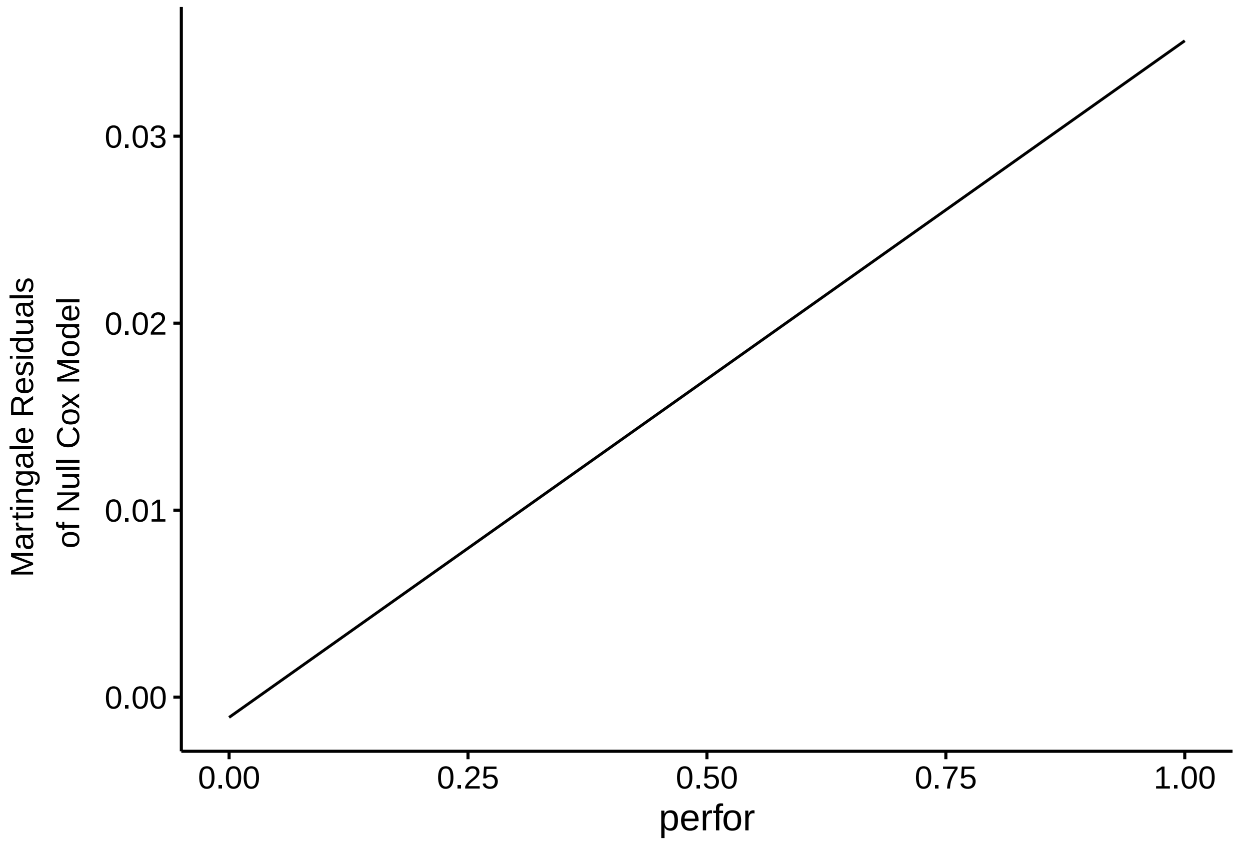


Looks linear.

perfor

```
ggcoxfunctional(Surv(time, status) ~ perfor, data = colonMod)
```

```
## Warning: arguments formula is deprecated; will be removed in the next version;
## please use fit instead.
```



Looks linear.

The Model Thus Far

```
fit <- coxph(Surv(time,status)~age+sex+rx+node4+extent+obstruct+surg+adhere+perfor,data=colonMod)
```

Removing Predictors, Testing Assumptions

It isn't necessarily the case that including all significant univariate predictors will produce a good model. This is just one way to go about it: include all of them, and then start removing ones that fail assumptions and aren't really improving the predictive power of the model. Here is the summary so far:

```
summary(fit)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex + rx + node4 +
##     extent + obstruct + surg + adhere + perfor, data = colonMod)
##
##   n= 1200, number of events= 601
##
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)
## age         0.0013604  1.0013613  0.0035111  0.387  0.69842
## sexmale    -0.0676497  0.9345878  0.0821189 -0.824  0.41005
```

```
## rxLev+5FU -0.4586667  0.6321259  0.1060001 -4.327 1.51e-05 ***
## rxObs     -0.0004284  0.9995717  0.0951504 -0.005  0.99641
## node4      0.9333459  2.5430036  0.0850518 10.974  < 2e-16 ***
## extent     0.4436738  1.5584220  0.1002857  4.424 9.68e-06 ***
## obstruct   0.1949277  1.2152232  0.1011518  1.927  0.05397 .
## surg       0.2407003  1.2721397  0.0898503  2.679  0.00739 **
## adhere     0.3114646  1.3654235  0.1037237  3.003  0.00267 **
## perfor    -0.0645637  0.9374764  0.2378956 -0.271  0.78609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0014     0.9986    0.9945    1.0083
## sexmale      0.9346     1.0700    0.7956    1.0978
## rxLev+5FU    0.6321     1.5820    0.5135    0.7781
## rxObs        0.9996     1.0004    0.8295    1.2045
## node4        2.5430     0.3932    2.1525    3.0043
## extent       1.5584     0.6417    1.2803    1.8969
## obstruct     1.2152     0.8229    0.9967    1.4817
## surg         1.2721     0.7861    1.0667    1.5171
## adhere       1.3654     0.7324    1.1142    1.6732
## perfor       0.9375     1.0667    0.5881    1.4944
##
## Concordance= 0.661  (se = 0.011 )
## Likelihood ratio test= 185.8  on 10 df,   p=<2e-16
## Wald test            = 190.7  on 10 df,   p=<2e-16
## Score (logrank) test = 200.8  on 10 df,   p=<2e-16
```

Let's first check model assumptions, that is, does the global model satisfy the Cox PH assumption, and do the individual predictors satisfy it as well?

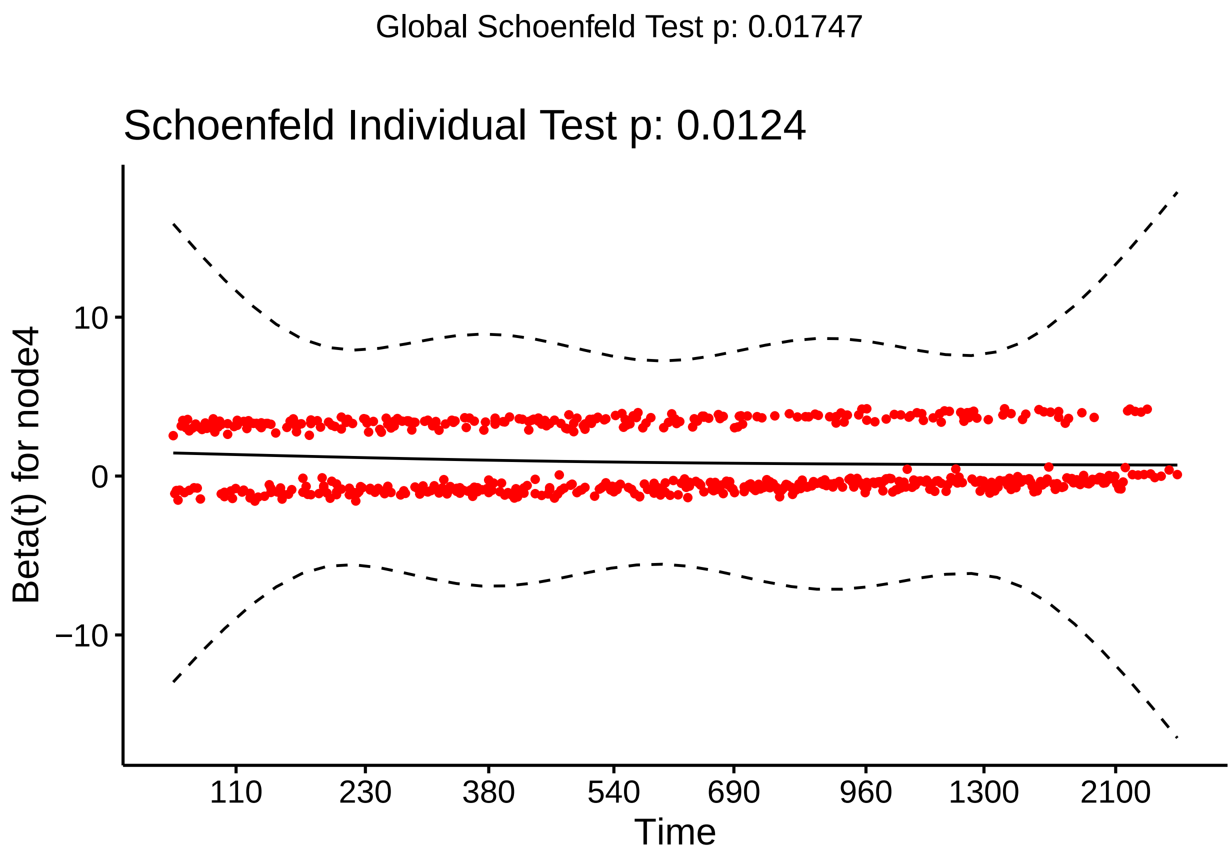
Testing the Cox PH Assumption

```
test.fit <- cox.zph(fit)
test.fit
```

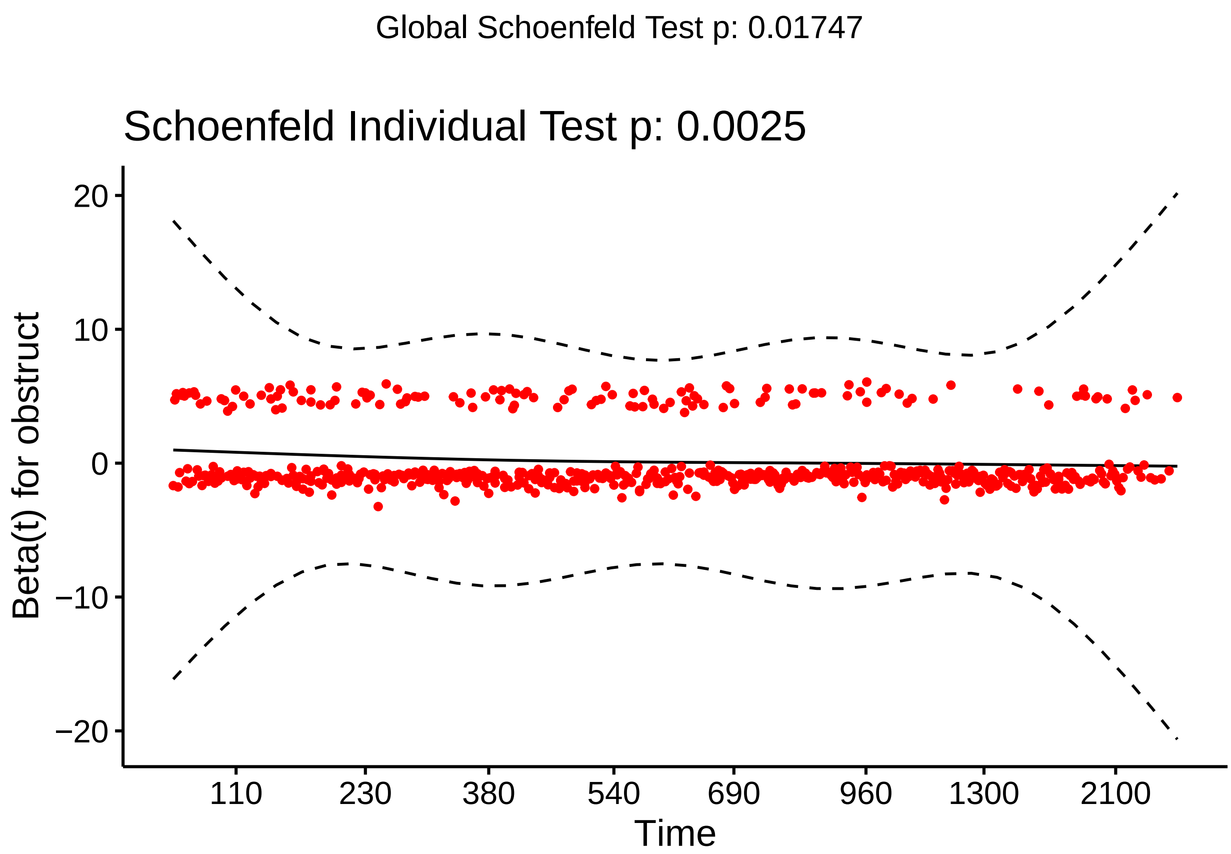
```
##           chisq df      p
## age       0.136  1 0.7119
## sex       1.723  1 0.1893
## rx        0.408  2 0.8155
## node4     6.255  1 0.0124
## extent    1.148  1 0.2839
## obstruct  9.111  1 0.0025
## surg      0.244  1 0.6211
## adhere    3.066  1 0.0800
## perfor    0.017  1 0.8963
## GLOBAL   21.567 10 0.0175
```

Here, the null hypothesis is that the Cox PH assumption is satisfied. So, we don't have evidence to say that the global model fails the PH assumption, but we do see that node4, obstruct probably do. Let's look a bit more in detail to see that:

```r
ggcoxzph(test.fit,var="node4")
```




```r
ggcoxzph(test.fit,var="obstruct")
```



You can detect a slight fluctuation in these with respect to time (it is very minimal, but it should be completely flat). So, it safe to remove these and forget about them. Here is the new model:

```
fit <- coxph(Surv(time,status)~age+sex+rx+extent+surg+adhere+perfor,data=colonMod)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex + rx + extent +
##     surg + adhere + perfor, data = colonMod)
##
##   n= 1200, number of events= 601
##
##                  coef exp(coef)  se(coef)      z Pr(>|z|)
## age         -0.002787  0.997217  0.003436 -0.811  0.41727
## sexmale     -0.082913  0.920431  0.081956 -1.012  0.31169
## rxLev+5FU -0.442736  0.642277  0.105898 -4.181 2.91e-05 ***
## rxObs        0.015521  1.015642  0.094831  0.164  0.86999
## extent       0.453294  1.573487  0.095459  4.749 2.05e-06 ***
## surg         0.254454  1.289757  0.089458  2.844  0.00445 **
## adhere       0.293550  1.341180  0.104630  2.806  0.00502 **
## perfor      -0.106609  0.898877  0.237345 -0.449  0.65331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          0.9972     1.0028    0.9905    1.0040
```

```
## sexmale      0.9204     1.0864    0.7838    1.0808
## rxLev+5FU    0.6423     1.5570    0.5219    0.7904
## rxObs        1.0156     0.9846    0.8434    1.2231
## extent       1.5735     0.6355    1.3050    1.8972
## surg         1.2898     0.7753    1.0823    1.5369
## adhere       1.3412     0.7456    1.0925    1.6465
## perfor       0.8989     1.1125    0.5645    1.4313
##
## Concordance= 0.597  (se = 0.011 )
## Likelihood ratio test= 73.1  on 8 df,   p=1e-12
## Wald test            = 70.34  on 8 df,   p=4e-12
## Score (logrank) test = 70.31  on 8 df,   p=4e-12
```

Removing Covariates

The model is suggesting that age, sex and perfor indicator may not be so useful when including rx, extent, surg and adhere. Let's do tests to see if we should drop them, and also use AIC in our analysis.

```
fit <- coxph(Surv(time,status)~age+sex+rx+extent+surg+adhere+perfor,data=colonMod)
fit2 <- coxph(Surv(time,status)~sex+rx+extent+surg+adhere+perfor,data=colonMod)
anova(fit,fit2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age + sex + rx + extent + surg + adhere + perfor
##  Model 2: ~ sex + rx + extent + surg + adhere + perfor
##    loglik  Chisq Df P(>|Chi|)
## 1 -4017.6
## 2 -4017.9 0.6541  1    0.4187
```

```
fit3 <- coxph(Surv(time,status)~age+rx+extent+surg+adhere+perfor,data=colonMod)
anova(fit,fit3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age + sex + rx + extent + surg + adhere + perfor
##  Model 2: ~ age + rx + extent + surg + adhere + perfor
##    loglik  Chisq Df P(>|Chi|)
## 1 -4017.6
## 2 -4018.1 1.0226  1    0.3119
```

```
fit4 <- coxph(Surv(time,status)~age+sex+rx+extent+surg+adhere,data=colonMod)
anova(fit,fit4)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age + sex + rx + extent + surg + adhere + perfor
##  Model 2: ~ age + sex + rx + extent + surg + adhere
##    loglik  Chisq Df P(>|Chi|)
## 1 -4017.6
## 2 -4017.7 0.2081  1    0.6482
```

```r
AIC(fit,fit2,fit3,fit4)
```

```
##      df      AIC
## fit   8 8051.183
## fit2  7 8049.837
## fit3  7 8050.205
## fit4  7 8049.391
```

Neither hypothesis test suggests that we add their term in the model (the first test is for adding age, the second is for adding sex and the third is for adding perfor). However, fit4 seems to do better on AIC. So, let's remove perfor first, and then see what happens with age and sex.

```r
fit <- coxph(Surv(time,status)~age+sex+rx+extent+surg+adhere,data=colonMod)
fit2 <- coxph(Surv(time,status)~sex+rx+extent+surg+adhere,data=colonMod)
anova(fit,fit2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age + sex + rx + extent + surg + adhere
##  Model 2: ~ sex + rx + extent + surg + adhere
##    loglik Chisq Df P(>|Chi|)
## 1 -4017.7
## 2 -4018.0 0.611  1    0.4344
```

```r
fit3 <- coxph(Surv(time,status)~age+rx+extent+surg+adhere,data=colonMod)
anova(fit,fit3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age + sex + rx + extent + surg + adhere
##  Model 2: ~ age + rx + extent + surg + adhere
##    loglik  Chisq Df P(>|Chi|)
## 1 -4017.7
## 2 -4018.2 1.0548  1    0.3044
```

```r
AIC(fit,fit2,fit3)
```

```
##      df      AIC
## fit   7 8049.391
## fit2  6 8048.002
## fit3  6 8048.446
```

```r
fit <- coxph(Surv(time,status)~sex+rx+extent+surg+adhere,data=colonMod)
fit3 <- coxph(Surv(time,status)~rx+extent+surg+adhere,data=colonMod)
anova(fit,fit3)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ sex + rx + extent + surg + adhere
##  Model 2: ~ rx + extent + surg + adhere
##    loglik  Chisq Df P(>|Chi|)
## 1 -4018.0
## 2 -4018.6 1.1158  1    0.2908
```

```r
AIC(fit,fit3)
```

```
##      df      AIC
## fit   6 8048.002
## fit3  5 8047.118
```

lower AIC indicates a better model (on that measure), so in terms of AIC, fit3 is better (no sex variable)

```r
test.fit <- cox.zph(fit3)
test.fit
```

```
##        chisq df     p
## rx     0.261  2 0.878
## extent 0.756  1 0.385
## surg   0.060  1 0.807
## adhere 2.749  1 0.097
## GLOBAL 4.364  5 0.498
```

Looks good. So, my personal choice is to go with fit3. Here is the model output:

```r
summary(fit3)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + extent + surg + adhere,
##     data = colonMod)
##
##   n= 1200, number of events= 601
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev+5FU -0.44038   0.64379  0.10552 -4.174 3.00e-05 ***
## rxObs      0.01936   1.01955  0.09471  0.204  0.83803
## extent     0.45294   1.57293  0.09542  4.747 2.07e-06 ***
## surg       0.25143   1.28586  0.08938  2.813  0.00491 **
## adhere     0.28062   1.32395  0.10293  2.726  0.00640 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev+5FU    0.6438     1.5533    0.5235    0.7917
## rxObs        1.0195     0.9808    0.8468    1.2275
## extent       1.5729     0.6358    1.3046    1.8964
## surg         1.2859     0.7777    1.0792    1.5320
## adhere       1.3240     0.7553    1.0821    1.6199
##
## Concordance= 0.594  (se = 0.011 )
## Likelihood ratio test= 71.16  on 5 df,   p=6e-14
## Wald test            = 68.19  on 5 df,   p=2e-13
## Score (logrank) test = 68.02  on 5 df,   p=3e-13
```

Final Notes

A few comments:

- other measures of model performance should probably also be considered, especially considering how close our final two models were.
- we didn't consider interaction terms. If any of the relationships between these covariates and survival time are hypothesized to change depending on values of other covariates, this should be tested.