veteran dataset

We will still use `veteran` data set in `survival` package from Veterans' Administration Lung Cancer study.

```
head(veteran)
```

```
##   trt celltype time status karno diagtime age prior
## 1   1 squamous   72      1    60        7  69     0
## 2   1 squamous  411      1    70        5  64    10
## 3   1 squamous  228      1    60        3  38     0
## 4   1 squamous  126      1    60        9  63    10
## 5   1 squamous  118      1    70       11  65    10
## 6   1 squamous   10      1    20        5  49     0
```

- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status

See R: Veterans' Administration Lung Cancer study for more details.

Review: Cox PH Model

We want to see if there is a significant effect from cancer cell types.

```
fit = coxph(Surv(time,status)~celltype,data=veteran)
summary(fit)
```

Likelihood ratio test:

p-value = 2e-05. (it has significant effect)

Each individual test: e.g.

p-value = 7.83e-05. (`celltypesmallcell` is significantly different from `celltysquamous`)

Review: Cox PH Model

```
Call:
coxph(formula = Surv(time, status) ~ celltype, data = veteran)

  n= 137, number of events= 128

                    coef exp(coef) se(coef)     z Pr(>|z|)
celltypesmallcell 1.0013    2.7217   0.2535 3.950 7.83e-05 ***
celltypeadeno     1.1477    3.1510   0.2929 3.919 8.90e-05 ***
celltypelarge     0.2301    1.2588   0.2773 0.830    0.407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
celltypesmallcell     2.722     0.3674     1.656     4.473
celltypeadeno         3.151     0.3174     1.775     5.594
celltypelarge         1.259     0.7944     0.731     2.168

Concordance= 0.608  (se = 0.029 )
Rsquare= 0.166   (max possible= 0.999 )
Likelihood ratio test= 24.85  on 3 df,   p=2e-05
Wald test            = 24.09  on 3 df,   p=2e-05
Score (logrank) test = 25.51  on 3 df,   p=1e-05
```

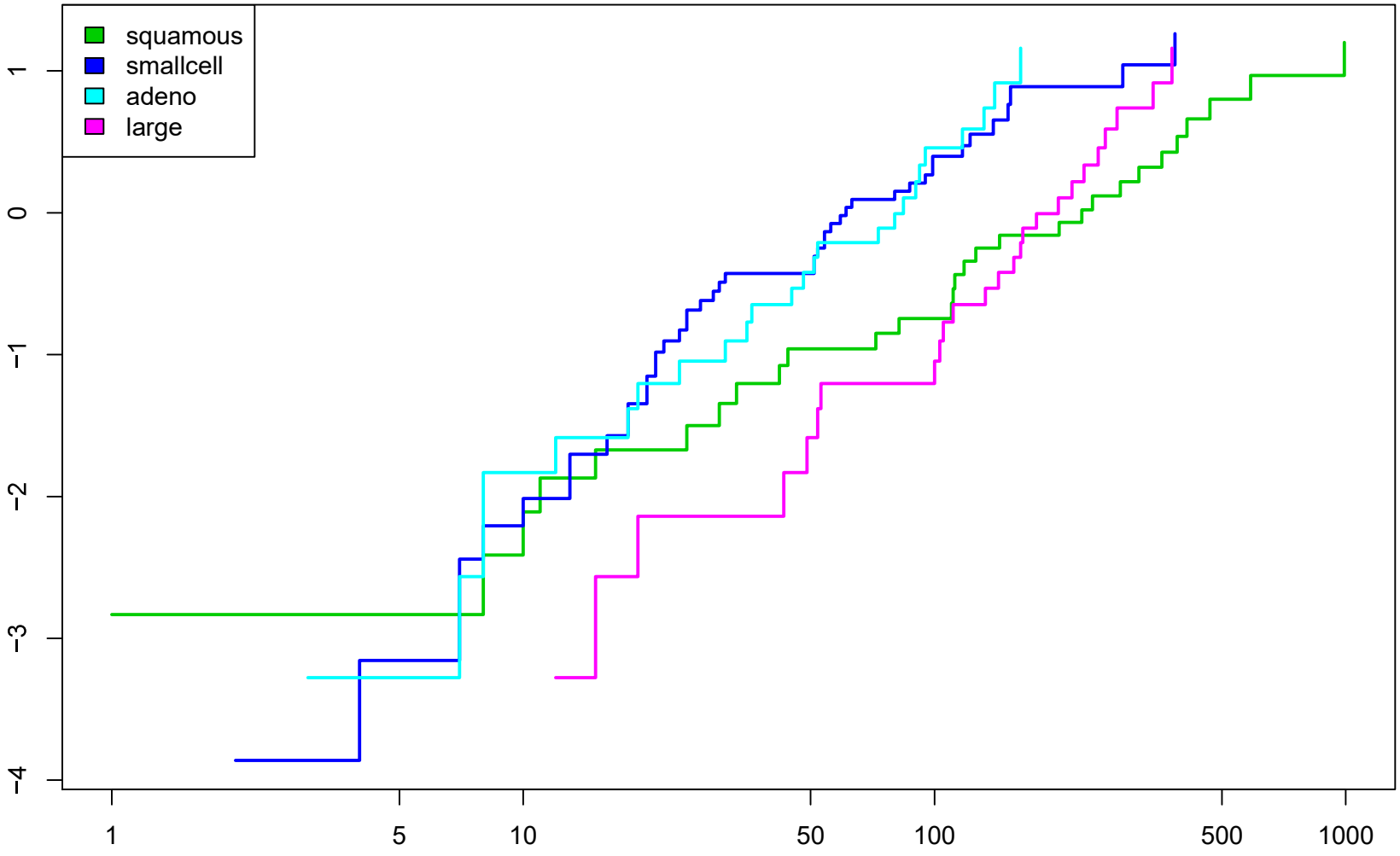
Figure 1: summary(fit)

Review: Log-Log Plot

We can use the following codes to draw the log-log plot:

```
veteran.km = survfit(Surv(time,status)~celltype,
                     data=veteran)
plot(veteran.km, fun='cloglog', col=3:6,lwd=2)
#levels(veteran$celltype)
legend('topleft',c("squamous","smallcell",
                   "adeno","large"),fill = 3:6)
```

Review: Log-Log Plot



Review: Cox PH Model (with covariates)

We want to test the effect of `celltype`, controlling the `diagtime` covariate.

```
fit2 = coxph(Surv(time,status)~celltype+diagtime,
             data=veteran)
fit3 = coxph(Surv(time,status)~diagtime,data=veteran)
#summary(fit2)
#summary(fit3)
```

Review: Cox PH Model (with covariates)

```
Call:
coxph(formula = Surv(time, status) ~ celltype + diagtime, data = veteran)

  n= 137, number of events= 128

                    coef exp(coef) se(coef)     z Pr(>|z|)
celltypesmallcell 0.982017  2.669835 0.254398 3.860 0.000113 ***
celltypeadeno     1.180827  3.257068 0.294902 4.004 6.22e-05 ***
celltypelarge     0.234520  1.264302 0.277552 0.845 0.398133
diagtime          0.009137  1.009179 0.008539 1.070 0.284562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  exp(coef) exp(-coef) lower .95 upper .95
celltypesmallcell     2.670     0.3746    1.6216     4.396
celltypeadeno         3.257     0.3070    1.8273     5.806
celltypelarge         1.264     0.7910    0.7338     2.178
diagtime              1.009     0.9909    0.9924     1.026

Concordance= 0.622  (se = 0.03 )
Rsquare= 0.172   (max possible= 0.999 )
Likelihood ratio test= 25.86  on 4 df,   p=3e-05
Wald test            = 25.38  on 4 df,   p=4e-05
Score (logrank) test = 26.86  on 4 df,   p=2e-05
```

Figure 2: summary(fit2)

Review: Cox PH Model (with covariates)

```
Call:
coxph(formula = Surv(time, status) ~ diagtime, data = veteran)

  n= 137, number of events= 128

             coef exp(coef) se(coef)     z Pr(>|z|)
diagtime 0.009100  1.009142 0.008978 1.014    0.311

         exp(coef) exp(-coef) lower .95 upper .95
diagtime     1.009     0.9909    0.9915     1.027

Concordance= 0.509  (se = 0.03 )
Rsquare= 0.007   (max possible= 0.999 )
Likelihood ratio test= 0.91  on 1 df,   p=0.3
Wald test            = 1.03  on 1 df,   p=0.3
Score (logrank) test = 1.02  on 1 df,   p=0.3
```

Figure 3: summary(fit3)

Review: Cox PH Model (with covariates)

Degree of Freedom is:

$$\text{df} = 4 - 1 = 3.$$

Two ways to compute the likelihood ratio statistic:

```
#lrt2 = 2*(fit2$loglik[2]-fit3$loglik[2])
#pchisq(lrt2, df=3, lower.tail = FALSE)
lrt1 = summary(fit2)$logtest[1] - summary(fit3)$logtest[1]
pchisq(lrt1, df=3, lower.tail = FALSE)
```

```
##          test
## 1.583109e-05
```

p-value is small $\Longrightarrow$ the effect of `celltype` is significant if we consider `diagtime`

Review: Cox PH Model (PH assumption)

`cox.zph()` is used to test the Proportional Hazards Assumption of a Cox Regression.

```
cox.zph(fit2)
```

```
##                         rho   chisq      p
## celltypesmallcell 0.05683 0.43383 0.5101
## celltypeadeno     0.14724 2.93832 0.0865
## celltypelarge     0.20260 5.32714 0.0210
## diagtime          0.00401 0.00221 0.9625
## GLOBAL                 NA 7.08153 0.1316
```

p-value is small $(0.0210 < 0.05)$ $\Longrightarrow$ **the PH assumption is violated.**

Review: Stratified Cox PH Model

According to our analysis, the `celltype` may rely on time t (or the baseline may rely on `celltype`). For different `celltype`, we use different baseline. (the parameters of `diagtime` are same.)

```
fitSC = coxph(Surv(time,status)~diagtime+strata(celltype),
              data=veteran)
summary(fitSC)
```

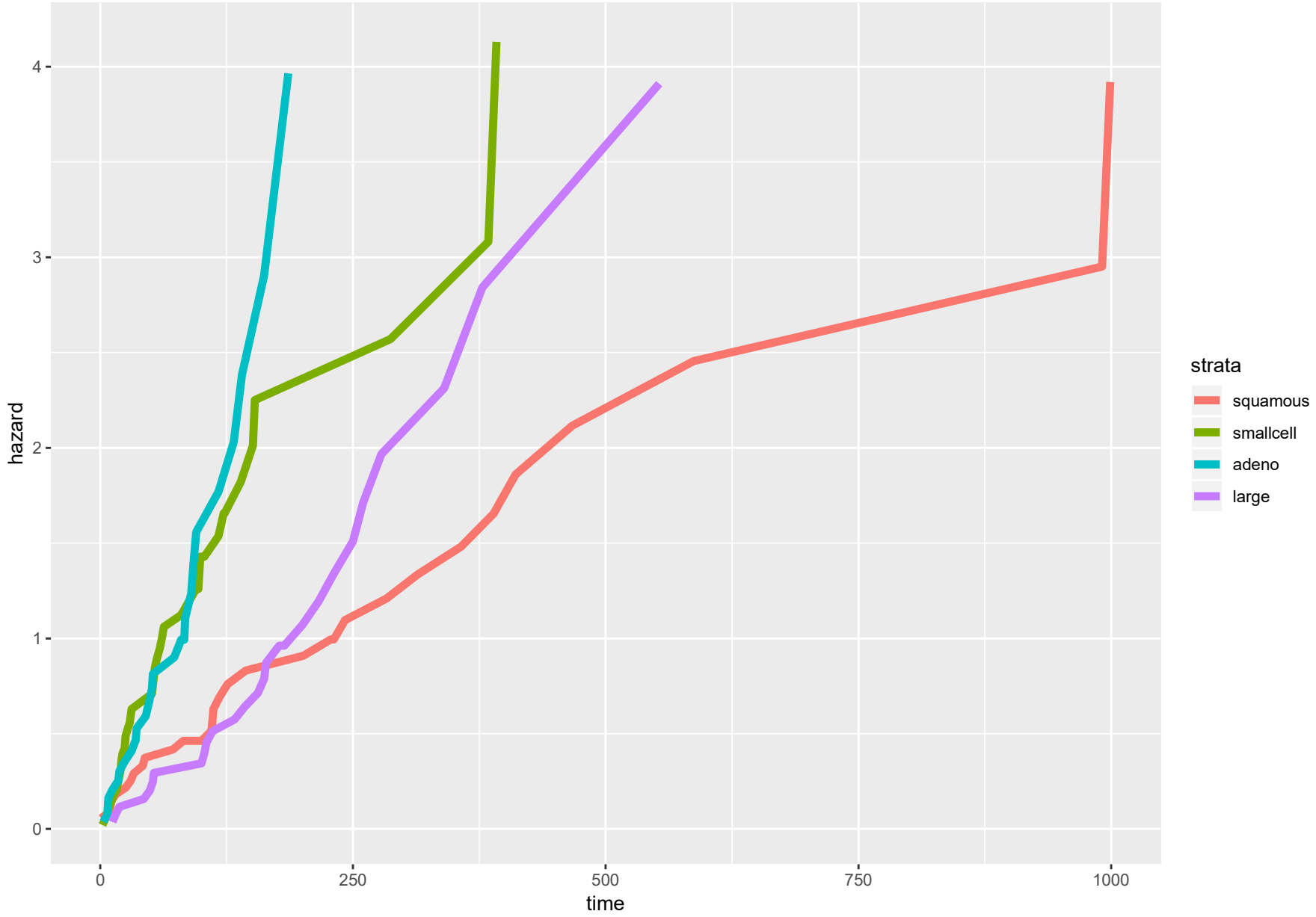
Review: Stratified Cox PH Model

```
Call:
coxph(formula = Surv(time, status) ~ diagtime + strata(celltype),
    data = veteran)

  n= 137, number of events= 128

             coef exp(coef) se(coef)     z Pr(>|z|)
diagtime 0.009883  1.009932 0.008323 1.187    0.235

         exp(coef) exp(-coef) lower .95 upper .95
diagtime      1.01     0.9902    0.9936     1.027

Concordance= 0.533  (se = 0.059 )
Rsquare= 0.009   (max possible= 0.993 )
Likelihood ratio test= 1.23  on 1 df,   p=0.3
Wald test            = 1.41  on 1 df,   p=0.2
Score (logrank) test = 1.42  on 1 df,   p=0.2
```

Figure 4: summary(fitSC)

Review: Stratified Cox PH Model

We can plot their baseline hazard function using `basehaz()` function:



Review: Stratified Cox PH Model

Codes:

```
bhaz = basehaz(fitSC)
ggplot(bhaz)+
   geom_line(aes(x=time,y=hazard,colour=strata), size=2)
```

Review: Stratified Cox PH Model

Finally, we want to see if there is a significant interaction between diagtime and celltype.

```
fitSC = coxph(Surv(time,status)~diagtime+strata(celltype),
#Fit the stratified model with interaction :
fitSC.int = coxph(Surv(time,status)~
                    strata(celltype)*diagtime, data=veteran
#Compute GLRT statistic:
lrt = 2*(fitSC.int$loglik[2]-fitSC$loglik[2])
#p-value:
pchisq(lrt,df=3,lower.tail = FALSE)
```

```
## [1] 0.1739869
```

p-value is large $\implies$ the interaction term is not significant

survSplit() function

```
help("survSplit")
```

Description

Given a survival data set and a set of specified cut times, split each record into multiple subrecords at each cut time. The new data set will be in 'counting process' format, with a start time, stop time, and event status for each record.

Usage

survSplit(formula, data, subset)

survSplit() function

Now we want to construct a new data frame with additional rows that split the time variable into before and after $t = 90$.

```
veteran2 <- survSplit(Surv(time, status) ~1, veteran,
                      cut=90, episode ="timegroup")

#veteran2 <- survSplit(Surv(time, status) ~1, veteran,
#                      cut=c(90,120), episode ="timegroup",
```

survSplit() function

New data frame has an additional colomn tstart.

```
names(veteran)
```

```
## [1] "trt"      "celltype" "time"     "status"   "karno"
## [7] "age"      "prior"
```

```
names(veteran2)
```

```
## [1] "tstart"    "time"      "status"    "timegroup"
```

survSplit() function

(first row: time$=72$) When time≤ 90, we do nothing.

(second row: time$=411$) When time>90, divide this observation into two rows.

```
head(veteran,3)
```

```
##   veteran.time veteran.status
## 1           72              1
## 2          411              1
## 3          228              1
```

```
head(veteran2,4)
```

```
##   tstart time status timegroup
## 1      0   72      1         1
## 2      0   90      0         1
## 3     90  411      1         2
## 4      0   90      0         1
```

survSplit() function

```
nrow(veteran)
```

```
## [1] 137
```

```
nrow(veteran2)
```

```
## [1] 198
```

survSplit() function

Fit Cox PH model using the new data frame.

```
fit4 = coxph(Surv(tstart, time, status)~celltype:strata(tim
cox.zph(fit4)
```

	rho	chisq	p
diagtime	0.012056	0.018335	0.892
celltypesquamous:strata(timegroup)timegroup=1	-0.033792	0.144453	0.704
celltypesmallcell:strata(timegroup)timegroup=1	0.010383	0.013971	0.906
celltypeadeno:strata(timegroup)timegroup=1	0.021831	0.061545	0.804
celltypelarge:strata(timegroup)timegroup=1	NA	NaN	NaN
celltypesquamous:strata(timegroup)timegroup=2	-0.036766	0.171661	0.679
celltypesmallcell:strata(timegroup)timegroup=2	-0.046462	0.268001	0.605
celltypeadeno:strata(timegroup)timegroup=2	-0.000942	0.000109	0.992
celltypelarge:strata(timegroup)timegroup=2	NA	NaN	NaN
GLOBAL	NA	1.045402	0.999

Figure 5: cox.zph(fit4)