Today

- `quantile()` function
- Confident Interval
  - Use `survfit()` function
  - Greenwood formula
- `survfit()` for different groups
- Lab B: 3(d). Hypothesis test

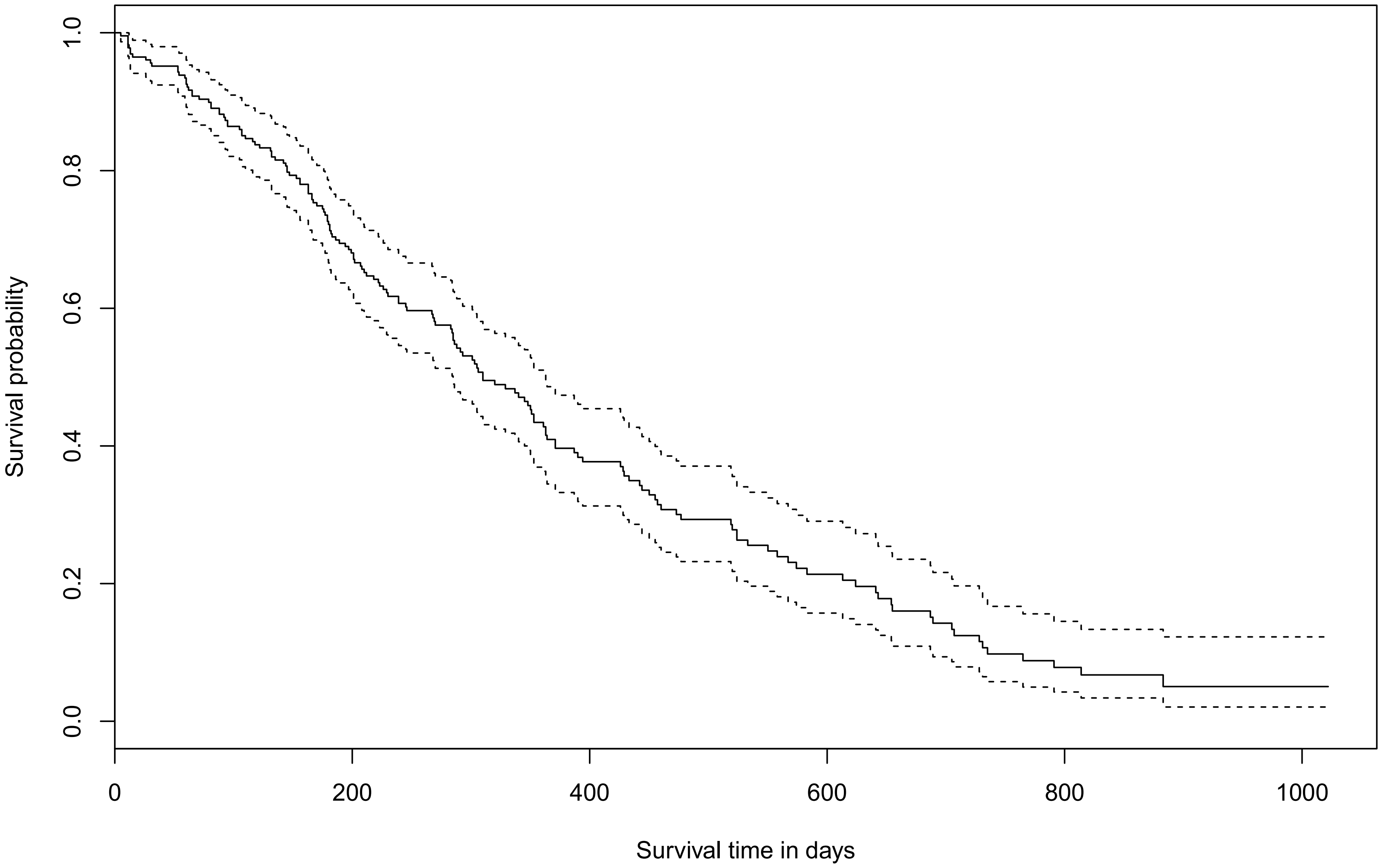
```r
heroin = read.table("Heroin.txt")
heroin.time = heroin$Time
heroin.cns = heroin$Status
heroin.surv <- Surv(heroin.time, heroin.cns)
heroin$Group <- ifelse(heroin$Time <= 365, 1, 0)
#factor(heroin$Group, levels=c(1,2))
#survdiff(heroin.surv ~ heroin$Group, rho=0)
#km = survfit(heroin.surv ~ heroin$Group)

sum(heroin$Group)
```

Review: `survfit()` function

Create a survival object and plot KM estimator with 95% CI.

```r
lungkm = survfit(Surv(time,status)~1,data=lung)
plot(lungkm,xlab="Survival time in days",
     ylab="Survival probability" )
```

Review: `survfit()` function



`quantile()` function

Compute the 10th quantile (the first time the survival function is below .9):

```r
min(lungkm$time[lungkm$surv < .9])
```

```
## [1] 79
```

quantile() function

Use quantile() funtion in R

```r
quantile(lungkm,probs = .1, conf.int = FALSE)
```

```
## 10
## 79
```

quantile() function

Use quantile() funtion in R

```
quantile(lungkm,probs =c(.1,.2,.75), conf.int = FALSE)
```

```
##  10  20  75
##  79 145 550
```

Confident Interval

Confident interval for estimated survival probability:

```r
summary(lungkm)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = l
##
##  time n.risk n.event survival std.err lower 95% CI upper
##     5    228       1   0.9956 0.00438       0.9871
##    11    227       3   0.9825 0.00869       0.9656
##    12    224       1   0.9781 0.00970       0.9592
##    13    223       2   0.9693 0.01142       0.9472
##    15    221       1   0.9649 0.01219       0.9413
##    26    220       1   0.9605 0.01290       0.9356
##    30    219       1   0.9561 0.01356       0.9299
##    31    218       1   0.9518 0.01419       0.9243
##    53    217       2   0.9430 0.01536       0.9134
##    54    215       1   0.9386 0.01590       0.9079
##    59    214       1   0.9342 0.01642       0.9026
```

Confident Interval

```
s = summary(lungkm)
names(s)
```

```
##  [1] "n"               "time"            "n.risk"          "n
##  [5] "n.censor"        "surv"            "type"            "st
##  [9] "lower"           "upper"           "conf.type"       "co
## [13] "call"            "table"           "rmean.endtime"
```

```
#s$lower
#s$upper
```

Greenwood formula

95% CI for $\log S(t)$:

$$\log \hat{S}(t) \pm 1.96 \hat{S}(t) \sqrt{\sum_{j=1}^{k} \frac{m_j}{n_j(n_j - m_j)}}$$

95% CI for $S(t)$:

$$\hat{S}(t) \times \exp\left[\pm 1.96 \hat{S}(t) \sqrt{\sum_{j=1}^{k} \frac{m_j}{n_j(n_j - m_j)}} \right]$$

Compute it by hand

```
mj = lungkm$n.event
nj = lungkm$n.risk
```

```
Vj = mj/nj/(nj-mj)
cVj = cumsum(Vj)
```

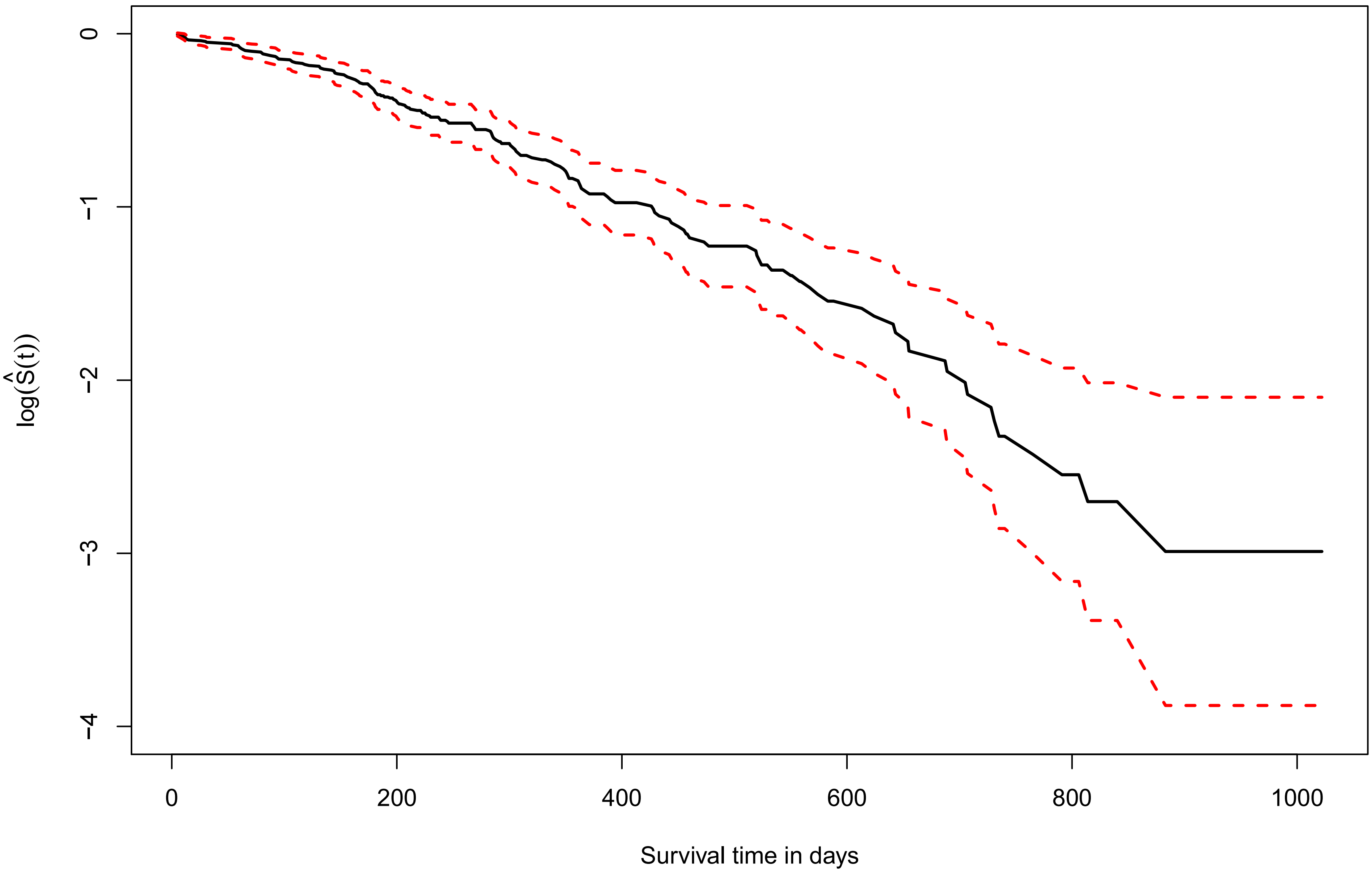
Compute it by hand

Reminder: 95% CI for $\log S(t)$:

$$\log \hat{S}(t) \pm 1.96 \hat{S}(t) \sqrt{\sum_{j=1}^{k} \frac{m_j}{n_j(n_j - m_j)}}$$

```
lowerCI = log(lungkm$surv) - 1.96*sqrt(cVj)
upperCI = log(lungkm$surv) + 1.96*sqrt(cVj)
```

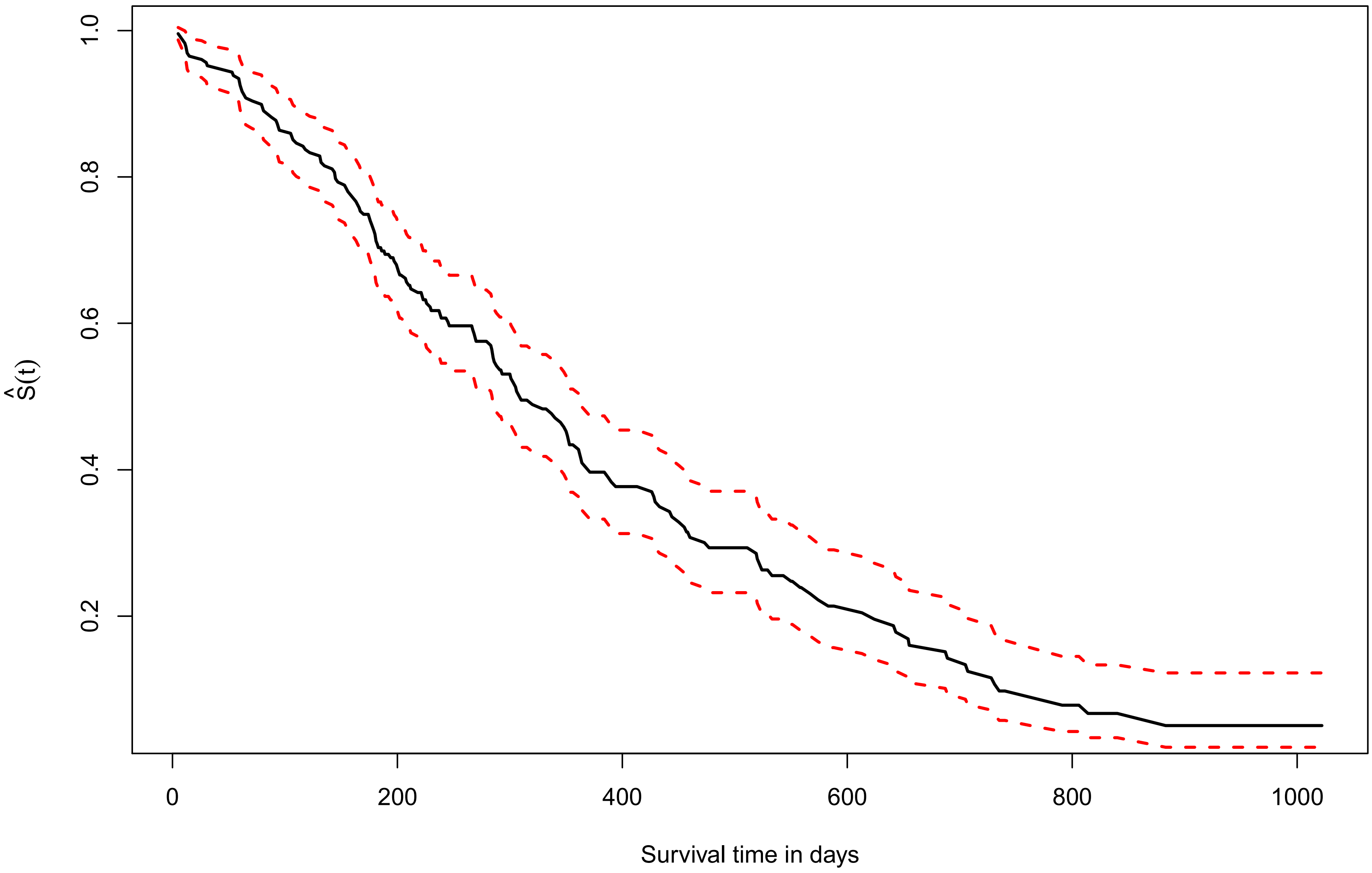
Plot it



Transform it into CI for S(t)

```
par(mar=c(5,5,4,2))
plot(lungkm$time,lungkm$surv,lwd=2,type="l",
xlab="Survival time in days",ylab=expression(hat(S)(t)))
lines(lungkm$time,exp(lowerCI),lty=2,col=2,lwd=2)
lines(lungkm$time,exp(upperCI),lty=2,col=2,lwd=2)
```

Transform it into CI for S(t)



survfit() for different groups

Divide data into to part. Treat them separetely

```
g1 = lung[lung$sex==1,]
g2 = lung[lung$sex==2,]
kmg1 = survfit(Surv(time,status)~1,data=g1)
kmg2 = survfit(Surv(time,status)~1,data=g2)
```

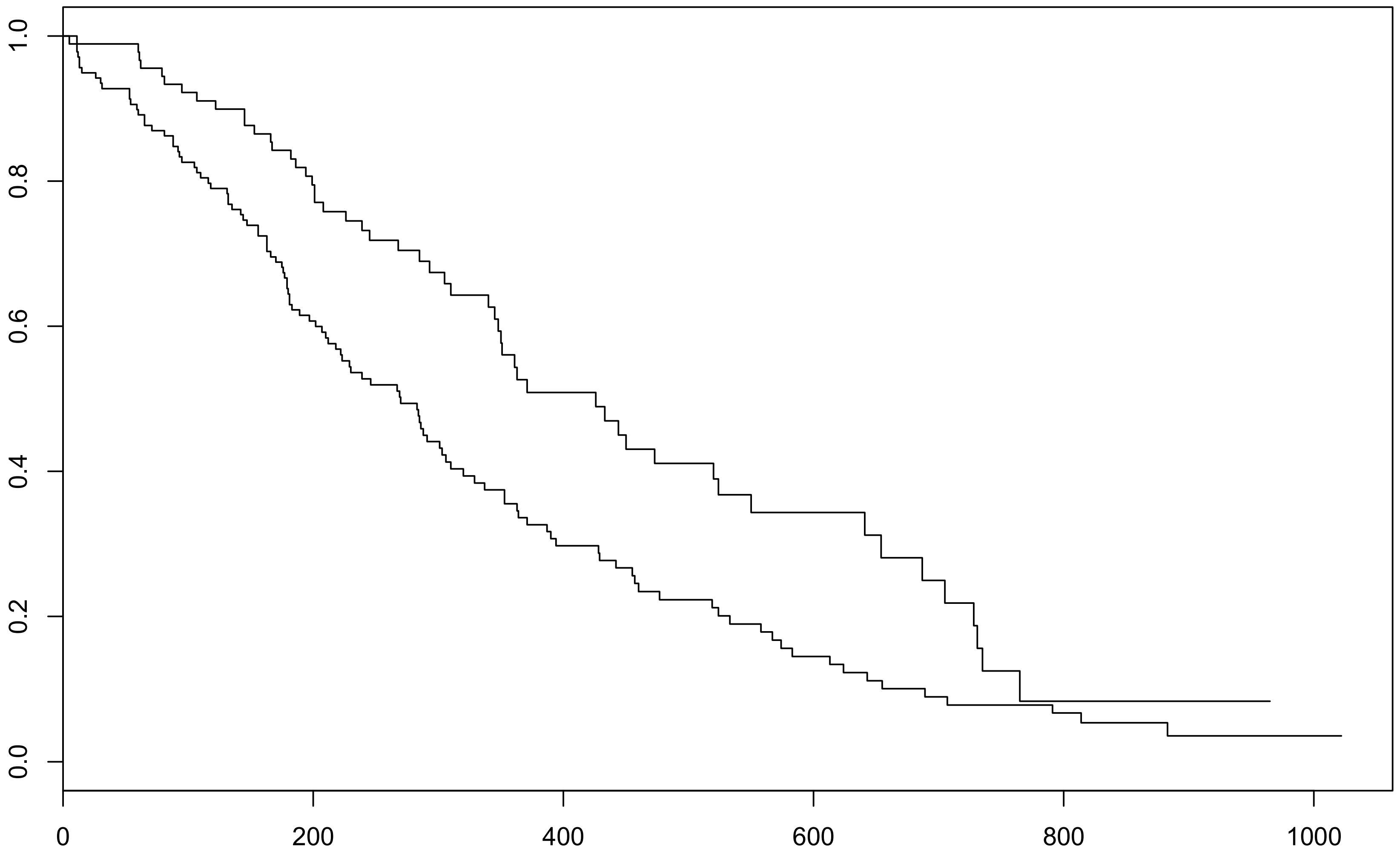
surfit() function

Change 1 to sex:

```
km = survfit(Surv(time,status)~sex,data=lung)
```

Plot

```
plot(km)
```



Last part: 3(d)

The file `heroin.Rdt` contains data from a study on in-patient methadone treatment clinics in Australia. The columns are labeled `Status` and `Time` which gives the number of days each subject spent in the clinics. Censored observations were generally subjects who were still in the clinics at the end of the study period.

(d)New recommendations for clinic administration are that, in order to save money, at least 50% of the patients should be discharged within one year. Is there significant evidence that most patients from this study population are in the clinics for more than one year? Perform a hypothesis test using the relevant statistic and an approximation to its standard error. Should we use a one-sided or a two-sided alternative? Compute an approximate P-value for the test.

veteran dataset

In today's section, we 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
```

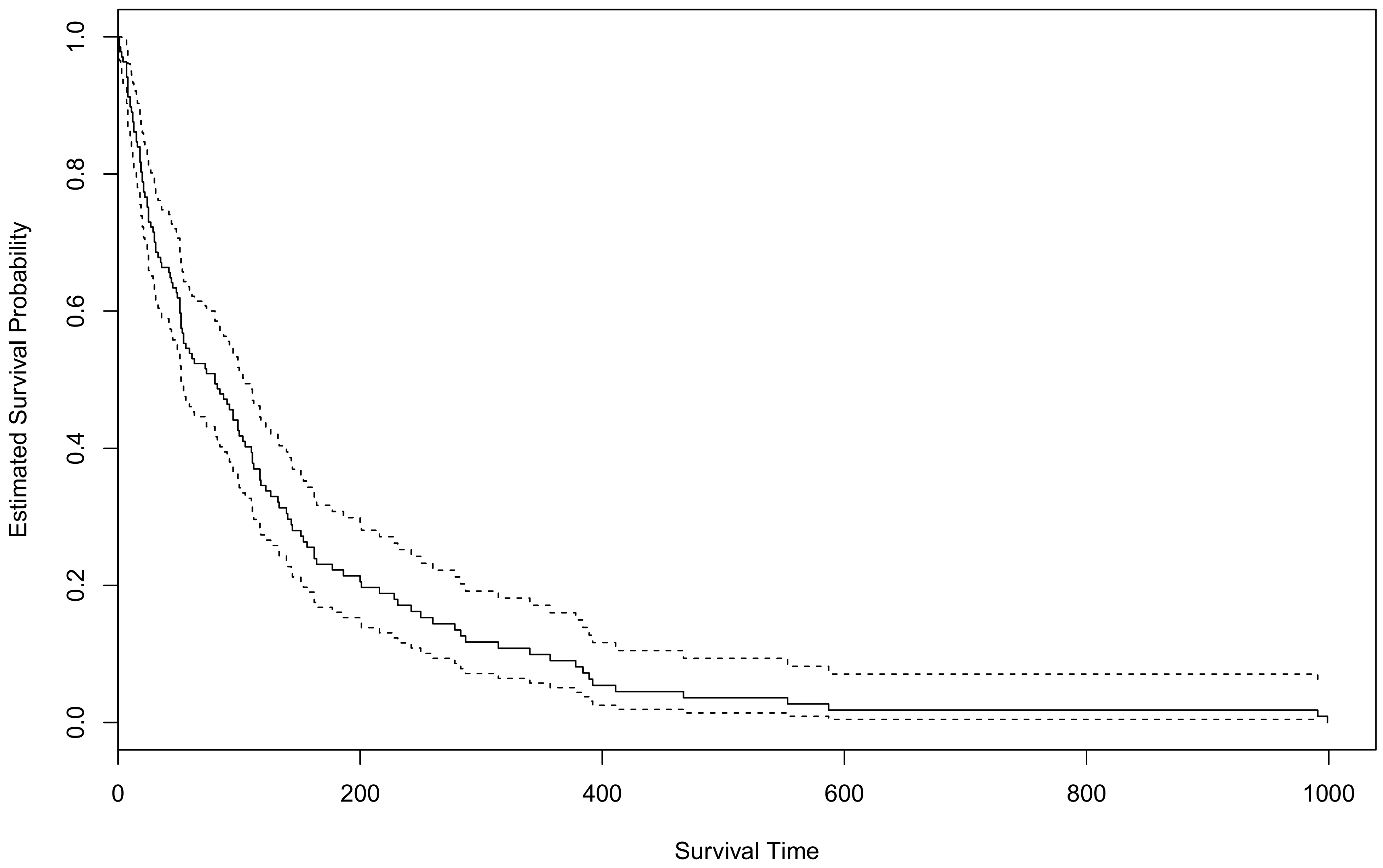
veteran dataset

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 10=yes

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

Review: K-M estimate

We only consider `time` and `status`. Plot the Kaplan–Meier estimate of the survivor function.



Review: K-M estimate

Code:

```
veteran.km = survfit(Surv(time,status)~1, data=veteran)
plot(veteran.km,xlab="Survival Time",
     ylab="Estimated Survival Probability" )
```

Review: Compare two groups

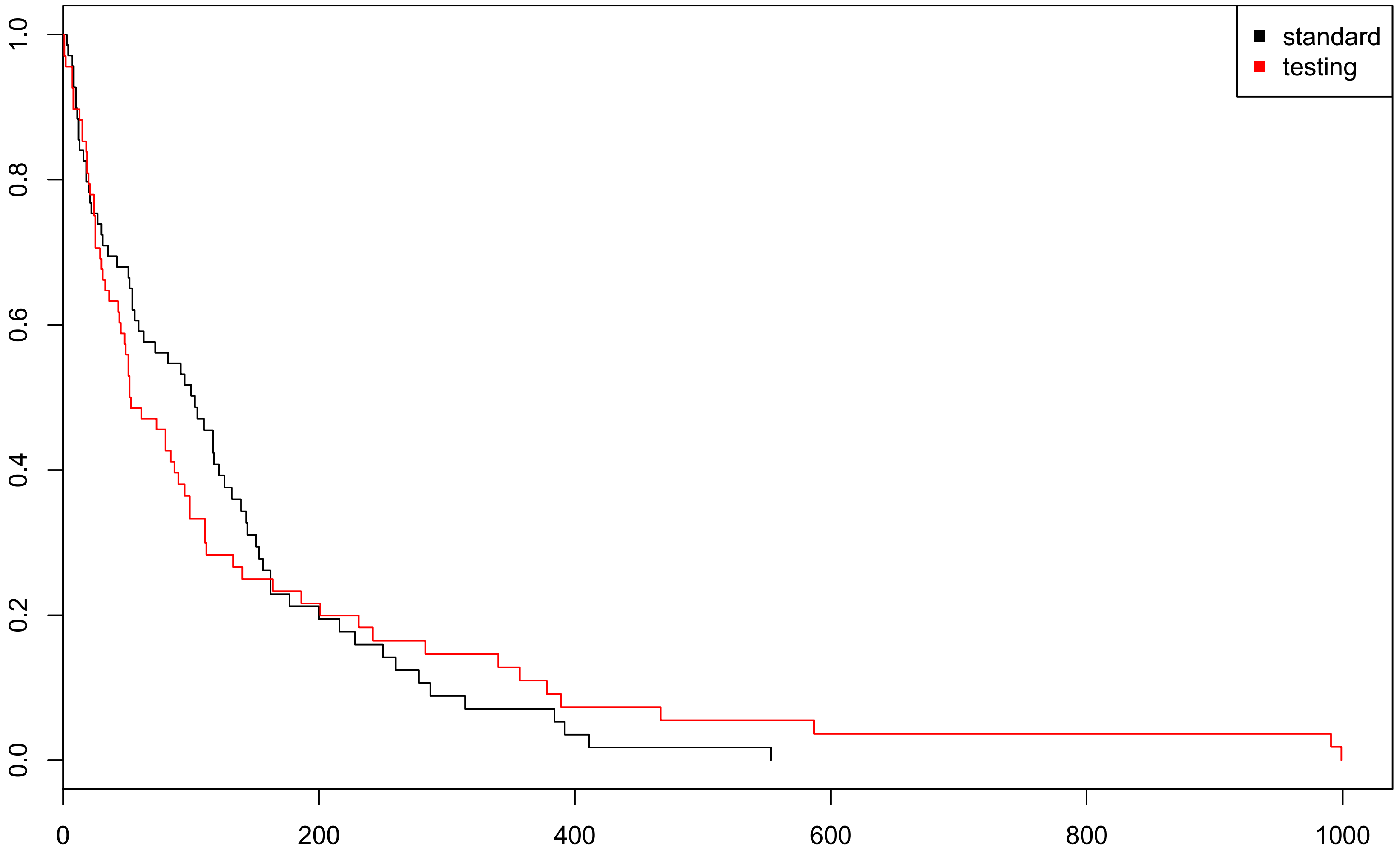
Goal: To study the effect of treatment. Divide all observations into two groups based on trt.

```
survdiff(Surv(time,status)~trt, data=veteran)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ trt, data = vete
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69       64     64.5   0.00388   0.00823
## trt=2 68       64     63.5   0.00394   0.00823
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.9
```

Review: Compare two groups

The p-value is 0.9. It means there is no significant difference between the standard group and the test group. Now, plot the K-M estimators for both groups:



Review: Compare two groups

- When $t > 200$, the estimated survival probability in testing group is always larger than that in standard group.
- Log-rank test claims that two groups are same.

coxph() function

Use coxph() function to compare two groups. Do not consider covariates.

```
coxph(Surv(time,status)~trt,data=veteran)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt, data = veteran
##
##        coef exp(coef) se(coef)   z    p
## trt 0.0177    1.0179   0.1807 0.1 0.92
##
## Likelihood ratio test=0.01  on 1 df, p=0.9
## n= 137, number of events= 128
```

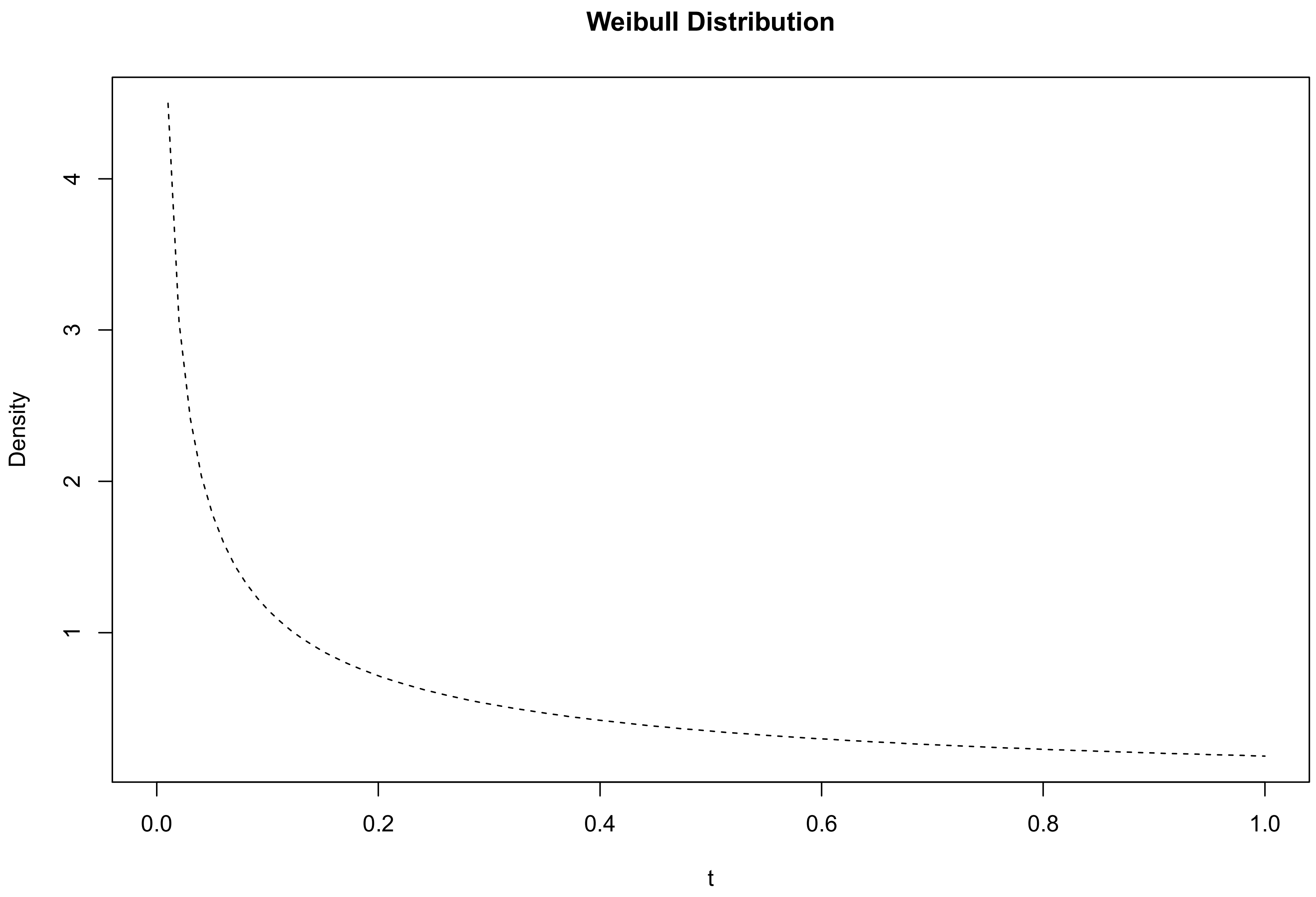
coxph() function

Use coxph() function to compare two groups. Consider covarites age, celltype, karno, and diagtime to control for differences in the groups.

```
#including `trt`
cox = coxph(Surv(time,status)~age+celltype+karno+diagtime
      +trt, data=veteran)
#not including `trt`
cox2 = coxph(Surv(time,status)~age+celltype+karno+diagtime
      ,data=veteran)
```

Cox proportional hazards model

Reminder: Hazard function can be considered as the risk of dying at time t. For example, $h(t)$ for leukemia patients has Weibull distribution.



Cox proportional hazards model

Semiparametric model for hazard function:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}.$$

- $h_0(t)$ is called *the baseline hazard function*.
- Proportional hazards assumption: $h_0(t)$ only relies on t.
- Time-independence.

Likelihood ratio test

```
cox$loglik
```

```
## [1] -505.4491 -474.4443
```

The first one is for null model where there is no covariate. The second one is what we need.

Likelihood ratio test

Reminder:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L_\theta(x)}{\sup_{\theta \in \Theta} L_\theta(x)}$$

By monotonicity of log function:

$$\log \lambda(x) = \sup_{\theta \in \Theta_0} l_\theta(x) - \sup_{\theta \in \Theta} l_\theta(x)$$

Recall the asymptotic property:

$$-2 \log \lambda \xrightarrow{D} \chi_1^2$$

Likelihood ratio test

- Compute the Likelihood ratio

```
lrt = 2*(cox$loglik[2]-cox2$loglik[2])
lrt
```

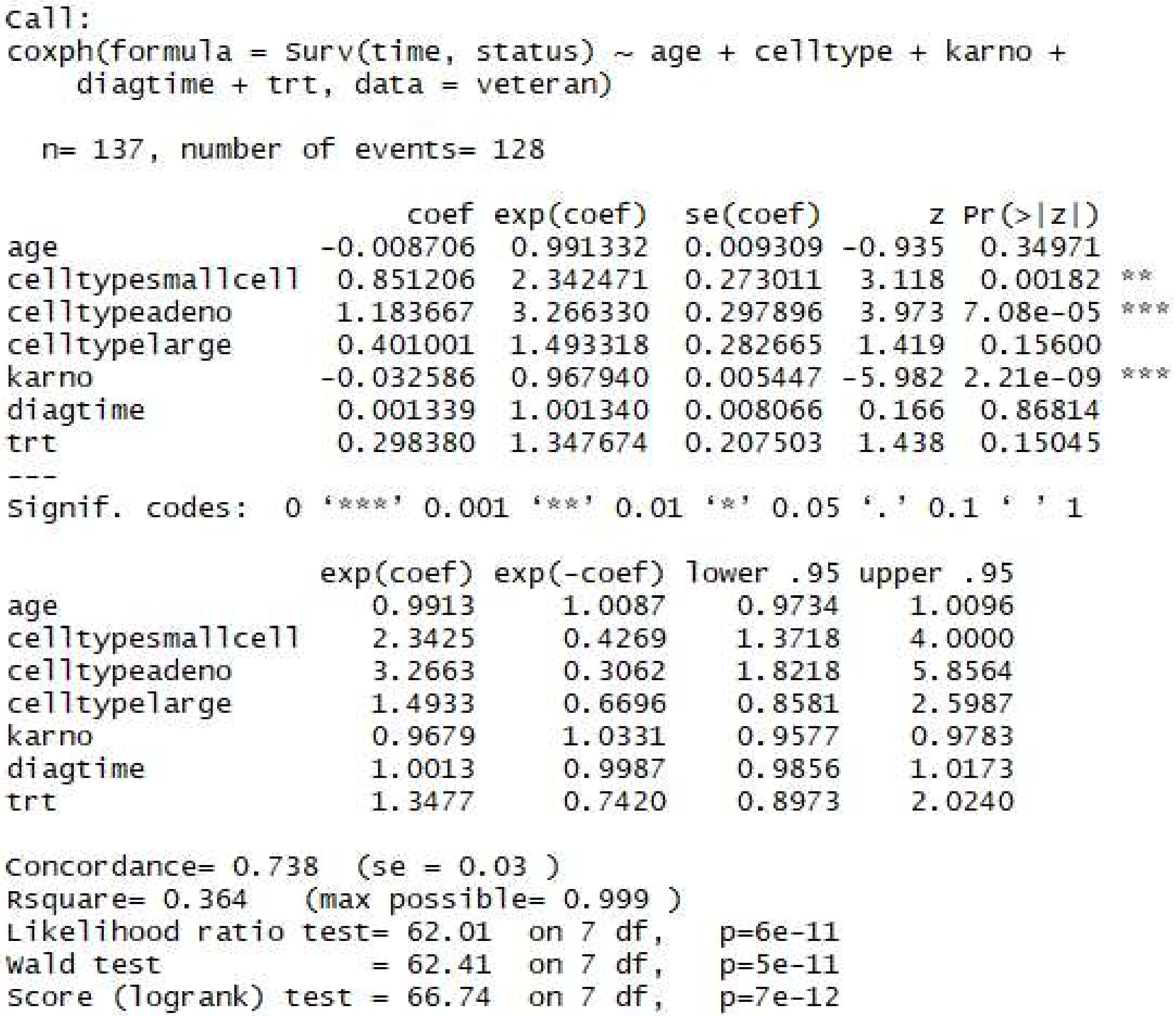
```
## [1] 2.071351
```

- It has approxmiated chi-square distribution with degree 1

```
pchisq(lrt,df=1,lower.tail=FALSE)
```

```
## [1] 0.1500885
```

Comment: coxph() function



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

  n= 137, number of events= 128

                       coef exp(coef)  se(coef)      z Pr(>|z|)
age               -0.008706  0.991332  0.009309 -0.935  0.34971
celltypesmallcell  0.851206  2.342471  0.273011  3.118  0.00182 **
celltypeadeno      1.183667  3.266330  0.297896  3.973 7.08e-05 ***
celltypelarge      0.401001  1.493318  0.282665  1.419  0.15600
karno             -0.032586  0.967940  0.005447 -5.982 2.21e-09 ***
diagtime           0.001339  1.001340  0.008066  0.166  0.86814
trt                0.298380  1.347674  0.207503  1.438  0.15045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
age                  0.9913     1.0087    0.9734    1.0096
celltypesmallcell    2.3425     0.4269    1.3718    4.0000
celltypeadeno        3.2663     0.3062    1.8218    5.8564
celltypelarge        1.4933     0.6696    0.8581    2.5987
karno                0.9679     1.0331    0.9577    0.9783
diagtime             1.0013     0.9987    0.9856    1.0173
trt                  1.3477     0.7420    0.8973    2.0240

Concordance= 0.738  (se = 0.03 )
Rsquare= 0.364   (max possible= 0.999 )
Likelihood ratio test= 62.01  on 7 df,   p=6e-11
Wald test            = 62.41  on 7 df,   p=5e-11
Score (logrank) test = 66.74  on 7 df,   p=7e-12
```

Figure 1: summary(cox)

Construct CI for parameters

```r
confint(cox,level=.95)
```

```
##                         2.5 %       97.5 %
## age               -0.02695186  0.009540464
## celltypesmallcell  0.31611466  1.386298277
## celltypeadeno      0.59980068  1.767533126
## celltypelarge     -0.15301235  0.955013877
## karno             -0.04326240 -0.021908783
## diagtime          -0.01446981  0.017148119
## trt               -0.10831833  0.705079044
```

veteran dataset

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

```r
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
```

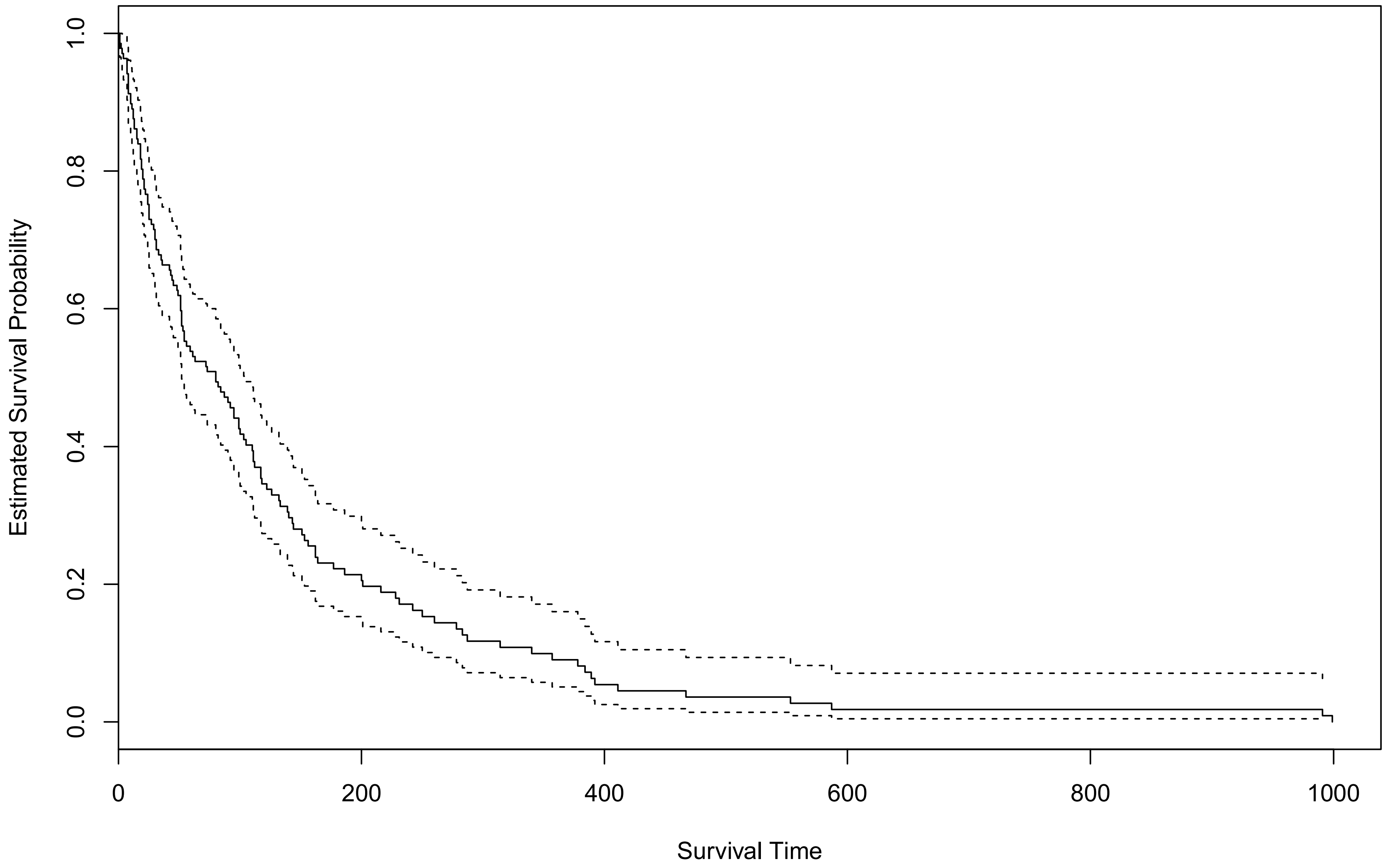
veteran dataset

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 10=yes

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

K-M estimate

We only consider `time` and `status`. Plot the Kaplan–Meier estimate of the survivor function.



Review: K-M estimate

Code:

```
veteran.km = survfit(Surv(time,status)~1, data=veteran)
plot(veteran.km,xlab="Survival Time",
     ylab="Estimated Survival Probability" )
```

Review: Cox proportional hazards model

Semiparametric model for hazard function:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}.$$

- $h_0(t)$ is called *the baseline hazard function*.
- Proportional hazards assumption: $h_0(t)$ only relies on t.
- Time-independence.

Review: `coxph()` function

Description: Fits a Cox proportional hazards regression model. (Run `help('coxph')` for more details.)

```r
cox = coxph(Surv(time,status)~trt,data=veteran)

#Use `cox$loglik` to get log likelihood ratio
```

Review: Construct CI for parameters

```r
# .95 confident interval for `exp(coef)` (harzard ratio)
summary(cox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt, data = veteran
##
##   n= 137, number of events= 128
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## trt 0.01774   1.01790  0.18066 0.098    0.922
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt     1.018     0.9824    0.7144      1.45
##
## Concordance= 0.525  (se = 0.026 )
## Rsquare= 0   (max possible= 0.999 )
## Likelihood ratio test= 0.01  on 1 df,   p=0.9
## Wald test            = 0.01  on 1 df,   p=0.9
```

Log-Log Plots

Our Propotional Harzard Model:

$$h(t) = h_0(t)e^{\beta x}.$$

Reminder (Textbook pg. 15):

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

So we have:

$$\log S(t) = -\int_0^t h_0(t)e^{\beta x}dt = -e^{\beta x}\int_0^t h_0(t)dt.$$

Log-Log Plots

Log again:

$$\log\left(-\log(S(t))\right) = \beta x + \log \int_0^t h_0(t)dt.$$

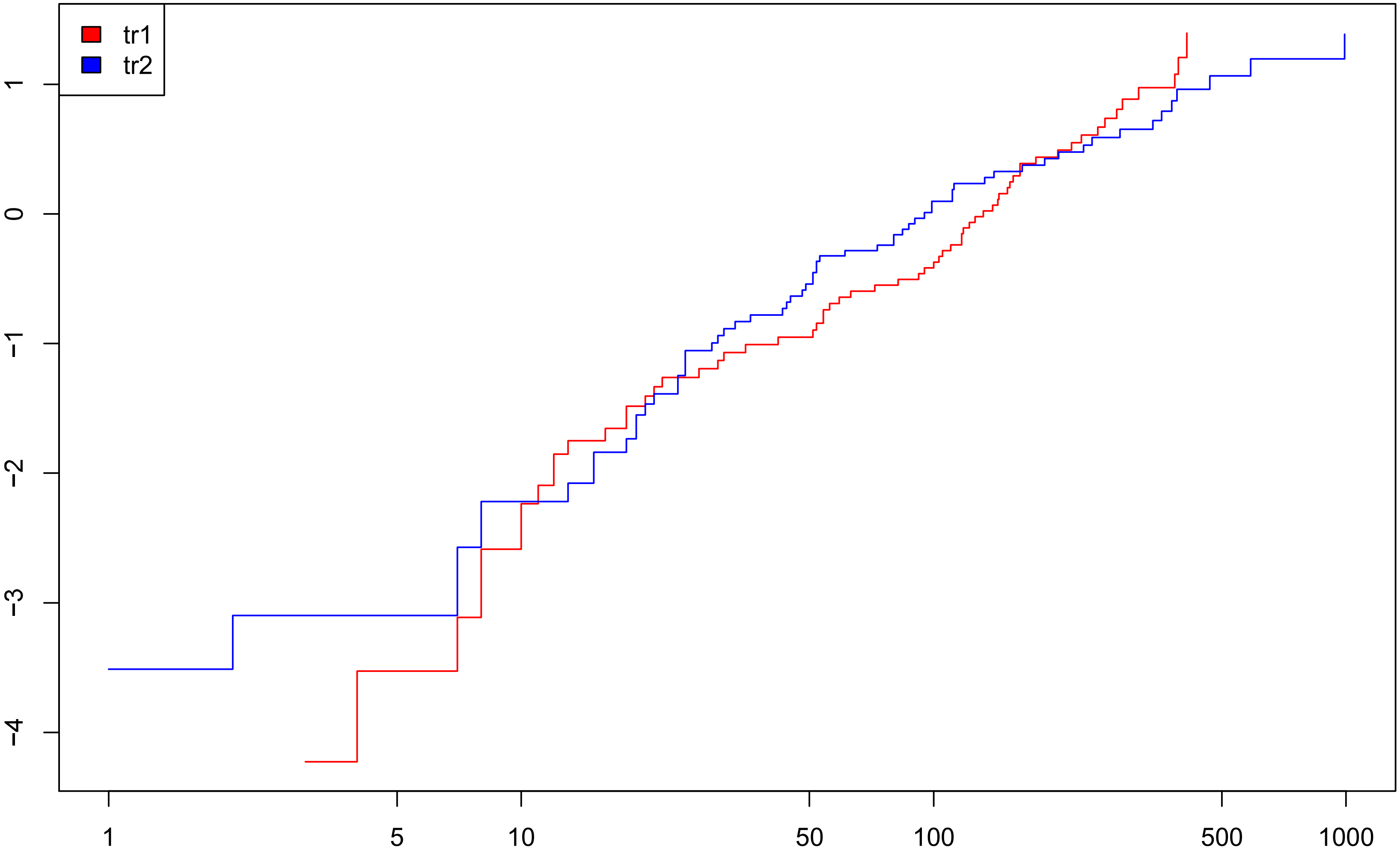
Comment: it should be linear in x.

Log-Log Plots

Let fun='cloglog'. Code:

```r
veteran.km = survfit(Surv(time,status)~trt, data=veteran)
plot(veteran.km, fun='cloglog', col=c('red','blue'))
legend('topleft',c("tr1","tr2"),fill = c("red","blue"))
```

Log-Log Plots



cox.zph() function

Description: Test the proportional hazards assumption for a Cox regression model fit (coxph).

(ref. cox.zph in R Documentation)

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

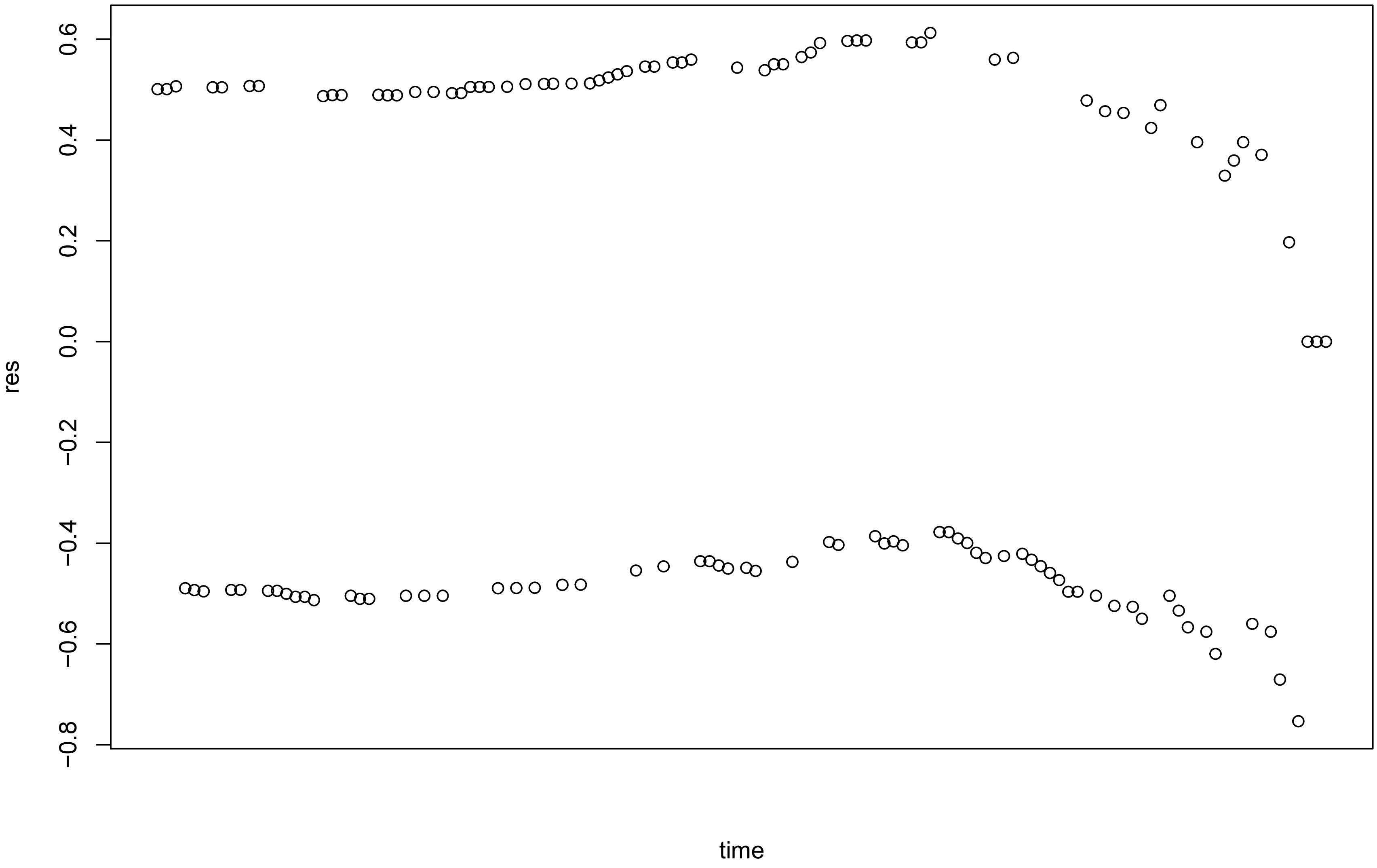
```
##       rho chisq      p
## trt -0.16   3.3 0.0691
```

`cox.zph()` function

When p is small, it means there are time dependent coefficients.

The scaled Schoenfeld residuals are used in the cox.zph function. (ref: `residuals.coxph` in R documentation)

Schoenfeld Residuals



veteran dataset

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

```r
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
```

veteran dataset

- trt: 1=standard 2=test
- celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time
- status: censoring status
- karno: Karnofsky performance score (100=good)
- diagtime: months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 10=yes

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

Review: Cox PH Model

Semiparametric model for hazard function:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}.$$

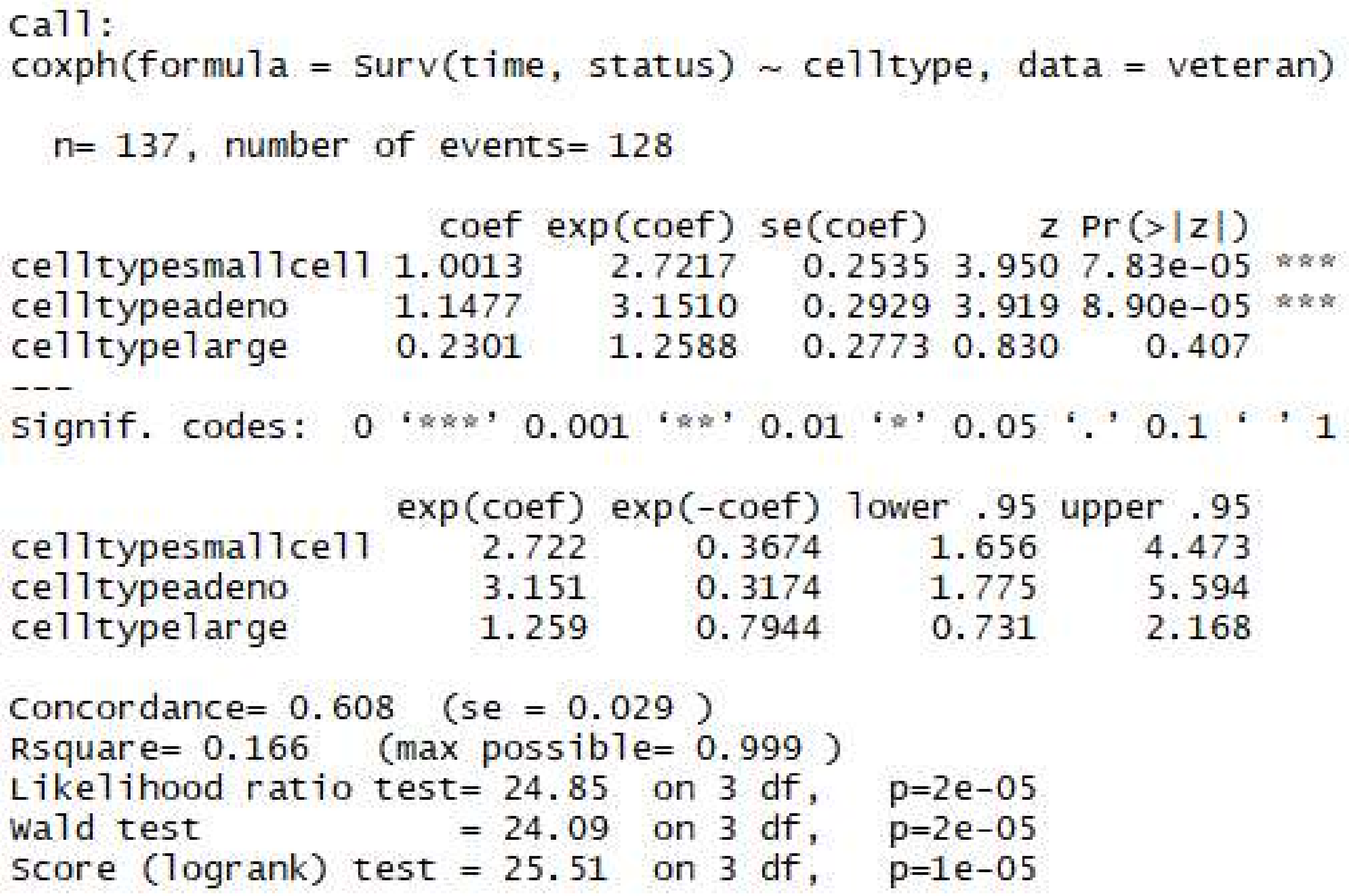
- $h_0(t)$ is called *the baseline hazard function*.
- $h_0(t)$ only relies on t.
- Time-independence.

Review: Cox PH Model

We want to see if there is a significant difference between cancer cell types.

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

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
```

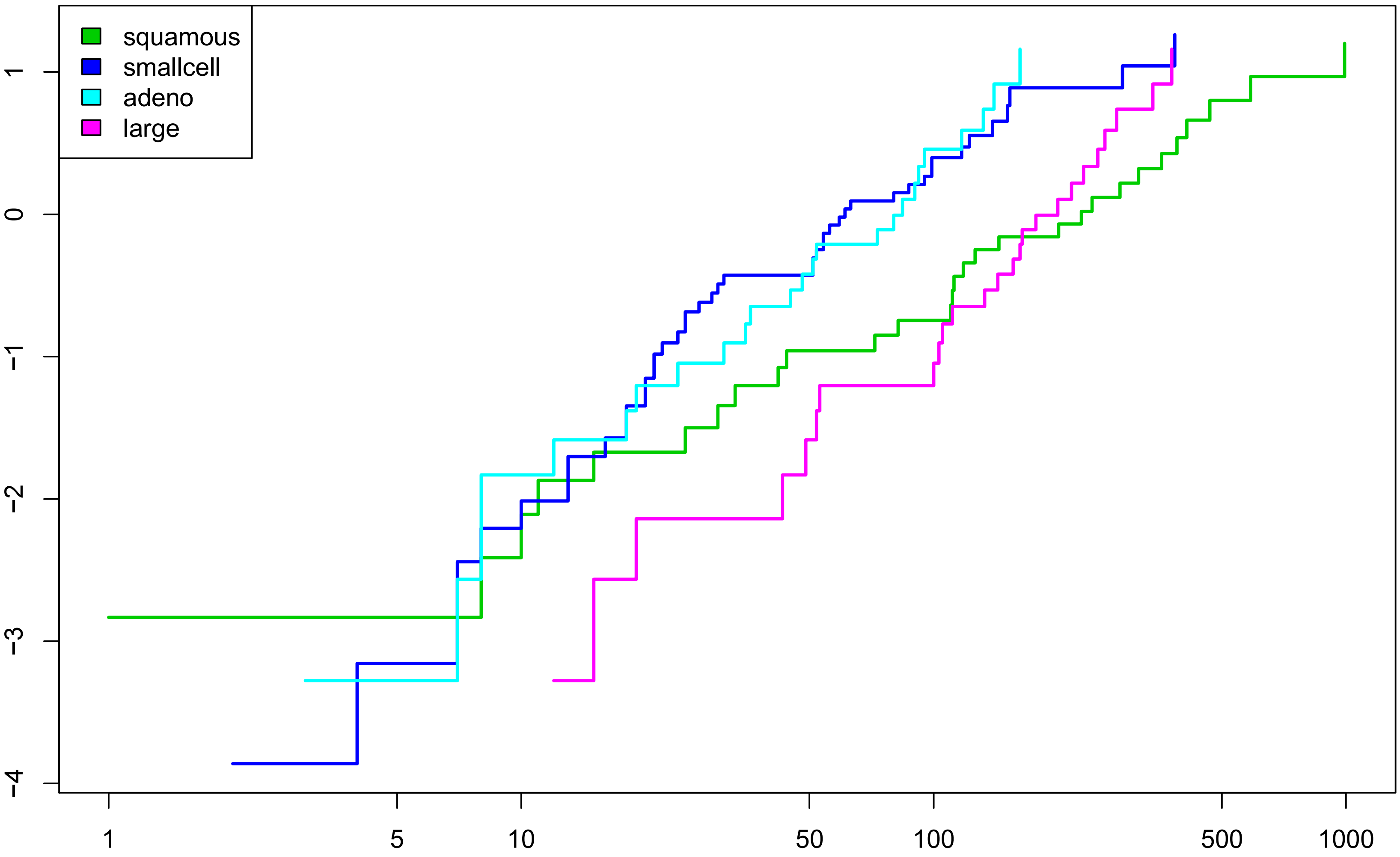
Figure 1: summary(fit)

Review: Log-Log Plot

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

```r
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

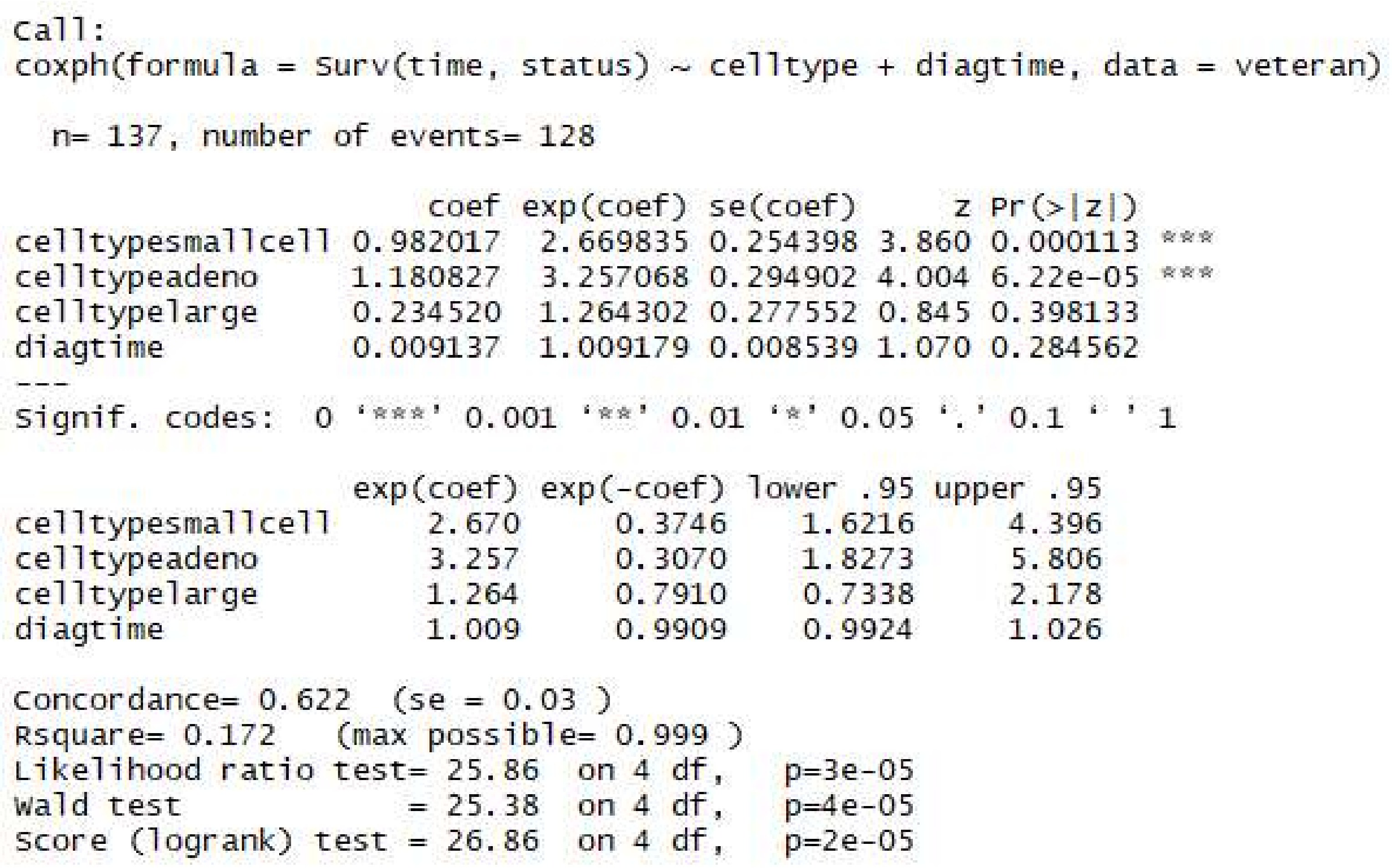


Review: Cox PH Model (with covariates)

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)
```

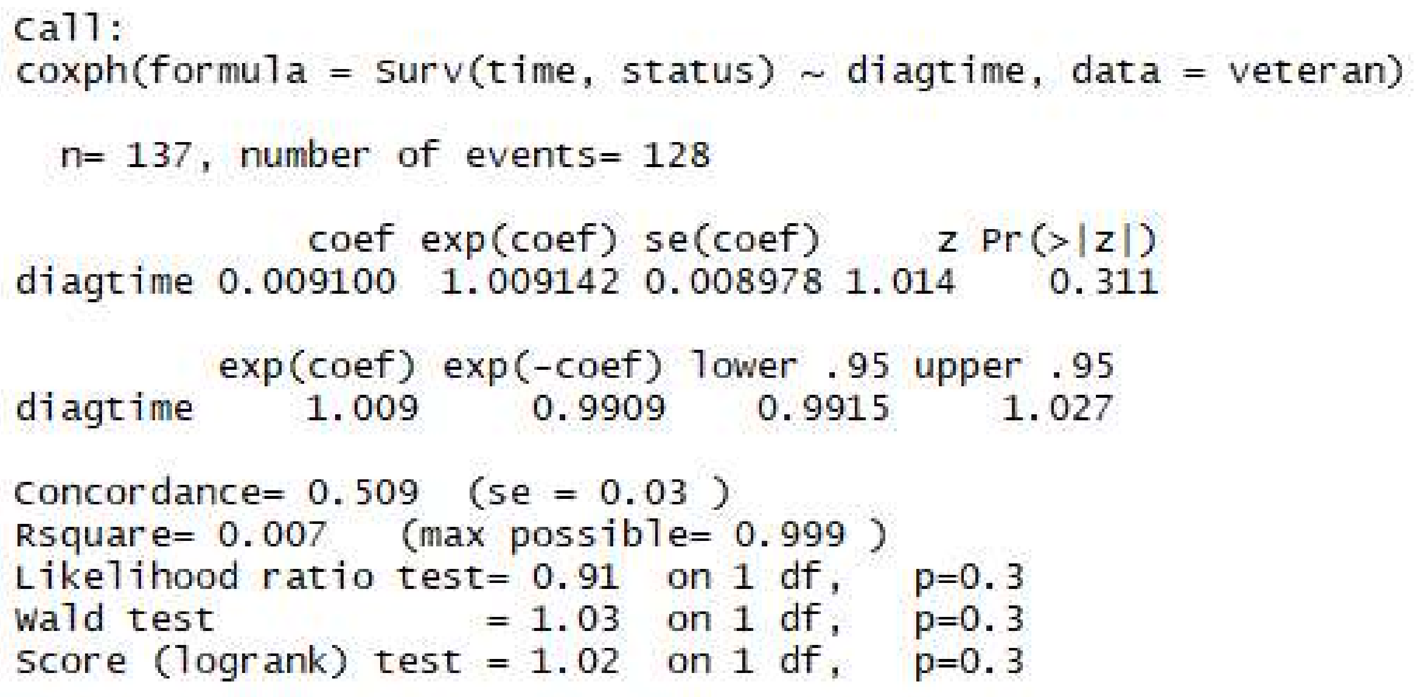
Review: Cox PH Model (with covariates)



```
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:

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

Two ways to compute the likelihood ratio statistic:

```r
#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
```

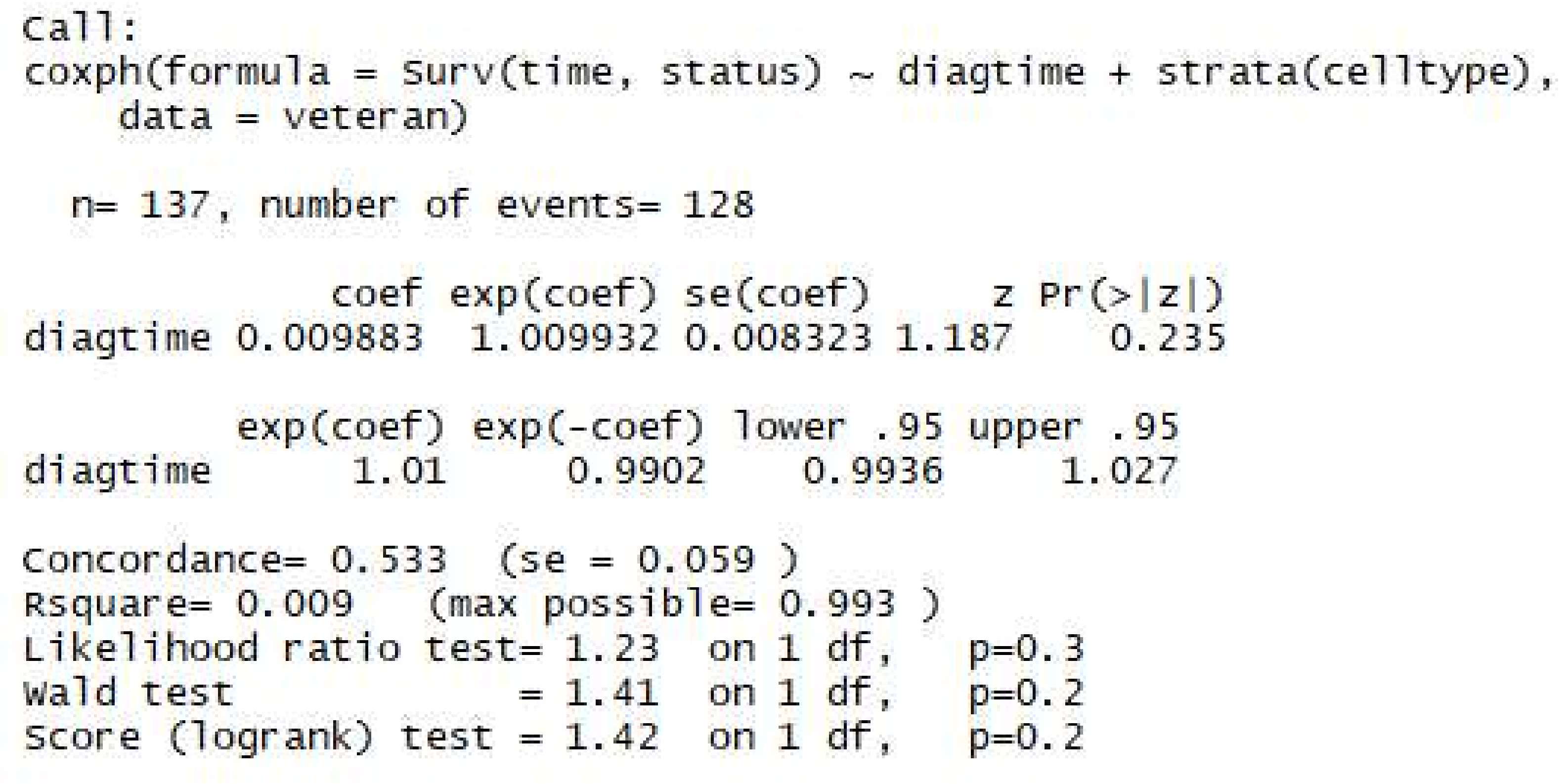
p-value is small $(0.0210 < 0.05) \implies$ **the PH assumption is violated.**

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.)

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

Stratified Cox PH Model

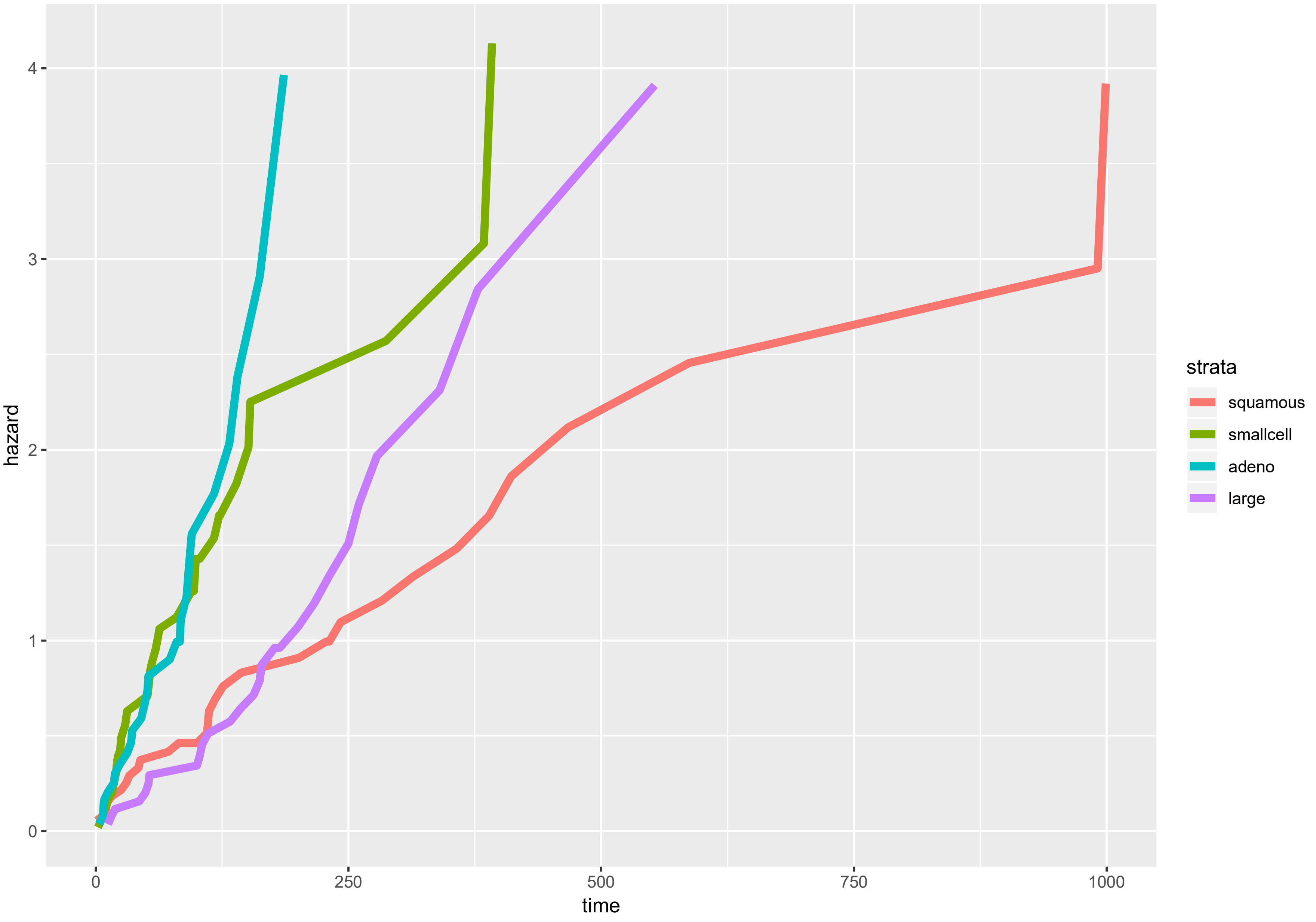


```
Call:
coxph(formula = Surv(time, status) ~ diagtime + strata(celltype),
    data = veteran)

  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)

Stratified Cox PH Model

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



Stratified Cox PH Model

Codes:

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

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 $\Longrightarrow$ the interaction term is not significant

veteran dataset

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

```r
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)
```

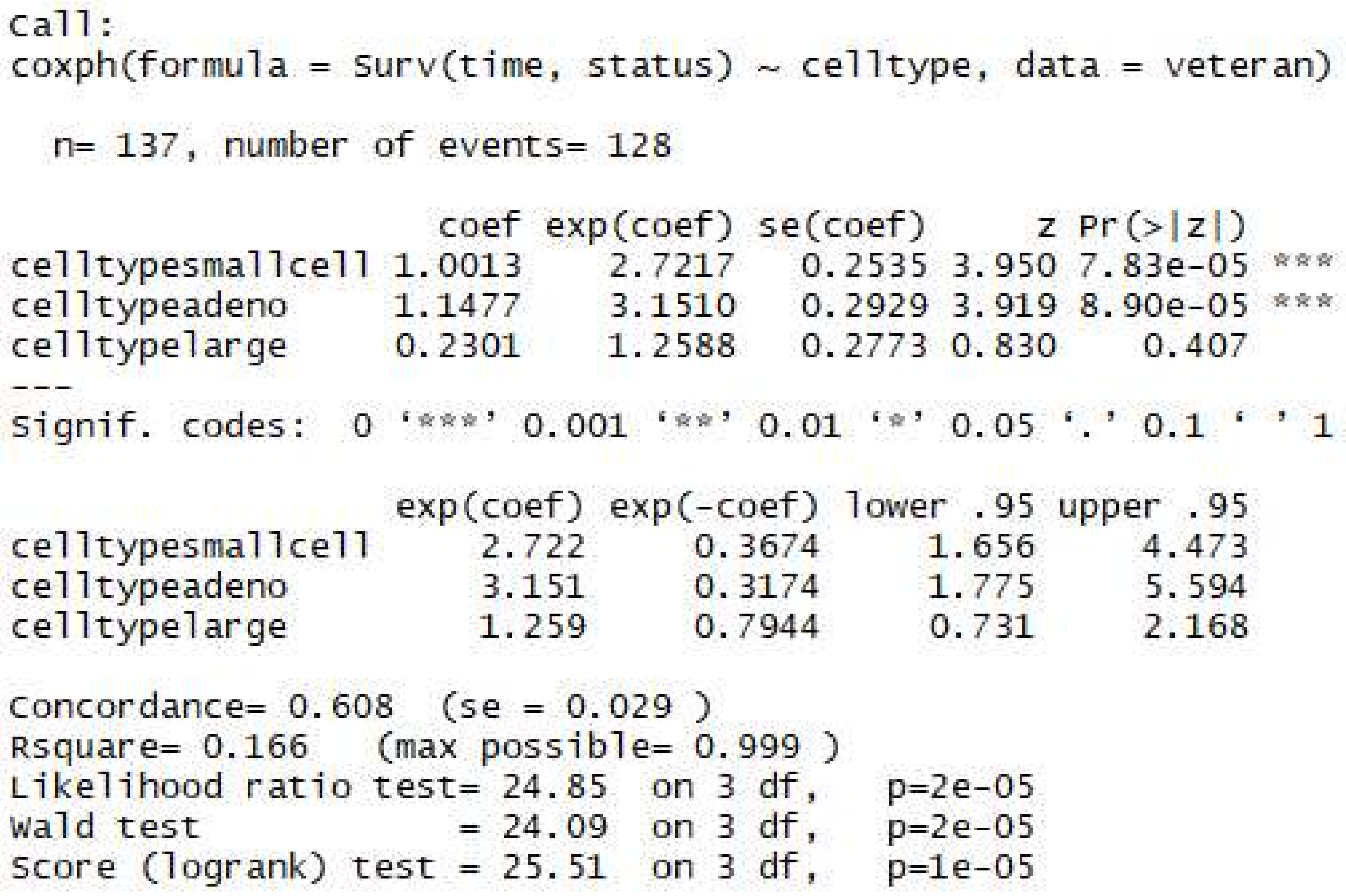
Likelihood ratio test:

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
```

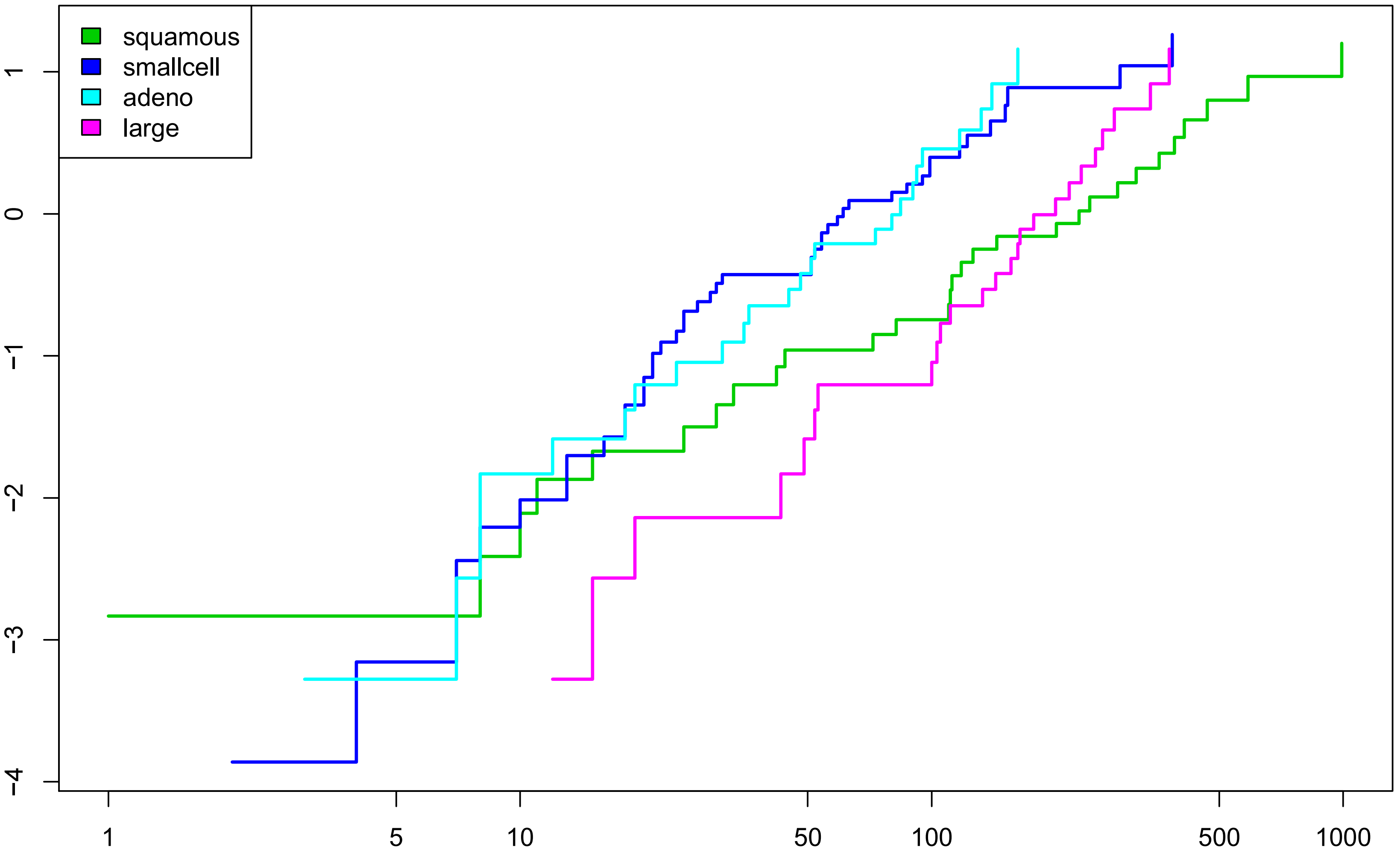
Figure 1: summary(fit)

Review: Log-Log Plot

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

```r
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

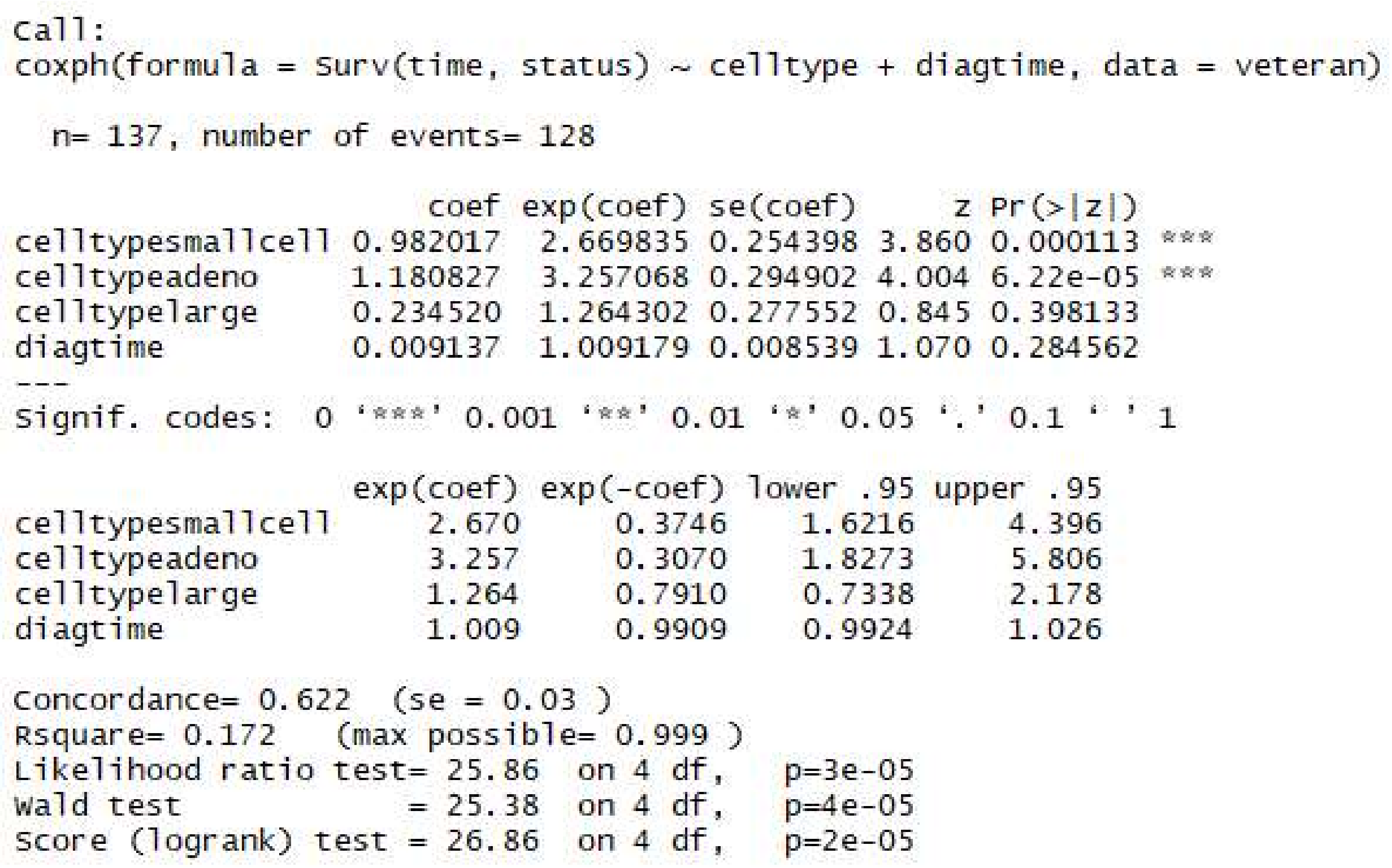


Review: Cox PH Model (with covariates)

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)
```

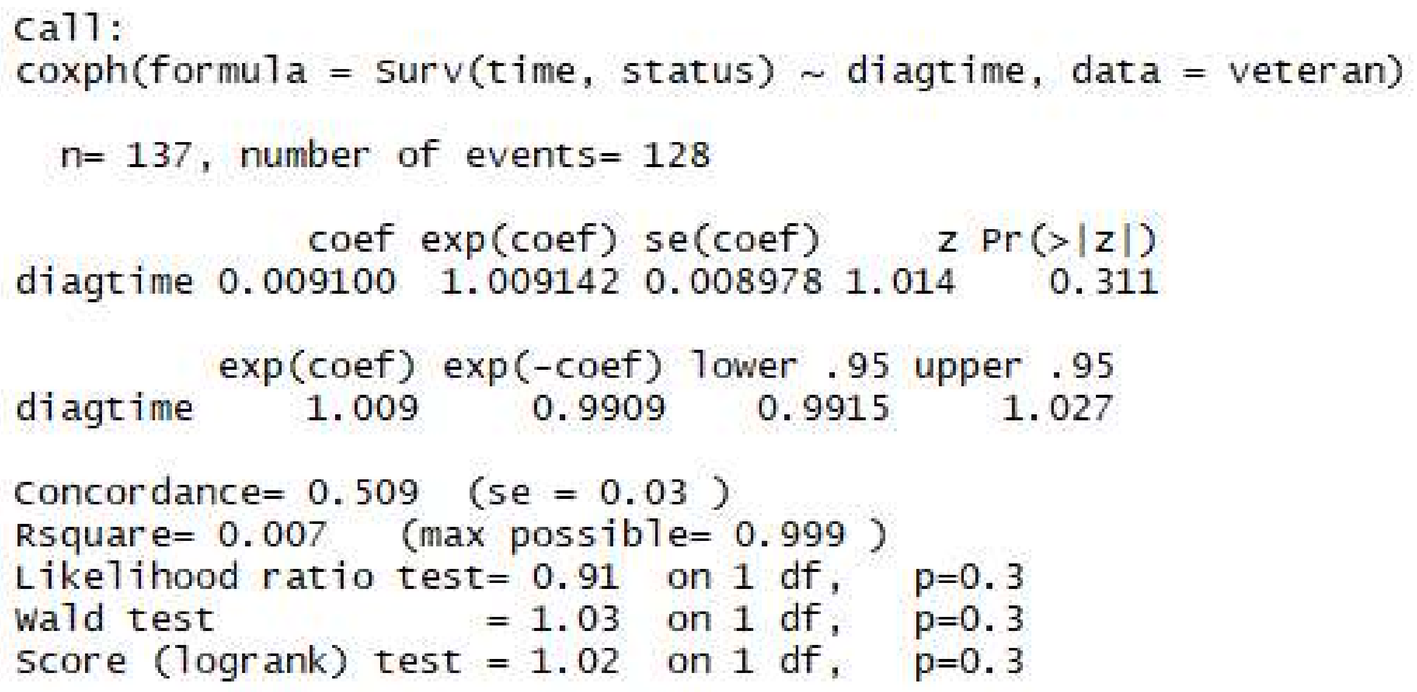
Review: Cox PH Model (with covariates)



```
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:

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

Two ways to compute the likelihood ratio statistic:

```r
#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.

```r
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
```

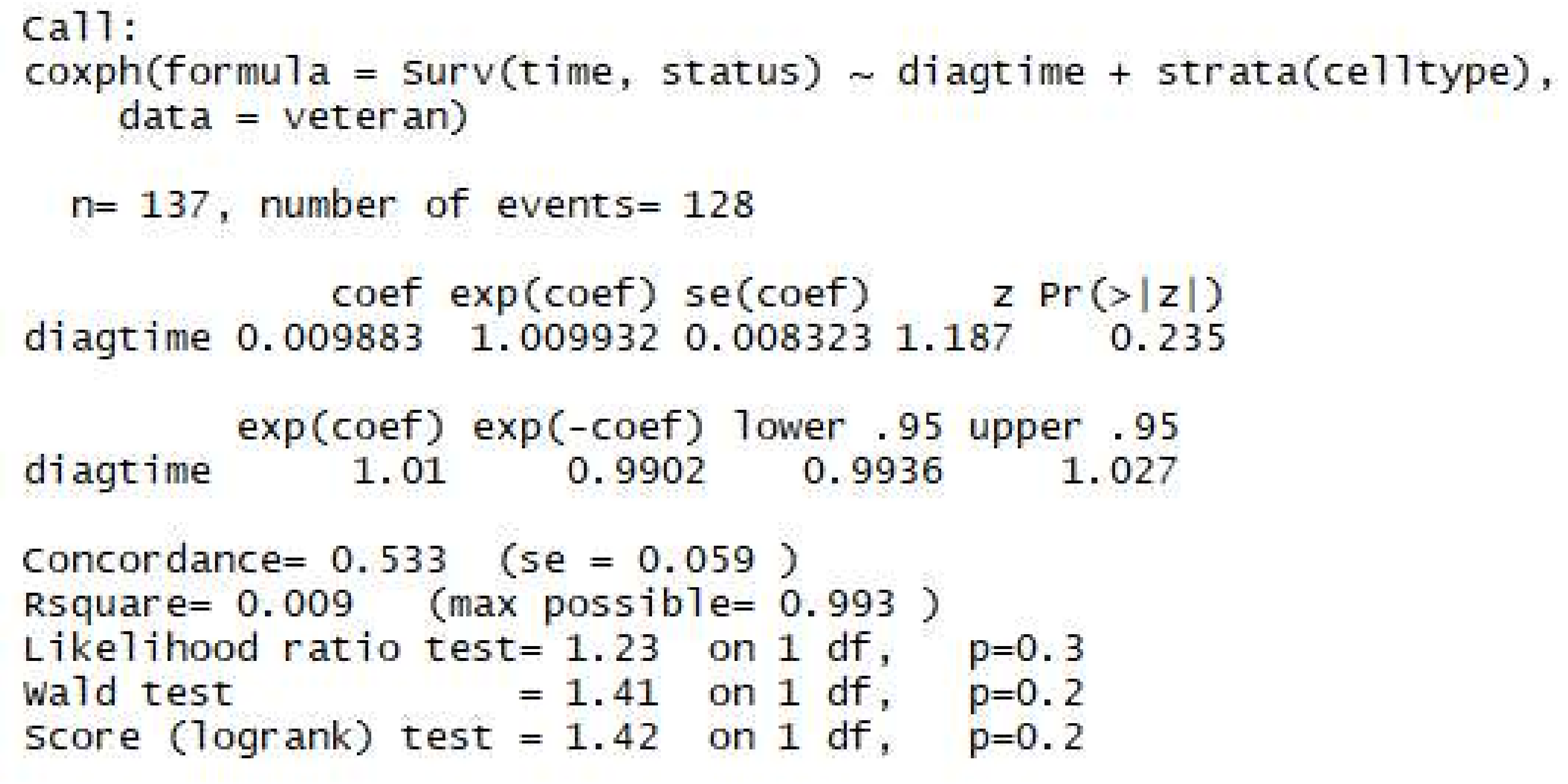
p-value is small $(0.0210 < 0.05) \implies$ **the PH assumption is violated.**

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.)

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

Review: Stratified Cox PH Model

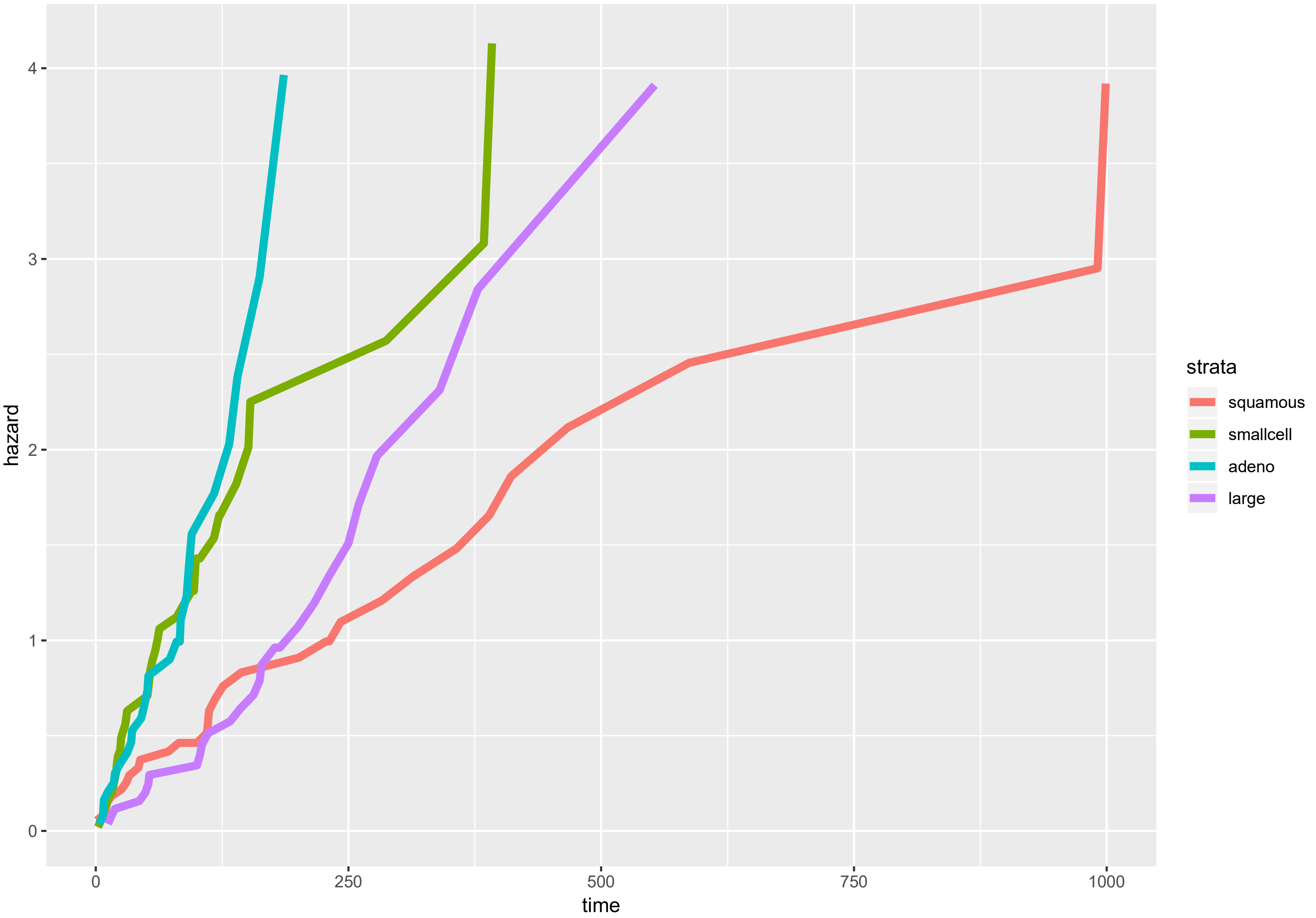


```
Call:
coxph(formula = Surv(time, status) ~ diagtime + strata(celltype),
    data = veteran)

  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=vetera
#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 $\Longrightarrow$ the interaction term is not significant

`survSplit()` function

```r
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.

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

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

```r
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

```r
nrow(veteran)
```

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

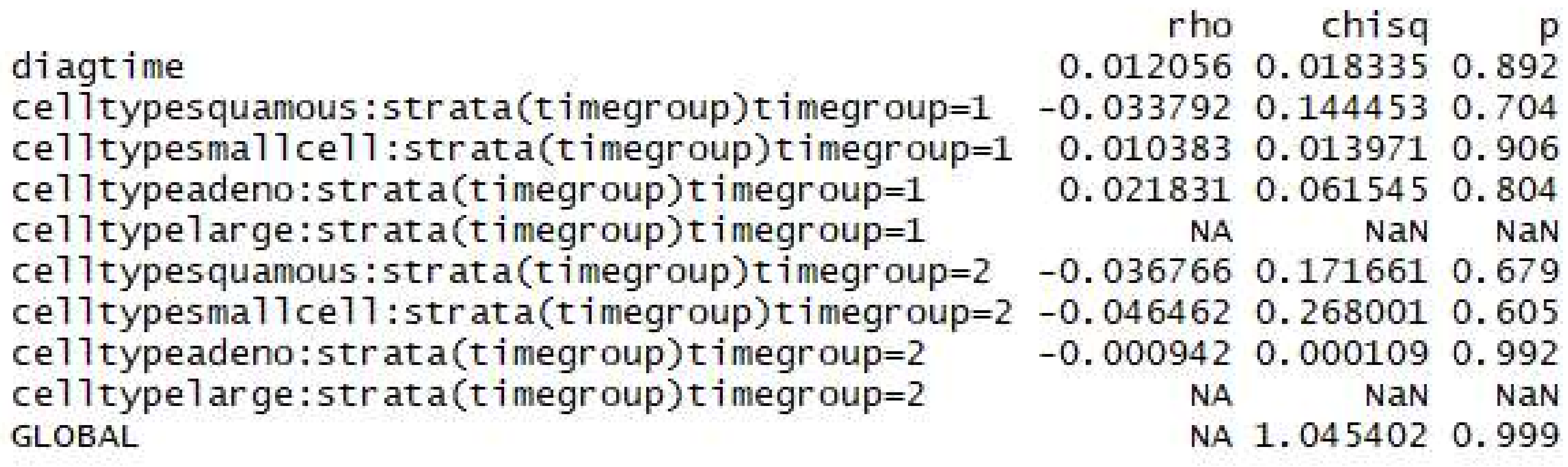
```r
nrow(veteran2)
```

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

survSplit() function

Fit Cox PH model using the new data frame.

```
fit4 = coxph(Surv(tstart, time, status)~celltype:strata(ti
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)