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
```

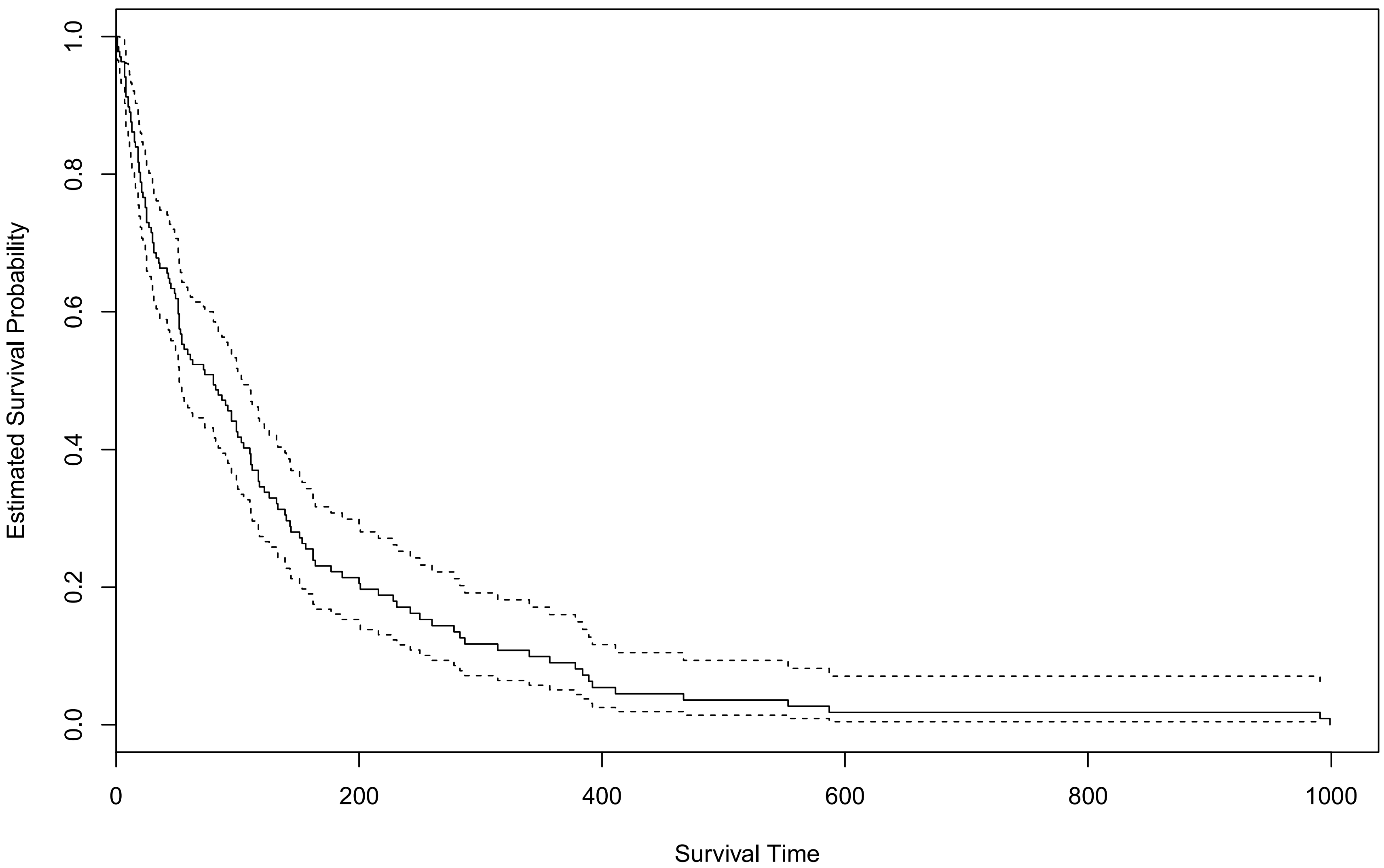
veteran dataset

- 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.)

```
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 = veterar
##
##   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:

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

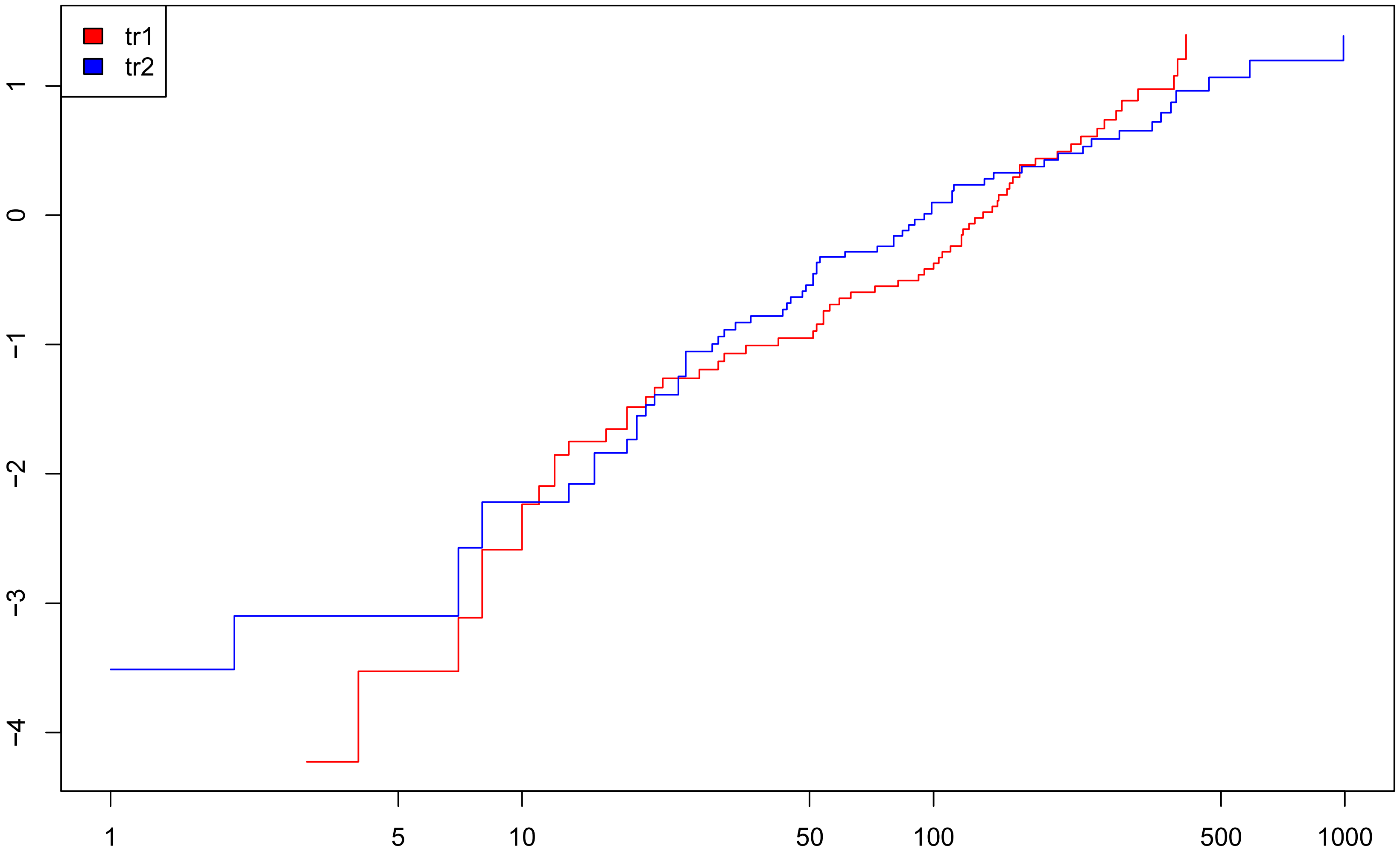
Comment: it should be linear in x.

Log-Log Plots

Let `fun='cloglog'`. Code:

```
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)

```r
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