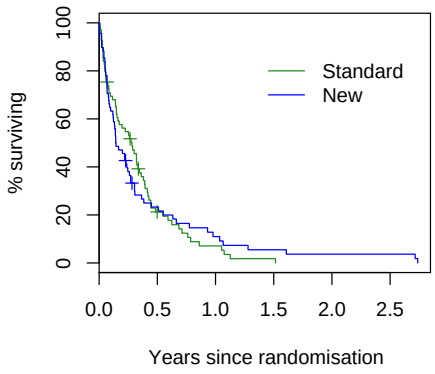

Figure 1: Survival distributions for two lung cancer treatments

Proportional hazards models

The mainstay of survival analysis in the medical world is the Cox proportional hazards model and its extensions. This expresses the hazard (or rate) of events as an unspecified baseline hazard function multiplied by a function of the predictor variables.

Writing $h(t;z)$ for the hazard at time t with predictor variables $Z = z$ the Cox model specifies

$$\log h(t,z) = \log h_0(t) e^{\beta z}.$$

Somewhat unusually for a semiparametric model, there is very little loss of efficiency by leaving $h_0(t)$ unspecified, and computation is, if anything, easier than for parametric models.

A standard example of the Cox model is one constructed at the Mayo Clinic to predict survival in patients with primary biliary cirrhosis, a rare liver disease. This disease is now treated by liver transplantation, but at the same there was no effective treatment. The model is based on data from 312 patients in a randomised trial.

```
> data(pbc)
> mayomodel<-coxph(Surv(time,status)~edtrt+
                       log(bili)+log(protime)+
                       age+platelet,
                 data=pbc, subset=trt>0)
> mayomodel
Call:
coxph(formula = Surv(time, status) ~ edtrt +
          log(bili) + log(protime) +
           age + platelet, data = pbc,
           subset = trt > 0)


                coef exp(coef)
edtrt         1.02980     2.800
log(bili)     0.95100     2.588
log(protime)  2.88544    17.911
age           0.03544     1.036
platelet     -0.00128     0.999
              se(coef)     z       p
edtrt         0.300321  3.43 0.00061
log(bili)     0.097771  9.73 0.00000
log(protime) 1.031908  2.80 0.00520
age           0.008489  4.18 0.00003
platelet      0.000927 -1.38 0.17000

Likelihood ratio test=185  on 5 df, p=0  n= 312
```

The `survexp` function can be used to compare predictions from a proportional hazards model to actual survival. Here the comparison is for 106 patients who did not participate in the randomised trial. They are divided into two groups based on whether they had edema (fluid accumulation in tissues), an important risk factor.

```
> plot(survfit(Surv(time,status)~edtrt,
       data=pbc,subset=trt==-9))
> lines(survexp(~edtrt+
              ratetable(edtrt=edtrt,bili=bili,
                    platelet=platelet,age=age,
                    protime=protime),
              data=pbc,
              subset=trt==-9,
              ratetable=mayomodel,
              cohort=TRUE),
        col="purple")
```

The `ratetable` function in the model formula wraps the variables that are used to match the new sample to the old model.

Figure 2 shows the comparison of predicted survival (purple) and observed survival (black) in these 106 patients. The fit is quite good, especially as people who do and do not participate in a clinical trial are often quite different in many ways.

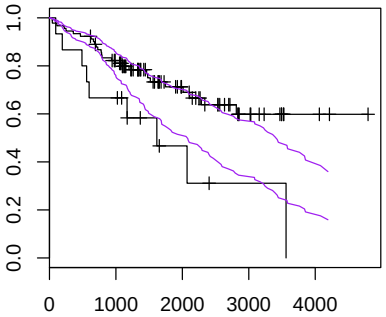

Figure 2: Observed and predicted survival

The main assumption of the proportional hazards model is that hazards for different groups are in fact proportional, *i.e.* that β is constant over time. The