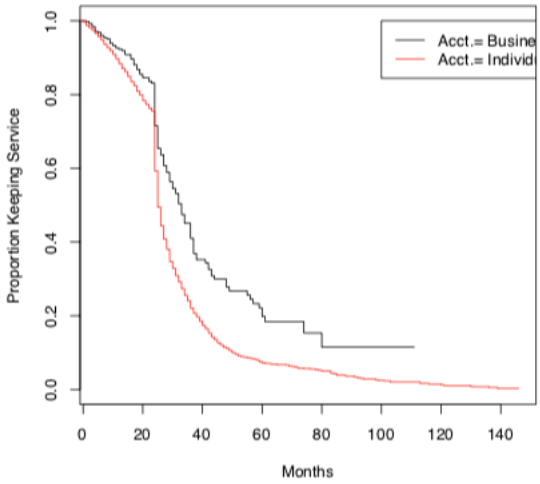
10.8 (Cellphone data: Kaplan-Meier curves and logrank test)

(a) The Kaplan-Meier curves are shown below.



Both plots show steep drops at approximately 24 months. The explanation is that service contracts used to be for 24 months which is

when many people changed their service. Also note that individual customers change service much more frequently than business customers. This is to be expected since businesses sign long term contracts.

(b) Here is the output for the log rank test. We see that the difference between business customers and individual customers is highly significant.

```
survdiff(formula = Surv(Months, Churn) ~ Account_Type,
        data = cell)

n=4912, 88 observations deleted due to missingness.

                   N Observed Expected (O-E)^2/E (O-E)^2/V
Account_Type=B   418      143      225     29.73      35.6
Account_Type=I 4494     2271     2189      3.05      35.6

 Chisq= 35.6  on 1 degrees of freedom, p= 2.37e-09
```

10.9 (Cellphone data: Cox model): The output of the first Cox regression model is shown below.

```
Call:
coxph(formula = Surv(Months, Churn) ~ Account_Type, data = cell)

  n= 4912, number of events= 2414
   (88 observations deleted due to missingness)

                 coef exp(coef) se(coef)     z Pr(>|z|)
Account_TypeI 0.50811   1.66215  0.08633 5.886 3.97e-09
---
              exp(coef) exp(-coef) lower .95 upper .95
Account_TypeI     1.662     0.6016     1.403     1.969

Concordance= 0.516  (se = 0.004 )
Rsquare= 0.008   (max possible= 0.999 )
Likelihood ratio test= 40.29  on 1 df,   p=2.19e-10
Wald test            = 34.64  on 1 df,   p=3.968e-09
Score (logrank) test = 35.39  on 1 df,   p=2.701e-09
```

Note that Account-Type is highly significant.

The output of the second Cox regression model is shown below. Note that Account-Type is no longer significant. The reason is that Account-Type and Line-Count are highly correlated since business customers have many lines whereas individual customers have few.

```
coxph(formula = Surv(Months, Churn) ~ Account_Type + Line_Count,
    data = cell)

  n= 4912, number of events= 2414
   (88 observations deleted due to missingness)

                  coef exp(coef) se(coef)       z Pr(>|z|)
Account_TypeI -0.04573   0.95530  0.08690  -0.526    0.599
Line_Count    -0.51815   0.59562  0.02432 -21.302   <2e-16
---

              exp(coef) exp(-coef) lower .95 upper .95
Account_TypeI    0.9553      1.047    0.8057    1.1327
Line_Count       0.5956      1.679    0.5679    0.6247

Concordance= 0.652  (se = 0.007 )
Rsquare= 0.134   (max possible= 0.999 )
Likelihood ratio test= 706.6  on 2 df,   p=0
Wald test            = 457.6  on 2 df,   p=0
Score (logrank) test = 501.5  on 2 df,   p=0
```

10.10 (Air Miles Reward Program: Cox model): Out of 5330 observations, 22 have negative t values and 1 has a zero t value. I deleted these 23 observations since negative t values don't make sense and are not allowed by any survival distribution model. Incidentally, whether you work with t or $\log(t)$, the Cox model gives the same results since the actual survival times don't enter the partial likelihood function, only their ordered values through the risk sets R_i's.

The R output for the Cox model using t as the response variable is as follows.

```
Call:
coxph(formula = Surv(t, censored) ~ totredeem + prevcat
    + prefood +     pregas + prebank + preretail + preother
```

```
    + prebonus + baselen,     data = AMRP)

  n= 5259, number of events= 1697
   (48 observations deleted due to missingness)

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
totredeem        -1.046e-04  9.999e-01  1.431e-05 -7.315 2.58e-13
prevcatentertain -2.484e-02  9.755e-01  8.404e-02 -0.296  0.76753
prevcatmerch      1.646e-01  1.179e+00  6.638e-02  2.480  0.01312
prevcattravel     1.942e-01  1.214e+00  6.462e-02  3.006  0.00265
prefood          -2.585e-01  7.722e-01  6.007e-02 -4.303 1.69e-05
pregas           -1.649e-01  8.479e-01  9.107e-02 -1.811  0.07011 .
prebank          -2.779e-02  9.726e-01  1.745e-02 -1.592  0.11133
preretail         7.132e-03  1.007e+00  5.609e-02  0.127  0.89881
preother          2.914e-02  1.030e+00  2.817e-02  1.035  0.30090
prebonus          1.798e-01  1.197e+00  6.664e-02  2.698  0.00698
baselen           2.513e-02  1.025e+00  5.756e-04 43.657  < 2e-16
```

Although `prevcatentertain` has a very high p-value we can't drop it since `entertainment` is one of the four categories of the `prevcat` variable, and it is not significantly different from the reference category `gift certificate`; the other two categories differ significantly from the reference category `gift certificate`. We can drop `pregas`, `prebank`, `preretain` and `preother`. The new fitted model is as follows.

```
Call:
coxph(formula = Surv(t, censored) ~ totredeem + prevcat + prefood +
    prebonus + baselen, data = AMRP)

  n= 5259, number of events= 1697
   (48 observations deleted due to missingness)

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
totredeem        -1.173e-04  9.999e-01  1.236e-05 -9.494  < 2e-16
prevcatentertain -3.344e-02  9.671e-01  8.397e-02 -0.398  0.69041
prevcatmerch      1.487e-01  1.160e+00  6.600e-02  2.253  0.02425
prevcattravel     1.750e-01  1.191e+00  6.415e-02  2.728  0.00638
prefood          -2.208e-01  8.018e-01  5.606e-02 -3.940 8.16e-05
prebonus          1.467e-01  1.158e+00  6.260e-02  2.343  0.01911
```

```
baselen           2.511e-02  1.025e+00  5.747e-04 43.691  < 2e-16
```

Now all variables are significant and we may use this model. The strongest predictors are `prevcatmerch, prevcattravel, prefood, prebonus, baselen`.

10.11 (Recidivism study: Time-dependent employment status): When the time dependent covariate `employed` is included in the model, we see that `Aid` is now even less significant with the p-value nearly doubled ($p = 0.1260$).

```
Call:
coxph(formula = Surv(tstart, tstop, Arrest) ~ Aid + Age + Race +
    Work + Married + Parole + Education + employed, data = sdata)

               coef exp(coef) se(coef)     z       p
Aid         -0.1758    0.8388   0.1149 -1.53  0.1260
Age         -0.0902    0.9137   0.0149 -6.07 1.3e-09
Race        -0.3484    0.7058   0.1915 -1.82  0.0688
Work         0.1411    1.1516   0.1256  1.12  0.2612
Married      0.0950    1.0996   0.2074  0.46  0.6470
Parole       0.0201    1.0203   0.1207  0.17  0.8676
Education   -0.2319    0.7930   0.0798 -2.91  0.0036
employedyes -0.2069    0.8131   0.1227 -1.69  0.0917

Likelihood ratio test=77.1  on 8 df, p=1.84e-13
n= 1405, number of events= 309
```