Analysis of Time-to-Event Data: Study Sheet 8,

Submission by: René-Marcel Kruse

Exercise 1:

Let T denote survival time with survival function $S_T(t)$. Simulate a sample of size $n = 1500$ from a Cox model with hazard rate.

$$h(t; x) = t \cdot exp\{0.5x\}.$$

Use the inverse transform sampling method developed in exercise 5, study sheet 7. Simulate the covariate x_1 from a uniform distribution on the interval `[-3; 3]` and the censoring times from a uniform distribution on the interval `[0; 6]`. Plot the Cox-Snell residuals against the cumulative hazard rate to check the overall goodness-of-fit of the fitted model. For the derivation of the distribution of the Cox-Snell residuals use the distribution of $Y = -ln(S_T(T)) \sim E(\lambda = 1)$

Answer:

Draw samples:

```
x <- runif(1500, -3, 3)
censortimes <- runif(1500, 0, 6)
u <- runif(1500)
```

Insverse sampling method:

```
samples <- sqrt(-2 * log(1 - u) * exp(-0.5 * x))
```

Event times and censoring:

```
time <- pmin(samples, censortimes)
delta <- (samples <= censortimes) * 1
```

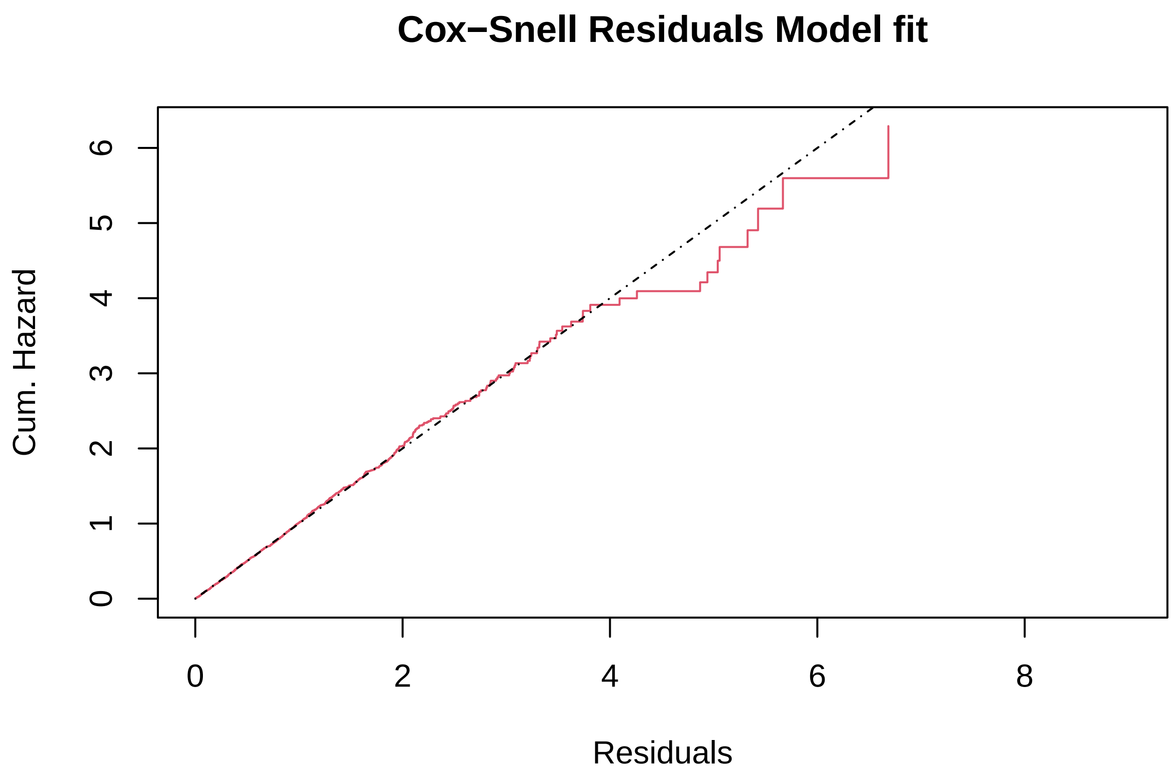
Cox Model:

```
coxmodel <- coxph(Surv(time, delta) ~ x)
summary(coxmodel)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ x)
##
##   n= 1500, number of events= 1112
##
##     coef exp(coef) se(coef)     z Pr(>|z|)
## x 0.4977    1.6449   0.0212 23.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##   exp(coef) exp(-coef) lower .95 upper .95
## x     1.645     0.6079     1.578     1.715
##
## Concordance= 0.719  (se = 0.008 )
## Likelihood ratio test= 584.9  on 1 df,   p=<2e-16
```

```
## Wald test            = 551  on 1 df,   p=<2e-16
## Score (logrank) test = 603.4  on 1 df,   p=<2e-16
```

Cox-Snell residuals:



Exercise 2:

Simulate a sample of size $n = 1500$ from a Cox model with hazard rate

$$\text{(a)} \quad h(t; x) = t \cdot exp\{sin(x_1) + 0.5x_2\}$$

$$\text{(b)} \quad h(t; x) = t \cdot exp\{x_1^2 + 0.5x_2\}$$

Use the inverse transform sampling method developed in exercise 5, study sheet 7. Simulate the covariate x_1 and x_2 from a uniform distribution on the interval `[-3; 3]` and the censoring times from a uniform distribution on the interval `[0; 6]`. Obtain the martingale residuals and deviance residuals and check whether one can use them to make conclusions about the functional form of the covariate x_1. The `loess()` function can be used to smooth the residuals.

Answer:

```r
x1 <- runif(1500, -3, 3)
x2 <- runif(1500, -3, 3)
censortimes <- runif(1500, 0, 6)
u <- runif(1500)
```

(a):

Insverse sampling method:

```
samples <- sqrt(-2 * log(u) * exp(-sin(x1)-0.5*x2))
```

Event times and censoring:

```
time <- pmin(samples, censortimes)
delta <- (samples <= censortimes) * 1
```

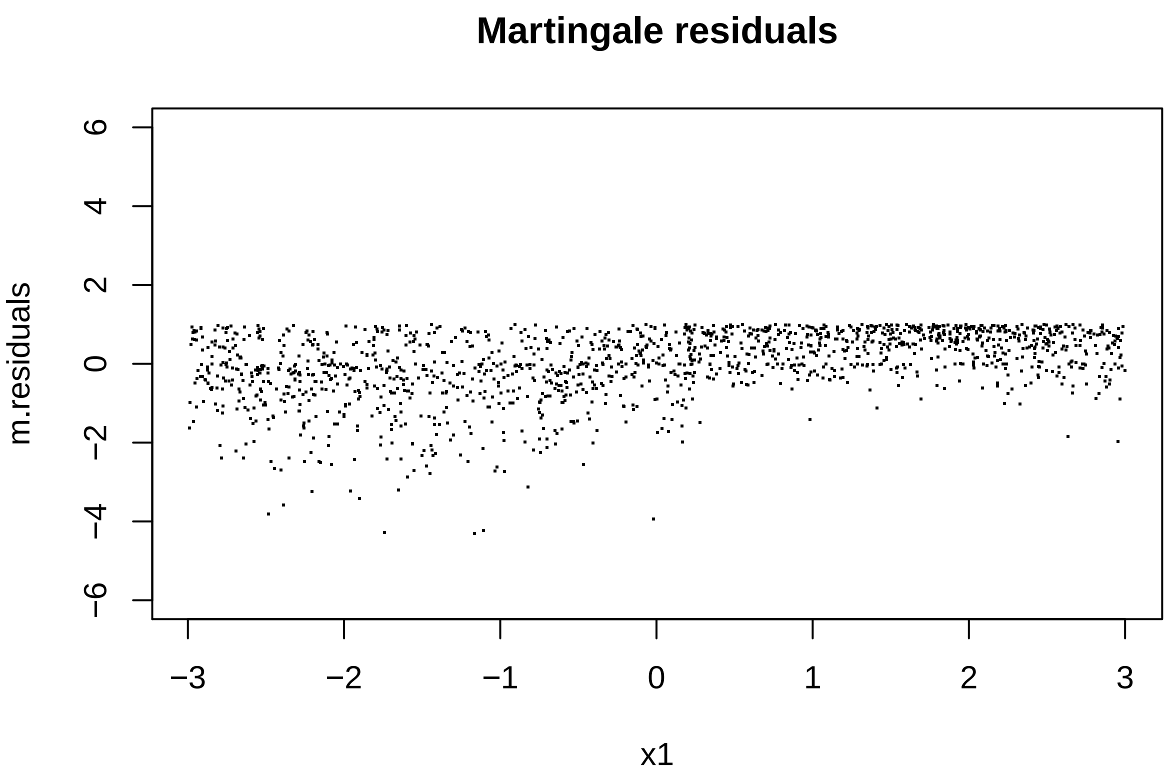
Cox Model:

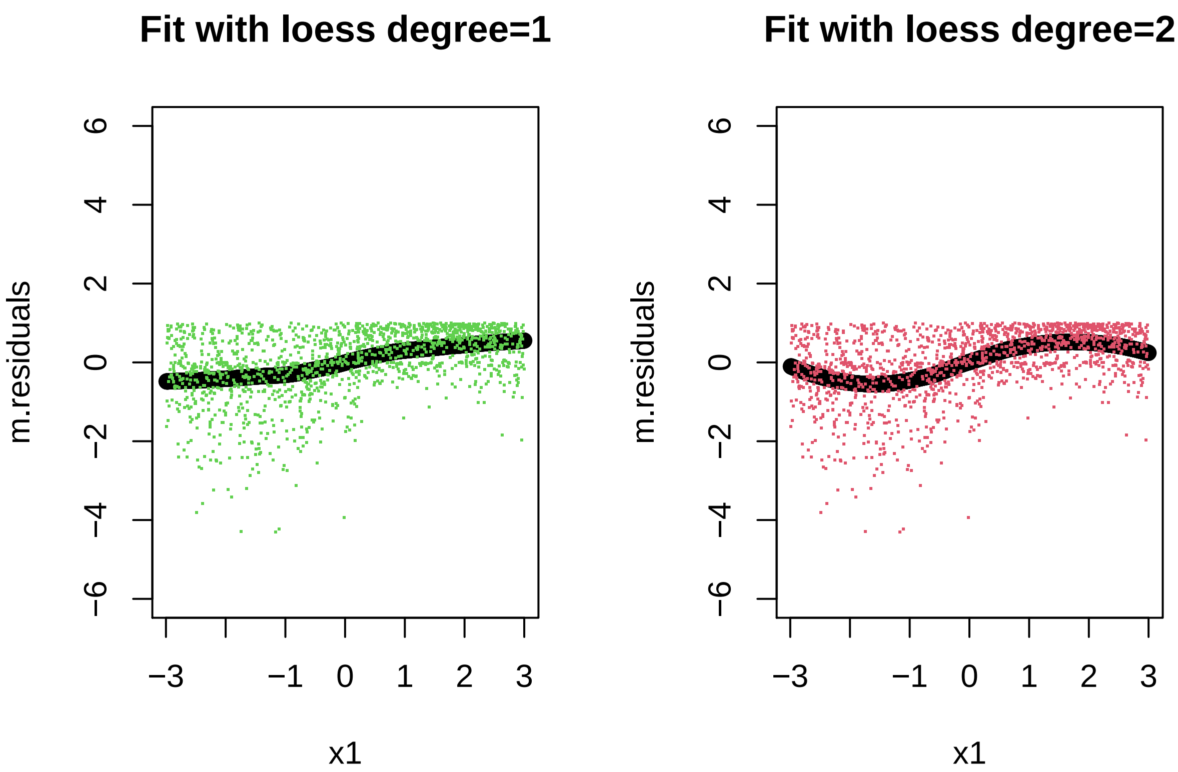
```
coxmodel <- coxph(Surv(time, delta) ~ x2)
summary(coxmodel)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ x2)
##
##   n= 1500, number of events= 1143
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## x2 0.36878   1.44596  0.01925 19.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## x2     1.446     0.6916     1.392     1.502
##
## Concordance= 0.667  (se = 0.008 )
## Likelihood ratio test= 377.4  on 1 df,   p=<2e-16
## Wald test            = 366.9  on 1 df,   p=<2e-16
## Score (logrank) test = 388.8  on 1 df,   p=<2e-16
```

Martingale Residuals:

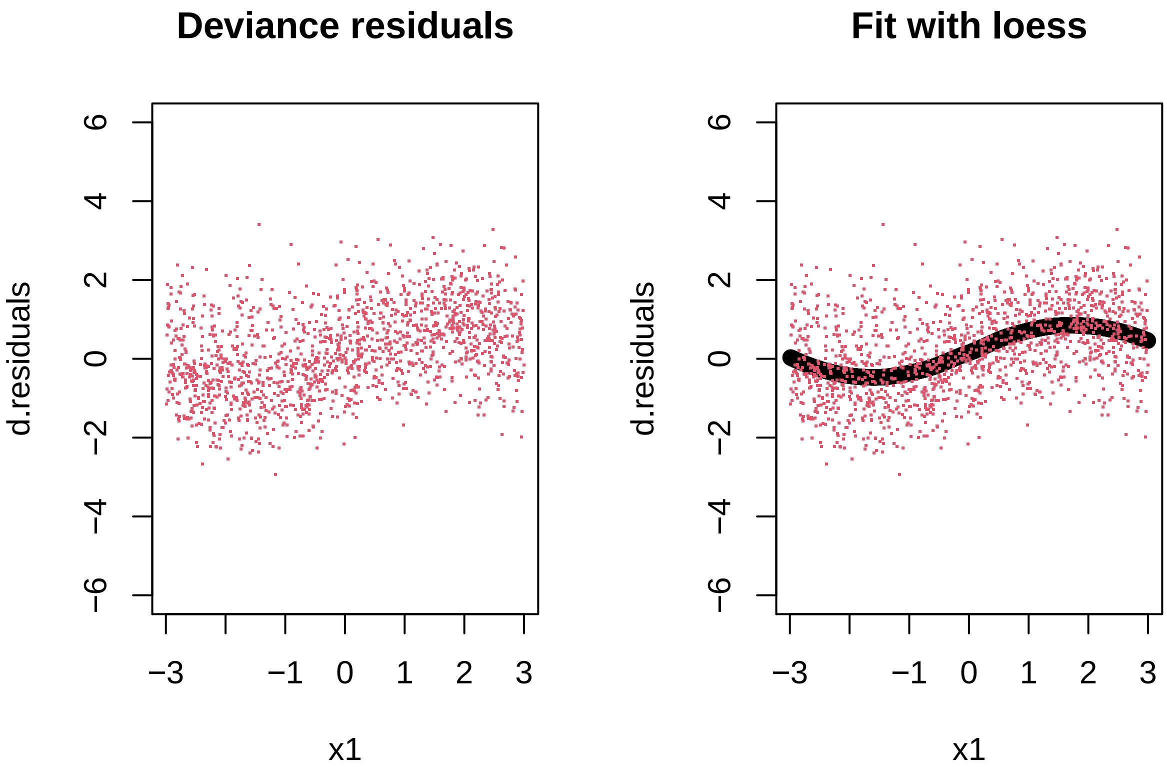
```
m.residuals <- residuals(coxmodel, type="martingale")
m.loess1 <- loess(m.residuals ~ x1, degree=1)
m.loess2 <- loess(m.residuals ~ x1, degree=2)
```





Deviance Residuals:

```
d.residuals <- residuals(coxmodel, type="deviance")
d.loess <- loess(d.residuals ~ x1, degree=2)
```



(b):

Insverse sampling method:

```
samples <- sqrt(-2 * log(u) * exp(- x1^2 - 0.5 * x2))
```

Event times and censoring:

```
time <- pmin(samples, censortimes)
delta <- (samples <= censortimes) * 1
```

Cox Model:

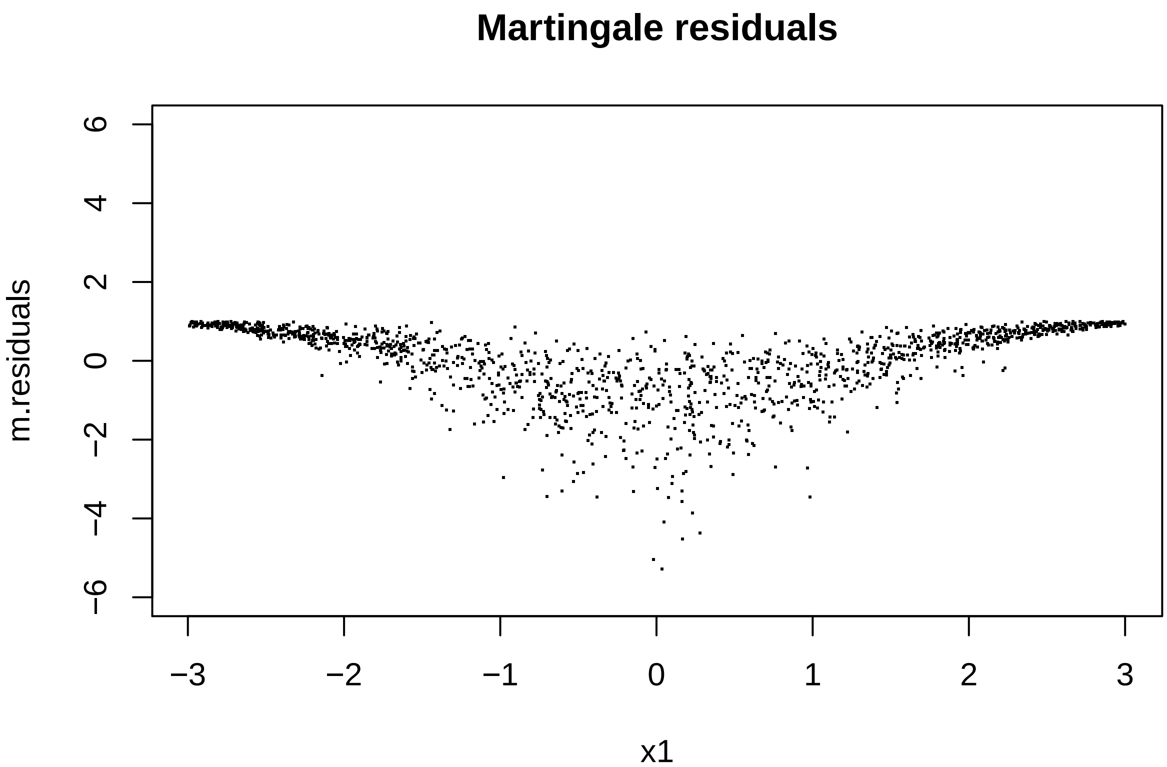
```
coxmodel <- coxph(Surv(time, delta) ~ x2)
summary(coxmodel)
```

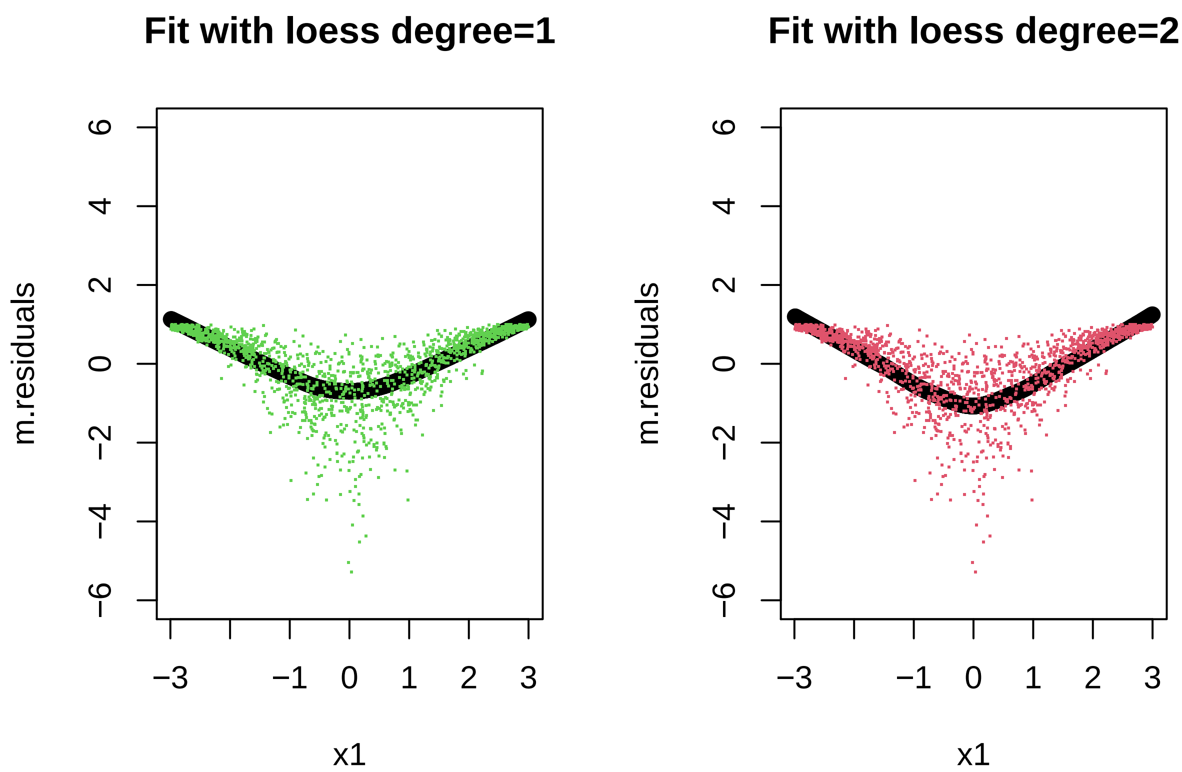
```
## Call:
## coxph(formula = Surv(time, delta) ~ x2)
##
##   n= 1500, number of events= 1362
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## x2 0.17220   1.18792  0.01665 10.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## x2     1.188     0.8418      1.15     1.227
##
```

```
## Concordance= 0.575  (se = 0.008 )
## Likelihood ratio test= 107.5  on 1 df,   p=<2e-16
## Wald test            = 107  on 1 df,   p=<2e-16
## Score (logrank) test = 108.7  on 1 df,   p=<2e-16
```

Martingale Residuals:

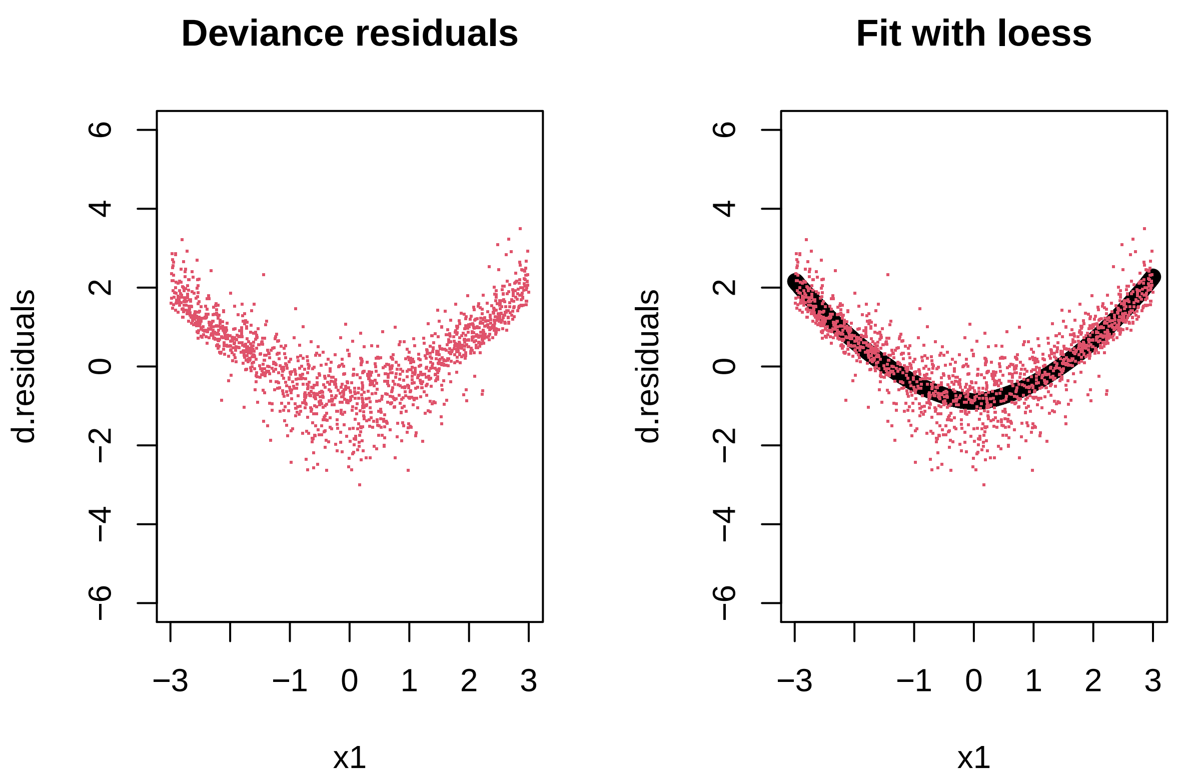
```
m.residuals <- residuals(coxmodel, type="martingale")
m.loess1 <- loess(m.residuals ~ x1, degree=1)
m.loess2 <- loess(m.residuals ~ x1, degree=2)
```





Deviance Residuals:

```
d.residuals <- residuals(coxmodel, type="deviance")
d.loess <- loess(d.residuals ~ x1, degree=2)
```



Exercise 3:

In the lectures, the martingale property has been stated as follows (see slide 7 of the set of slides "Refinements of the semiparametric proportional hazards model"):

$$E[dM(t)|F_{t-}] = 0 \quad \text{for all } t.$$

Show that equation (1) is equivalent to

$$E[M(t)|F_s] = M(s) \quad \text{for all } s.$$

Answer:

We do know that $M(t)$ is defined as follows

$$M(t) = N(t) - \Delta(t)$$

where as $N(t)$ represents a counting process and $\Delta(t)$ represents the cumulative intensity of the process itself.

We can now employing the law of total expectations reformulate as follows

$$\begin{aligned}
E[M(t)|F_s] - M(s) &= E[M(t) - M(s)|F_s] \\
&= E[\int_s^t dM(u)|F_s] \\
&= \int_s^t E[dm(u)|F_s] \\
&= \int_s^t E[E[dM(u)|F_s, F_{u-}]|F_s] \\
&= \int_s^t E[E[dM(u)|F_{u-}]|F_s] \\
&= 0
\end{aligned}$$

Exercise 4:

The file `resmelanoma.prn` that is available in the Stud.IP folder "Data" contains survival times from 30 resected melanoma patients (for a description of the data, see the file `resmelanomahelp.txt`).1 Let ageg denote the age group with ageg = 1 if age < 45 and ageg = 2 otherwise. Fit the survival times with an ageg-stratified Cox proportional hazards model with the covariates sex and treatment received.

Answer:

```
##   ID AGE SEX INI2 INI3A INI3B INI4A TRT RTIME CRT STIME CST
## 1  1  59   0    0     0     1     0   1  33.7   0  33.7   0
## 2  2  50   0    0     0     1     0   1   3.8   1   3.9   1
## 3  3  76   1    0     0     1     0   1   6.3   1  10.5   1
## 4  4  66   0    0     0     1     0   1   2.3   1   5.4   1
## 5  5  33   1    0     0     1     0   1   6.4   1  19.5   1
## 6  6  23   0    0     0     1     0   1  23.8   0  23.8   0

##   ID AGE SEX INI2 INI3A INI3B INI4A TRT RTIME CRT STIME CST AGEG
## 1  1  59   0    0     0     1     0   1  33.7   0  33.7   0    2
## 2  2  50   0    0     0     1     0   1   3.8   1   3.9   1    2
## 3  3  76   1    0     0     1     0   1   6.3   1  10.5   1    2
```

```
## 4  4  66   0    0     0     1     0   1   2.3   1   5.4   1    2
## 5  5  33   1    0     0     1     0   1   6.4   1  19.5   1    1
## 6  6  23   0    0     0     1     0   1  23.8   0  23.8   0    1
```

```
## Call:
## coxph(formula = Surv(STIME, CST) ~ SEX + TRT + strata(AGEG),
##     data = data)
##
##   n= 30, number of events= 10
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## SEX 1.1153    3.0505   0.7301 1.528    0.127
## TRT 0.8799    2.4106   0.6674 1.318    0.187
##
##     exp(coef) exp(-coef) lower .95 upper .95
## SEX     3.051     0.3278    0.7292    12.761
## TRT     2.411     0.4148    0.6517     8.916
##
## Concordance= 0.624  (se = 0.074 )
## Likelihood ratio test= 3.31  on 2 df,   p=0.2
## Wald test            = 3.26  on 2 df,   p=0.2
## Score (logrank) test = 3.42  on 2 df,   p=0.2
```

Exercise 5:

The file `prison.txt`, which is available in the Stud.IP folder "Data", contains data from an experimental study of recidivism of 432 male prisoners, who were observed for a year after being released from prison. Half of the prisoners were randomly given financial aid when they were released. The following table gives a description of the observed variables:

Variable	Description
week	week of first arrest after release, or censoring time
arrest	the event indicator, 1 = arrested , 0 = not
fin	1 = received financial aid, 0 = not
age	in years at the time of release
race	1 = black, 0 = others
wexp	1 = had full-time work experience, 0 = not
mar	1 = married, 0 = not
paro	1 = released on parole, 0 = not
prio	number of prior convictions
educ	codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary)
emp1 - emp52	1 = employed in the corresponding week, 0 = not

(a):

Fit a Cox model to these data. Use backward selection, which is implemented in the function `stepAIC()` function from the R package MASS, to find the best model according to the Akaike Information Criterion (AIC)

```
## Start:  AIC=1334.09
## Surv(week, arrest) ~ fin + age + race + wexp + mar + paro + prio +
```

```
##     as.factor(educ)
##
##                   Df    AIC
## - as.factor(educ)  4 1331.5
## - paro             1 1332.3
## - wexp             1 1332.4
## - mar              1 1333.5
## - race             1 1333.5
## <none>               1334.1
## - fin              1 1336.5
## - age              1 1338.2
## - prio             1 1338.7
##
## Step:  AIC=1331.5
## Surv(week, arrest) ~ fin + age + race + wexp + mar + paro + prio
##
##                   Df    AIC
## - paro             1 1329.7
## - wexp             1 1330.0
## - race             1 1330.6
## - mar              1 1330.9
## <none>               1331.5
## - fin              1 1333.5
## + as.factor(educ)  4 1334.1
## - age              1 1337.5
## - prio             1 1338.5
##
## Step:  AIC=1329.68
## Surv(week, arrest) ~ fin + age + race + wexp + mar + prio
##
##                   Df    AIC
## - wexp             1 1328.2
## - race             1 1328.8
## - mar              1 1329.2
## <none>               1329.7
## + paro             1 1331.5
## - fin              1 1331.6
## + as.factor(educ)  4 1332.3
## - age              1 1335.5
## - prio             1 1337.2
##
## Step:  AIC=1328.21
## Surv(week, arrest) ~ fin + age + race + mar + prio
##
##                   Df    AIC
## - race             1 1327.3
## - mar              1 1328.2
## <none>               1328.2
## + wexp             1 1329.7
## + paro             1 1330.0
## - fin              1 1330.1
## + as.factor(educ)  4 1330.7
## - age              1 1336.4
## - prio             1 1337.6
```

```
##
## Step:  AIC=1327.35
## Surv(week, arrest) ~ fin + age + mar + prio
##
##                    Df    AIC
## <none>                1327.3
## - mar              1 1327.7
## + race             1 1328.2
## + wexp             1 1328.8
## - fin              1 1329.0
## + paro             1 1329.2
## + as.factor(educ)  4 1330.2
## - age              1 1335.4
## - prio             1 1336.2
```

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + mar + prio,
##     data = prison, method = "efron")
##
##   n= 432, number of events= 114
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## fin  -0.36020   0.69753  0.19049 -1.891  0.05864 .
## age  -0.06042   0.94137  0.02085 -2.897  0.00376 **
## mar  -0.53312   0.58677  0.37276 -1.430  0.15266
## prio  0.09751   1.10243  0.02722  3.583  0.00034 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## fin     0.6975     1.4336    0.4802    1.0132
## age     0.9414     1.0623    0.9037    0.9806
## mar     0.5868     1.7042    0.2826    1.2183
## prio    1.1024     0.9071    1.0452    1.1628
##
## Concordance= 0.633  (se = 0.027 )
## Likelihood ratio test= 31.41  on 4 df,   p=3e-06
## Wald test            = 29.98  on 4 df,   p=5e-06
## Score (logrank) test = 31.25  on 4 df,   p=3e-06
```

(b):

In the file prisonlong.txt each row corresponds to one observation per person per week. Fit a Cox model with the time-dependent variable employed to these data.

```
## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + mar, data = prison.long)
##
##          coef exp(coef) se(coef)      z       p
## fin  -0.36020   0.69753  0.19049 -1.891 0.05864
## age  -0.06042   0.94137  0.02085 -2.897 0.00376
## prio  0.09751   1.10243  0.02722  3.583 0.00034
## mar  -0.53312   0.58677  0.37276 -1.430 0.15266
##
```

```
## Likelihood ratio test=31.41  on 4 df, p=2.528e-06
## n= 19809, number of events= 114
```

```
## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + mar + employed, data = prison.long)
##
##              coef exp(coef) se(coef)      z        p
## fin      -0.33898   0.71250  0.19037 -1.781  0.07498
## age      -0.04598   0.95507  0.02059 -2.233  0.02552
## prio      0.08419   1.08784  0.02775  3.034  0.00241
## mar      -0.36119   0.69684  0.37334 -0.967  0.33331
## employed -1.32897   0.26475  0.24979 -5.320 1.04e-07
##
## Likelihood ratio test=67.22  on 5 df, p=3.871e-13
## n= 19809, number of events= 114
```

(c):

Create a variable employed.lag1 which should contain information whether the person was employed in the previous week. Again, fit a Cox model using the variable employed.lag1 instead to employed.

```
## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + mar + employed.lag1, data = prison.long)
##
##                   coef exp(coef) se(coef)      z        p
## fin           -0.34806   0.70606  0.19040 -1.828 0.067536
## age           -0.05118   0.95011  0.02075 -2.467 0.013635
## prio           0.08909   1.09318  0.02759  3.229 0.001243
## mar           -0.41964   0.65728  0.37373 -1.123 0.261507
## employed.lag1 -0.78896   0.45432  0.21700 -3.636 0.000277
##
## Likelihood ratio test=45.98  on 5 df, p=9.156e-09
## n= 19809, number of events= 114
```

(d):

How could you check the assumption of proportional hazards for all the variables of the best model found in (a) using interaction terms with time of observation?

```
## Warning in coxph(Surv(start, stop, arrest.time) ~ fin + fin:I(log(stop)) + : a
## variable appears on both the left and right sides of the formula
```

```
## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + fin:I(log(stop)) +
##     age + prio + mar + employed, data = prison.long)
##
##                      coef exp(coef) se(coef)      z        p
## fin              -0.80920   0.44521  0.86752 -0.933  0.35094
## age              -0.04585   0.95518  0.02057 -2.229  0.02582
## prio              0.08344   1.08702  0.02773  3.009  0.00262
## mar              -0.35752   0.69941  0.37339 -0.957  0.33832
## employed         -1.33307   0.26367  0.24996 -5.333 9.65e-08
## fin:I(log(stop))  0.14836   1.15993  0.26630  0.557  0.57745
##
## Likelihood ratio test=67.54  on 6 df, p=1.305e-12
```

```
## n= 19809, number of events= 114

## Warning in coxph(Surv(start, stop, arrest.time) ~ fin + age + age:I(log(stop))
## + : a variable appears on both the left and right sides of the formula

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     age:I(log(stop)) + prio + mar + employed, data = prison.long)
##
##                      coef exp(coef) se(coef)      z        p
## fin              -0.34600   0.70751  0.19042 -1.817  0.06921
## age               0.12967   1.13845  0.06549  1.980  0.04772
## prio              0.08515   1.08888  0.02778  3.065  0.00217
## mar              -0.32505   0.72249  0.37369 -0.870  0.38439
## employed         -1.32329   0.26626  0.24979 -5.298 1.17e-07
## age:I(log(stop)) -0.05790   0.94374  0.02169 -2.670  0.00758
##
## Likelihood ratio test=73.63  on 6 df, p=7.359e-14
## n= 19809, number of events= 114

## Warning in coxph(Surv(start, stop, arrest.time) ~ fin + age + prio +
## prio:I(log(stop)) + : a variable appears on both the left and right sides of the
## formula

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + prio:I(log(stop)) + mar + employed, data = prison.long)
##
##                        coef exp(coef)  se(coef)      z        p
## fin               -0.340670  0.711294  0.190633 -1.787   0.0739
## age               -0.045957  0.955083  0.020592 -2.232   0.0256
## prio               0.065373  1.067557  0.111385  0.587   0.5573
## mar               -0.362671  0.695815  0.373436 -0.971   0.3315
## employed          -1.328949  0.264755  0.249758 -5.321 1.03e-07
## prio:I(log(stop))  0.006205  1.006224  0.035436  0.175   0.8610
##
## Likelihood ratio test=67.26  on 6 df, p=1.492e-12
## n= 19809, number of events= 114

## Warning in coxph(Surv(start, stop, arrest.time) ~ fin + age + prio + mar + : a
## variable appears on both the left and right sides of the formula

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + mar + mar:I(log(stop)) + employed, data = prison.long)
##
##                      coef exp(coef) se(coef)      z        p
## fin              -0.33555   0.71495  0.19041 -1.762  0.07803
## age              -0.04640   0.95466  0.02061 -2.251  0.02438
## prio              0.08362   1.08721  0.02774  3.014  0.00258
## mar              -2.72922   0.06527  2.50260 -1.091  0.27547
## employed         -1.32677   0.26533  0.24983 -5.311 1.09e-07
## mar:I(log(stop))  0.71048   2.03497  0.71949  0.987  0.32341
##
## Likelihood ratio test=68.46  on 6 df, p=8.453e-13
## n= 19809, number of events= 114
```

```
## Warning in coxph(Surv(start, stop, arrest.time) ~ fin + age + prio + mar + : a
## variable appears on both the left and right sides of the formula

## Call:
## coxph(formula = Surv(start, stop, arrest.time) ~ fin + age +
##     prio + mar + employed + employed:I(log(stop)), data = prison.long)
##
##                           coef exp(coef) se(coef)      z       p
## fin                   -0.34069   0.71128  0.19043 -1.789 0.07361
## age                   -0.04591   0.95513  0.02060 -2.229 0.02583
## prio                   0.08391   1.08754  0.02774  3.025 0.00249
## mar                   -0.35983   0.69779  0.37331 -0.964 0.33510
## employed              -1.83520   0.15958  1.39457 -1.316 0.18819
## employed:I(log(stop))  0.15417   1.16669  0.41525  0.371 0.71044
##
## Likelihood ratio test=67.37  on 6 df, p=1.416e-12
## n= 19809, number of events= 114
```