EXERCISE 11.6

```
data tongue; /* From section 1.11 */
infile '/home/username/Survival analysis/Book data/Ploidy txt'
   firstobs=2;
input DNA survtime death;
run;

data tongue; set tongue;
label survtime='Survival time (weeks)';
run;

/* Either recode DNA to a dummy where 1=abnormal DNA, 0=normal DNA, or
use class statement with proc phreg */
data tongue; set tongue;
if DNA=1 then DNA_abnormal=1;
else if DNA=2 then DNA_abnormal=0;
run;

/* No id variable included, which will be needed when identifying
outliers. */
data tongue;
set tongue;
id=_n_;   * Row number used as id;
run;


/* a) Deviance residuals to investigate model accuracy for each
individual, i.e. to find any outliers */

proc phreg data=tongue noprint;
model survtime*death(0)=DNA_abnormal;
output out=Outp xbeta=Risk_score resdev=Deviance_res;
run;
/* or */
proc phreg data=tongue noprint;
class DNA(ref='2');
model survtime*death(0)= DNA;
output out=Outp xbeta=Risk_score resdev=Deviance_res;
run;
/* 'xbeta' gives the risk scores, 'resdev' gives deviance residuals */

proc sgplot data=Outp;
scatter y=Deviance_res x=Risk_score;
refline 0 / axis=y;
yaxis grid;
run;
/* 'refline' produces a reference line at the value 0 for the chosen
'axis'*/

/*Under light to moderate censoring, Dj should look like a sample of
Normally distributed noise.*/

/* Light or heavy censoring? */
proc lifetest data=tongue;
time survtime*death(0);
run;
/* 34%, i.e., moderate, which is good.
```

```
Under heavy censoring there will be a large collection of points near
zero which will distort the Normal approximation. */


/* Only one, binary, covariate here.
Not that easy to see if this looks like a sample of Normally distributed
noise.
Check for outliers, with "large" values (compare to the critical values
of a standard ND) */

/* A few possible outliers can be seen */


/* To find the largest values of the residuals, i.e. to investigate the
possible outliers, the data can be sorted */

proc sort data=outp;
by Deviance_res Risk_score;
run;

proc print data=outp noobs;
var id DNA_abnormal survtime death Risk_score Deviance_res;
run;

/* The deviance residual is used as some kind of measure of the
difference between the observed number of events and the expected number
of events the individual would experience according to the model. */

/* What does the linear predictor, or the risk score, actually mean?
Let's take a look at the fitted Cox model. */

proc phreg data=tongue;
model survtime*death(0)=DNA_abnormal;
run;

/* Parameter estimate = -0.46104, HR = 0.631*/
/* Individuals with abnormal DNA (DNA_abnormal=1 or DNA=1) have a 37%
lower risk of dying, compared to individuals with normal DNA
(DNA_abnormal=0 or DNA=2). This means that the estimated risk of dying
for individuals with normal DNA is high.
Estimated risk of experiencing the event for an individual with
DNA_abnormal=0 (normal DNA) = exp(b*0) = 1 */

/* In the 'outp' dataset we can see that an individual with normal DNA
and a long lifetime, which is expected to die, but is still alive, gets a
large (negative) residual (id=80, 79). */
/* And at the bottom of the 'outp' dataset we can also see that
individuals with abnormal DNA, with short lifetimes that are dead
(expected to be alive) also have large (positive) deviance residuals
(id=1, 53). */


/* What influence do these possible outliers have on the estimated model?
*/
```

```
/* Excluding one of the possible outliers, the one with the largest
|residual| */
data x1; set tongue;
if id=1 then delete;
run;

proc phreg data=x1;
model survtime*death(0)=DNA_abnormal;
run;
/* Parameter estimate = -0.49753, HR = 0.608, to be compared to the
estimate using the full dataset where Parameter estimate = -0.46104, HR =
0.631. Relatively small difference. */


data x2; set tongue;
if id=53 then delete;
run;

proc phreg data=x2;
model survtime*death(0)=DNA_abnormal;
run;
/* Parameter estimate = -0.41829, HR = 0.658 */
/* Also a relatively small difference to -0.46104, HR = 0.631 */


data x3; set tongue;
if id=80 then delete;
run;

proc phreg data=x3;
model survtime*death(0)=DNA_abnormal;
run;
/* Parameter estimate = -0.56823, HR = 0.567 */
/* Larger difference to -0.46104, HR = 0.631 - but still a relatively
small difference */


data x4; set tongue;
if id=79 then delete;
run;

proc phreg data=x4;
model survtime*death(0)=DNA_abnormal;
run;
/* Parameter estimate = -0.55484, HR = 0.574 */
/* Also a relatively small difference to -0.46104, HR = 0.631 */

/* Individuals no. 80 and 79 thus have the largest impact on the
parameter estimate, but the difference to the estimates from the full
dataset are quite small */
```

```
/* b) Schoenfeld residuals */

proc phreg data=tongue noprint;
model survtime*death(0)=DNA_abnormal;
output out=schres ressch=Schoenf_res;
run;

proc sort data=schres;
by survtime;
run;

data schres; set schres;
nr=_n_;   /* Creates observation numbers */
run;

proc sgplot data=schres;
scatter y= Schoenf_res x=nr;
refline 0 / axis=y;
yaxis grid;
label nr='Observation number';
run;
/* Look closer at e.g. obs. nr 40 in the 'schres' dataset */
/* Normal DNA, medium survival time, this individual should be alive but
is dead. */


****************** CLEAN SAS WORK DATASETS ******************;
proc datasets lib=work nolist memtype=data kill;
run; quit;
/*==================== End of Programme =======================*/
```