EXERCISE 11.2

```
data bf; /* From section 1.14 */
infile '/home/username/Survival analysis/Book data/Breastfeeding.txt'
   firstobs=2;
input time complete race poverty smoke alcohol age birthyear education
   prenatal_care;
run;

data bf; set bf;
label time ='Duration of breast feeding (weeks)';
run;



/* MARTINGALE RESIDUALS TO INVESTIGATE THE FUNCTIONAL FORM TO USE FOR
CONTINUOUS COVARIATES */

/* Fit the model without the covariate of interest. Since we don't know
which functional form to use for the continuous covariates yet, start with
including only the categorical ones*/
proc phreg data=bf noprint;
model time*complete(0)= smoke poverty /ties=exact;
output out = martingale resmart=mgale;
run;

/* Plot the Martingale residuals for mother's age at child birth */
proc loess data=martingale;
model mgale = age / smooth=0.6 direct;
label age ='Age of mother at birth (years)';
run;
/* "smooth=" gives the fraction of the data to be used as "nearby points".
Default for a model with at least 2 covariates: 0.5 */
/* "direct" fits a separate OLS for each local neighborhood. The default
option fits one OLS for a representative sample of points and interpolates
from that (less computationally demanding). */


/* Now including age, investigating education */
proc phreg data=bf noprint;
model time*complete(0)= smoke poverty age /ties=exact;
output out = martingale2 resmart=mgale;
run;
```

```
/* Plot the Martingale residuals for mother's education at child birth */
proc loess data=martingale2;
model mgale = education / smooth=0.7 direct;
label education ='Education of mother at birth (years)';
run;

/* Now including education, investigating birthyear */
proc phreg data=bf noprint;
model time*complete(0)= smoke poverty age education /ties=exact;
output out = martingale3 resmart=mgale;
run;

/* Birth year of the child */
proc loess data=martingale3;
model mgale = birthyear / smooth=0.6 direct;
label birthyear ='Birthyear of child';
run;

/* All numerical covariates can be included as they are, no transformations
are needed */


/* MULTICOLLINEARITY? */

proc corr data=bf spearman noprob;
var smoke poverty age education birthyear;
run;
/* Age is moderately strongly correlated with education and birthyear, as
expected */

data temp;
set bf;
fake=1;
run;

proc reg data=temp;
model fake=smoke poverty age education birthyear/vif;
run;
/* Highest VIF=2.25, which means that linear regression standard errors are
inflated only 1.5 times due to multicollinearity (sqrt(2.25)=1.5). How this
translates to Cox regression is yet uknown, we’ll use the VIF until more
research is available. Nothing to worry about */
/* CHECK THE PH ASSUMPTION */

/* Time-dependent covariate test */
proc phreg data=bf;
model time*complete(0)=smoke lntsmoke /ties=exact;
lntsmoke=log(time)*smoke;
run;

proc phreg data=bf;
model time*complete(0)=poverty lntPoverty /ties=exact;
lntPoverty=log(time)*poverty;
run;
```

```
proc phreg data=bf;
model time*complete(0)=age lntage /ties=exact;
lntage=log(time)*age;
run;

proc phreg data=bf;
model time*complete(0)=education lnteducation /ties=exact;
lnteducation=log(time)*education;
run;
/* Significant! */
/* Include a time-dependent covariate for education in the model? Or
stratify on categorized education? */

proc phreg data=bf;
model time*complete(0)=birthyear lntbirthyear /ties=exact;
lntbirthyear=log(time)*birthyear;
run;



/* Arjas plot */

/* Estimate the hazard without the covariate for which we want to check the
PH assumption. Education excluded below - to be included it in the way you
decide to handle the non-proportionality. */

/* Smoke */
proc phreg data=bf noprint;
model time*complete(0)= poverty age birthyear /ties=exact;
output out = hazarjas1 logsurv=ls;
run;

data hazarjas1; set hazarjas1;
cumhaz=-ls;
run;

proc sql noprint;
select count(time) into :t1-:t2         /* Number of 'times' is counted for
                    each category of the variable 'smoke', macro variables
                    are created containing those numbers */
   from bf          /* dataset to be used */
   group by smoke;
quit; /* run; not needed for proc sql, quit; shows the end of code */

%put _all_;
/* Prints all macro variables, including the values of t1 and t2 above -
check that the numbers are reasonable and which one is the largest (to be
used in the plot later on). T1 = 657, T2 = 270. */

proc sort data=hazarjas1;
by cumhaz;
run;
```

```
data arjas1; set hazarjas1;
retain n1 n2 h1 h2 c1 c2 0;
if cumhaz ne . then do;       /* h=logsurv (estimated above) */
   if smoke = 1 then do;      /* Calculations are being made separately for
          each stratum of the covariate to be checked */
   c1 = c1 + 1;     /* Starts at 0, increases by 1 for every observation in
          this stratum */
   n1 = n1 + complete; /* Counts the no. of events (complete=1 if event) */
   h1 = cumhaz + h1;    /* Sums the cumulative hazard up to each time */
   end;
else if smoke = 0 then do;
   c2 = c2 + 1;
   n2 = n2 + complete;
   h2 = cumhaz + h2;
   end;
end;
tot1 = h1 + cumhaz*(&t1-c1);  /* Total time on test within the stratum */
tot2 = h2 + cumhaz*(&t2-c2);
run;

proc sgplot data=arjas1 noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;
series x=n2 y=n2;
xaxis label='Number of Failures'  min=0 max=700; /* T1=657 to be included*/
yaxis label='Estimated Cumulative Hazard Rates'  min=0 max=700;
title 'Arjas plot - smoke';
run;

/* Clear non-linear deviations! */
/* Stratify on smoke? */


/* Poverty */
proc phreg data=bf noprint;
model time*complete(0)= age education birthyear /ties=exact;
output out = hazarjas2 logsurv=ls;
strata smoke;
run;

data hazarjas2; set hazarjas2;
  cumhaz=-ls;
run;

proc sql noprint;
select count(time) into :t1-:t2
   from bf
   group by poverty;
quit;

%put _all_;
/* T1 = 756, T2 = 171*/
```

```
proc sort data=hazarjas2;
by cumhaz;
run;

data arjas2; set hazarjas2;
retain n1 n2 h1 h2 c1 c2 0;
if cumhaz ne . then do;
   if poverty = 1 then do ;
   c1 = c1 + 1;
   n1 = n1 + complete;
   h1 = cumhaz + h1;
   end;
else if poverty = 0 then do;
   c2 = c2 + 1;
   n2 = n2 + complete;
   h2 = cumhaz + h2;
   end;
end;
tot1 = h1 + cumhaz*(&t1-c1);
tot2 = h2 + cumhaz*(&t2-c2);
run;

proc sgplot data=arjas2 noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;
series x=n2 y=n2;
xaxis label='Number of Failures'  min=0 max=800; /* T1=756 to be included*/
yaxis label='Estimated Cumulative Hazard Rates'  min=0 max=800;
title 'Arjas plot - poverty';
run;
/* Red line is straight, blue line shows some curvilinear departure but
perhaps only at the right end (where the estimates are not that
trustworthy) */


/* Continuous covariates have to be categorized to use the Arjas plot */

proc univariate data=bf noprint;
histogram age education birthyear;
run;

/* Arbitrary categorization of age */

data bf; set bf;
if age<=19 then agegr=1;
else if 19<age<=24 then agegr=2;
else if age>24 then agegr=3;
run;

proc phreg data=bf noprint;
model time*complete(0)= poverty education birthyear/ties=exact;
output out = hazarjas3 logsurv=ls;
strata smoke;
run;
```

```
data hazarjas3; set hazarjas3;
cumhaz=-ls;
run;

proc sql noprint;
select count(time) into :t1-:t3
from bf
group by agegr;
quit;

%put _all_;
/* T1 = 222, T2 = 561, T3 = 144 */

proc sort data=hazarjas3;
by cumhaz;
run;

data arjas3; set hazarjas3;
retain n1 n2 n3 h1 h2 h3 c1 c2 c3 0;
if cumhaz ne . then do;
   if agegr = 1 then do ;
      c1 = c1 + 1;
      n1 = n1 + complete;
      h1 = cumhaz + h1;
      end;
   else if agegr = 2 then do;
      c2 = c2 + 1;
      n2 = n2 + complete;
      h2 = cumhaz + h2;
      end;
   else if agegr = 3 then do;
      c3 = c3 + 1;
      n3 = n3 + complete;
      h3 = cumhaz + h3;
      end;
   end;
tot1 = h1 + cumhaz*(&t1-c1);
tot2 = h2 + cumhaz*(&t2-c2);
tot3 = h3 + cumhaz*(&t3-c3);
run;

proc sgplot data=arjas3 noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;
series x=n3 y=tot3;
series x=n2 y=n2;
xaxis label='Number of Failures' min=0 max=600; /* T2=561 to be included*/
yaxis label='Estimated Cumulative Hazard Rates' min=0 max=600;
title 'Arjas plot - age';
run;

/* All lines are straight and quite close to the 45 degree line - which is
an indication that age may be redundant in the model */
```

```
proc phreg data=bf;
model time*complete(0)= poverty education birthyear /ties=exact;
strata smoke;
run;
/* AIC = 5302.306 without age */

proc phreg data=bf;
model time*complete(0)= age poverty education birthyear /ties=exact;
strata smoke;
run;
/* AIC = 5303.280 with age, i.e. it increases which means that age is
unnecessary.
Age is also non-significant. Thus, age can be excluded from the model. */


/* Categorization of education */
data bf; set bf;
if 0<=education<=11 then educ=1;        *Not completed high school;
else if education=12 then educ=2;       *Completed high school but no
college/university;
else if education>=13 then educ=3;      *Some college/university education;
run;

proc phreg data=bf noprint;
model time*complete(0)= poverty birthyear /ties=exact;
output out = hazarjas4 logsurv=ls;
strata smoke;
run;

data hazarjas4; set hazarjas4;
cumhaz=-ls;
run;

proc sql noprint;
select count(time) into :t1-:t3
   from bf
   group by educ;
quit;

%put _all_;
/* T1 = 220, T2 = 438, T3 = 269 */


proc sort data=hazarjas4;
by cumhaz;
run;
```

```
data arjas4; set hazarjas4;
retain n1 n2 n3 h1 h2 h3 c1 c2 c3 0;
if cumhaz ne . then do;
   if educ = 1 then do ;
      c1 = c1 + 1;
      n1 = n1 + complete;
      h1 = cumhaz + h1;
      end;
   else if educ = 2 then do;
      c2 = c2 + 1;
      n2 = n2 + complete;
      h2 = cumhaz + h2;
      end;
   else if educ = 3 then do;
      c3 = c3 + 1;
      n3 = n3 + complete;
      h3 = cumhaz + h3;
      end;
   end;
tot1 = h1 + cumhaz*(&t1-c1);
tot2 = h2 + cumhaz*(&t2-c2);
tot3 = h3 + cumhaz*(&t3-c3);
run;

proc sgplot data=arjas4 noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;
series x=n3 y=tot3;
series x=n2 y=n2;
xaxis label='Number of Failures' min=0 max=500; /* T2=438 to be included*/
yaxis label='Estimated Cumulative Hazard Rates' min=0 max=500;
title 'Arjas plot - education';
run;

/* No curvilinear departure. But the time-dependent covariate test has
already indicated non-proportional hazards for education */


/* Arbirtary categorization of birthyear */
data bf; set bf;
if birthyear<=80 then birthygr=1;
else if 80<birthyear<=83 then birthygr=2;
else if birthyear>83 then birthygr=3;
run;

proc phreg data=bf noprint;
model time*complete(0)= poverty education /ties=exact;
output out = hazarjas5 logsurv=ls;
strata smoke;
run;

data hazarjas5; set hazarjas5;
cumhaz=-ls;
run;
```

```
proc sql noprint;
select count(time) into :t1-:t3
   from bf
   group by birthygr;
quit;

%put _all_;
/* T1 = 249, T2 = 414, T3 = 264 */

proc sort data=hazarjas5;
by cumhaz;
run;

data arjas5; set hazarjas5;
retain n1 n2 n3 h1 h2 h3 c1 c2 c3 0;
if cumhaz ne . then do;
   if birthygr = 1 then do ;
      c1 = c1 + 1;
      n1 = n1 + complete;
      h1 = cumhaz + h1;
      end;
   else if birthygr = 2 then do;
      c2 = c2 + 1;
      n2 = n2 + complete;
      h2 = cumhaz + h2;
      end;
   else if birthygr = 3 then do;
      c3 = c3 + 1;
      n3 = n3 + complete;
      h3 = cumhaz + h3;
      end;
   end;
tot1 = h1 + cumhaz*(&t1-c1);
tot2 = h2 + cumhaz*(&t2-c2);
tot3 = h3 + cumhaz*(&t3-c3);
run;


proc sgplot data=arjas5 noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;
series x=n3 y=tot3;
series x=n2 y=n2;
xaxis label='Number of Failures' min=0 max=500;  /* T2=414 to be included*/
yaxis label='Estimated Cumulative Hazard Rates' min=0 max=500;
title 'Arjas plot - birth year';
run;

/* No curvilinear departure */
/* INVESTIGATE MODEL FIT */

/* Cox-Snell residual plot */
proc phreg data=bf noprint;
model time*complete(0)= poverty education lnteducation
   birthyear/ties=exact;
lnteducation=log(time)*education;
output out = coxsnell LOGSURV = h;
strata smoke;
run;
/* 'WARNING: The OUTPUT data set has no observations due to the presence of
time-dependent explanatory variables.' */
/* Time-dependent covariates cannot be included. Ok to leave them out for
this part */

proc phreg data=bf noprint;
model time*complete(0)= poverty education birthyear/ties=exact;
output out = coxsnell LOGSURV = h;
strata smoke;
run;



data coxsnell; set coxsnell;
r=-h;     /* Cumulative hazards based on the Cox model, i.e. adjusting for
          covariate values */
run;


/*  Nelson-Aalen estimates based on r */
ods output ProductLimitEstimates=figure;
proc lifetest data=coxsnell nelson plots=none;
time r*complete(0);
run;

proc sort data=figure;
by r;
run;

proc sgplot data=figure;
step y=cumhaz x=r;
series y=r x=r;
xaxis label = "Residual";
yaxis label = "Estimated Cumulative Hazard Rates";
title "Cox-Snell residuals to check the overall fit of the model";
run;

/* Model seems to fit the data pretty well */
/* Generalized R-squared */
ods output FitStatistics=fit;
proc phreg data=bf;
model time*complete(0)= poverty education lnteducation
   birthyear/ties=exact;
lnteducation=log(time)*education;
strata smoke;
run;

data r2; set fit;
where criterion='-2 LOG L';
LRT=WithoutCovariates-WithCovariates;
R2=1-exp(-LRT/927);
run;

proc print data=r2;
var R2;
run;

/* -2LL without covariates = 5320.374, with covariates = 5286.401 */
/* Sample size in each strata: smoke 657, non-smoke 270 */

/* Generalized R2:
          LRT = 5320.374 -5286.401 = 33.973
          R2  = 1-exp(-LRT/n)= 1-exp(-33.973/927)  = 0.04
Thus, the covariates are very weakly associated with the duration of breast
feeding */


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