

```

/* input data */
proc import
  dbms=csv
  datafile='/home/u64388585/LungCancer/data/Sweden_Lung_Cancer_500.csv'
  out=lungdata
  replace;
run;

/* Assignment 2*/

/*
class: gender cancer_stage family_history smoking_status hypertension asthma cirrhosis other_cancer treatment_t;
*/

data lungdata2;
  set lungdata;
  /* =====
     Cancer stage (reference = stage 0)
     ===== */
  cancer_stagel = (cancer_stage = 1);
  cancer_stage2 = (cancer_stage = 2);
  cancer_stage3 = (cancer_stage = 3);

  /* =====
     Treatment type (reference = type 0)
     ===== */
  treatment_type1 = (treatment_type = 1);
  treatment_type2 = (treatment_type = 2);
  treatment_type3 = (treatment_type = 3);

  /* =====
     Smoking status (reference = status 0)
     ===== */
  smoking_status1 = (smoking_status = 1);
  smoking_status2 = (smoking_status = 2);
  smoking_status3 = (smoking_status = 3);
run;

/* Step 3*/
/* Test the assumption of PH for each covariate*/
/* Test all: age gender cancer_stage family_history smoking_status bmi cholesterol_level hypertension asthma ci:
proc phreg data=lungdata2;
model treatment_days*survived(1)= age lnt_age/ties=exact;
lnt_age=log(treatment_days)*age;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= gender lnt_gender/ties=exact;
lnt_gender=log(treatment_days)*gender;
run;
proc phreg data=lungdata2; /*cancer_stage Not satisfied*/
model treatment_days*survived(1)= cancer_stagel cancer_stage2 cancer_stage3
  lnt_cancer_stagel lnt_cancer_stage2 lnt_cancer_stage3 /ties=exact;
lnt_cancer_stagel=log(treatment_days)*cancer_stagel;
lnt_cancer_stage2=log(treatment_days)*cancer_stage2;
lnt_cancer_stage3=log(treatment_days)*cancer_stage3;
test lnt_cancer_stagel=lnt_cancer_stage2=lnt_cancer_stage3=0;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= family_history lnt_family_history/ties=exact;
lnt_family_history=log(treatment_days)*family_history;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= smoking_status1 smoking_status2 smoking_status3
  lnt_smoking_status1 lnt_smoking_status2 lnt_smoking_status3/ties=exact;
lnt_smoking_status1=log(treatment_days)*smoking_status1;
lnt_smoking_status2=log(treatment_days)*smoking_status2;
lnt_smoking_status3=log(treatment_days)*smoking_status3;
test lnt_smoking_status1=lnt_smoking_status2=lnt_smoking_status3=0;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= bmi lnt_bmi/ties=exact;
lnt_bmi=log(treatment_days)*bmi;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= cholesterol_level lnt_cholesterol_level/ties=exact;
lnt_cholesterol_level=log(treatment_days)*cholesterol_level;
run;
proc phreg data=lungdata2; /*hypertension Not satisfied*/
model treatment_days*survived(1)= hypertension lnt_hypertension/ties=exact;
lnt_hypertension=log(treatment_days)*hypertension;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= asthma lnt_asthma/ties=exact;
lnt_asthma=log(treatment_days)*asthma;
run;

```

```

proc phreg data=lungdata2;
model treatment_days*survived(1)= cirrhosis lnt_cirrhosis/ties=exact;
lnt_cirrhosis=log(treatment_days)*cirrhosis;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= other_cancer lnt_other_cancer/ties=exact;
lnt_other_cancer=log(treatment_days)*other_cancer;
run;
proc phreg data=lungdata2;
model treatment_days*survived(1)= treatment_type1 treatment_type2 treatment_type3
lnt_treatment_type1 lnt_treatment_type2 lnt_treatment_type3/ties=exact;
lnt_treatment_type1=log(treatment_days)*treatment_type1;
lnt_treatment_type2=log(treatment_days)*treatment_type2;
lnt_treatment_type3=log(treatment_days)*treatment_type3;
test lnt_treatment_type1=lnt_treatment_type2=lnt_treatment_type3=0;
run;

/*cancer_stage, hypertension Not satisfied PH*/

/* step 4*/
/* Investigate which functional form to use for continuous covariates */

/* 1) Fit Cox model WITHOUT bmi (covariate of interest) */
proc phreg data=lungdata2 noprint;
model treatment_days*survived(1) =
/* continuous */
age cholesterol_level

/* binary */
gender family_history hypertension asthma cirrhosis other_cancer

/* multi-category coded as dummies (reference = 0) */
cancer_stage1 cancer_stage2 cancer_stage3
smoking_status1 smoking_status2 smoking_status3
treatment_type1 treatment_type2 treatment_type3
/ ties=exact;

output out=martingale resmart=mgale;
run;
/* 2) LOESS smooth: mgale vs bmi */
proc loess data=martingale;
model mgale = bmi / smooth=0.7 direct;
ods output OutputStatistics=martplot;
run;

/* 1) Fit Cox model WITHOUT cholesterol_level (covariate of interest) */
proc phreg data=lungdata2 noprint;
model treatment_days*survived(1) =
/* continuous */
age bmi

/* binary */
gender family_history hypertension asthma cirrhosis other_cancer

/* multi-category coded as dummies (reference = 0) */
cancer_stage1 cancer_stage2 cancer_stage3
smoking_status1 smoking_status2 smoking_status3
treatment_type1 treatment_type2 treatment_type3
/ ties=exact;

output out=martingale resmart=mgale;
run;
/* 2) LOESS smooth: mgale vs cholesterol_level */
proc loess data=martingale;
model mgale = cholesterol_level / smooth=0.7 direct;
ods output OutputStatistics=martplot;
run;

/* 1) Fit Cox model WITHOUT cholesterol_level (covariate of interest) */
proc phreg data=lungdata2 noprint;
model treatment_days*survived(1) =
/* continuous */
bmi cholesterol_level

/* binary */
gender family_history hypertension asthma cirrhosis other_cancer

/* multi-category coded as dummies (reference = 0) */
cancer_stage1 cancer_stage2 cancer_stage3
smoking_status1 smoking_status2 smoking_status3
treatment_type1 treatment_type2 treatment_type3
/ ties=exact;

output out=martingale resmart=mgale;

```

```

run;
/* 2) LOESS smooth: mgale vs age */
proc loess data=martingale;
  model mgale = age / smooth=0.7 direct;
  ods output OutputStatistics=martplot;
run;

/*all of continuous variables don't need to transform*/

/* step 5*/
/*fit the best model
notice!! cancer_stage and hypertension violate PH assumption
*/

/*
L6, p55: Check stratified model assumption
H0: The covariate effects are the same across strata.
*/

/*
stratify on "cancer_stage" & set "hypertension" as time dependent covariate
*/
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age bmi cholesterol_level
    /* binary */
    gender family_history
    hypertension lnt_hypertension /*time dependent*/
    asthma cirrhosis other_cancer
    /* multi-category coded as dummies (reference = 0) */
    /*cancer_stage1 cancer_stage2 cancer_stage3*/
    smoking_status1 smoking_status2 smoking_status3
    treatment_type1 treatment_type2 treatment_type3
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
  strata cancer_stage ;
run;

proc sort data=lungdata2;
by cancer_stage;
run;

proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age bmi cholesterol_level
    /* binary */
    gender family_history
    hypertension lnt_hypertension /*time dependent*/
    asthma cirrhosis other_cancer
    /* multi-category coded as dummies (reference = 0) */
    /*cancer_stage1 cancer_stage2 cancer_stage3*/
    smoking_status1 smoking_status2 smoking_status3
    treatment_type1 treatment_type2 treatment_type3
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
  by cancer_stage ;
run;
/*
Stratified by cancer (-2LogL):2898.296
cancer stage 0 (-2LogL) = 679.825
cancer stage 1 (-2LogL) = 592.524
cancer stage 2 (-2LogL) = 809.415
cancer stage 3 (-2LogL) = 771.962

X2 = 2898.296-(679.825+592.524+809.415+771.962)=44.56999999999971
df = (4-1)*16 = 48
p-value=0.62: A stratified model is appropriate.
*/

/*
stratify on "hypertension" & set "cancer_stage" as time dependent covariate
*/
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age bmi cholesterol_level
    /* binary */
    gender family_history
    asthma cirrhosis other_cancer
    /* cancer stage: main + time-dependent */
    cancer_stage1 cancer_stage2 cancer_stage3
    lnt_stage1 lnt_stage2 lnt_stage3
    /* multi-category coded as dummies (reference = 0) */

```

```

        smoking_status1 smoking_status2 smoking_status3
        treatment_type1 treatment_type2 treatment_type3
    / ties=exact;
    lnt_stage1 = cancer_stage1 * log(treatment_days);
    lnt_stage2 = cancer_stage2 * log(treatment_days);
    lnt_stage3 = cancer_stage3 * log(treatment_days);
    strata hypertension ;
run;

proc sort data=lungdata2;
by hypertension;
run;

proc phreg data=lungdata2;
    model treatment_days*survived(1) =
        /* continuous */
        age bmi cholesterol_level
        /* binary */
        gender family_history
        asthma cirrhosis other_cancer
        /* cancer stage: main + time-dependent */
        cancer_stage1 cancer_stage2 cancer_stage3
        lnt_stage1 lnt_stage2 lnt_stage3
        /* multi-category coded as dummies (reference = 0) */
        smoking_status1 smoking_status2 smoking_status3
        treatment_type1 treatment_type2 treatment_type3
    / ties=exact;
    lnt_stage1 = cancer_stage1 * log(treatment_days);
    lnt_stage2 = cancer_stage2 * log(treatment_days);
    lnt_stage3 = cancer_stage3 * log(treatment_days);
    by hypertension;
run;
/*
Likelihood ratio chi square test
Stratified by hypertension (-2LogL):3435.912
hypertension 0 (-2LogL) = 597.204
hypertension 1 (-2LogL) = 2808.804

X2 = 3435.912-(597.204+2808.804)=29.903
df = (2-1)*20 = 20
p-value=~0.07: A stratified model is appropriate
*/

/*
Decide to use:
stratify on "cancer_stage" & set "hypertension" as time dependent covariate
since the p-value is larger (0.62)
*/

/*fit the best model*/
/*set backward selection criteria 0.3*/
proc phreg data=lungdata2;
    model treatment_days*survived(1) =
        /* continuous */
        age bmi cholesterol_level
        /* binary */
        gender family_history
        asthma cirrhosis other_cancer
        /* multi-category coded as dummies (reference = 0) */
        smoking_status1 smoking_status2 smoking_status3
        treatment_type1 treatment_type2 treatment_type3

        /*time dependent*/
        hypertension lnt_hypertension
    / ties=exact
    selection=backward slstay=0.3;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
    strata cancer_stage;
run;
/*
keep: age, cirrhosis, treatment_type3, hypertension, lnt_hypertension*/

/*fit again the best model*/
proc phreg data=lungdata2;
    model treatment_days*survived(1) =
        /* continuous */
        age
        /* binary */
        cirrhosis
        /* multi-category coded as dummies (reference = 0) */
        treatment_type1 treatment_type2 treatment_type3
        /*time dependent*/
        hypertension lnt_hypertension
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
    strata cancer_stage;
run;
/*-2logL=2902.128*/

```

```

/*check possible interaction*/
/*treatment * Age*/
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age
    /* binary */
    cirrhosis
    /* multi-category coded as dummies (reference = 0) */
    treatment_type1 treatment_type2 treatment_type3
    /*time dependent*/
    hypertension lnt_hypertension

    /* interaction terms */
    age*treatment_type1
    age*treatment_type2
    age*treatment_type3
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
  strata cancer_stage;
run;
/*age*treatment_type2, age*treatment_type3 pvalue<0.05*/
/*-2logL=2894.452*/

/*
X2=2902.128-2894.452=7.676
df=3
p-value=0.053

The data are insufficient to support the interaction of age * treatment, so use the model without interaction e:
*/

/*treatment * Hypertension*/
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age
    /* binary */
    cirrhosis
    /* multi-category coded as dummies (reference = 0) */
    treatment_type1 treatment_type2 treatment_type3
    /*time dependent*/
    hypertension lnt_hypertension

    /* interaction terms */
    hypertension*treatment_type1
    hypertension*treatment_type2
    hypertension*treatment_type3
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
  strata cancer_stage;
run;
/*no interaction effect*/

/*Our best model*/
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age
    /* binary */
    cirrhosis
    /* multi-category coded as dummies (reference = 0) */
    treatment_type1 treatment_type2 treatment_type3
    /*time dependent*/
    hypertension lnt_hypertension
    / ties=exact;
    lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time dependent covariate*/
  strata cancer_stage;
run;

/* step 6. */
/* Test the assumption of PH again (by graph)*/
/*
continuous: age
binary: cirrhosis
multi-category coded as dummies (reference = 0): treatment_type1 treatment_type2 treatment_type3
time dependent: hypertension lnt_hypertension
*/

/*

```

```

ARJAS PLOT
Checks PH for a BINARY covariate myvar (0/1)
===== */

%let mydata = lungdata2;
%let mytime = treatment_days;
%let mycens = survived;
%let censval = 1;

/* -----
Step 1) Fit Cox model WITHOUT the covariate of interest (&myvar)
----- */

/* Test cirrhosis */
%let myvar = cirrhosis;

/*-----*/
/* Wrong ?? can't we put time-dependent variable?? */
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    age
    treatment_type1 treatment_type2 treatment_type3
    hypertension lnt_hypertension
    / ties=exact;
  lnt_hypertension = hypertension * log(treatment_days);
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;
/*-----*/

/* Test cirrhosis*/
%let myvar = cirrhosis;
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    age
    treatment_type1 treatment_type2 treatment_type3
    hypertension

    / ties=exact;
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;

/*Test treatment 1*/
%let myvar = treatment_type1;
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    age
    cirrhosis
    treatment_type2 treatment_type3
    hypertension

    / ties=exact;
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;

/*Test treatment 2*/
%let myvar = treatment_type2;
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    age
    cirrhosis
    treatment_type1 treatment_type3
    hypertension

    / ties=exact;
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;

/*Test treatment 3*/
%let myvar = treatment_type3;
proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    age
    cirrhosis
    treatment_type1 treatment_type2
    hypertension

    / ties=exact;
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;

```

```

/*Test age*/
data lungdata2_age;
set lungdata2;
if age < 55 /*median*/ then age_old=1;
else age_old=0;
run;

%let mydata = lungdata2_age;
%let myvar = age_old;
proc phreg data=lungdata2_age;
  model treatment_days*survived(1) =
    cirrhosis
    treatment_type1 treatment_type2 treatment_type3
    hypertension
    / ties=exact;
  strata cancer_stage;
  output out=hazarjas logsurv=ls;
run;

/* -----
Step 2) Compute cumulative hazard from log-survival
----- */
data hazarjas;
set hazarjas;
cumhaz = -ls; /*cumhaz = -log(S(t)) */
run;

/* -----
Step 3) Count sample size in each category of &myvar (0 and 1)
Creates macro vars: t1 (for myvar=0), t2 (for myvar=1)
----- */
proc sql noprint;
select count(&mytime) into :t1-:t2
from &mydata
group by &myvar;
quit;

%put NOTE: t1 (myvar=0)=&t1 t2 (myvar=1)=&t2;

/* -----
Step 4) Sort by cumhaz (needed for Arjas calculations)
----- */
proc sort data=hazarjas;
by cumhaz;
run;

/* -----
Step 5) Build Arjas quantities (expected no. of events / TOT)
----- */
data arjas;
set hazarjas;

myevent = 1 - (&mycens=&censval);

retain n1 n2 h1 h2 c1 c2 0;

if cumhaz ne . then do;

  if &myvar = 1 then do;
    c1 = c1 + 1;
    n1 = n1 + myevent;
    h1 = cumhaz + h1;
  end;
else if &myvar = 0 then do;
    c2 = c2 + 1;
    n2 = n2 + myevent;
    h2 = cumhaz + h2;
  end;

end;

tot1 = h1 + cumhaz*(&t1-c1);
tot2 = h2 + cumhaz*(&t2-c2);
run;

/* -----
Step 6) Plot
NOTE: 45-degree line uses the stratum with larger max n.
You should adjust max= based on your output (largest n1/n2).
----- */
proc sgplot data=arjas noautolegend;
series x=n1 y=tot1;
series x=n2 y=tot2;

series x=n2 y=n2; /* 45-degree line (can swap to n1 if needed) */

```

```

axis label='Number of events' min=0 max=800;
yaxis label='Estimated Cumulative Hazard Rates' min=0 max=800;

title "Arjas plot for &myvar (PH check)";
run;

/*
cirrhosis Arjas plot looks very terrible
treatment1, 2, 3 Arjas plot looks very terrible
age_old also not looks good
*/

/* step 7. INVESTIGATE MODEL FIT*/
/* -----
Step 7A) Cox-Snell residuals (exclude time-dependent covariates)
Final model has time-dependent lnt_hypertension, so omit it here.
Keep the same strata(cancer_stage).
----- */

proc phreg data=lungdata2 noprint;
  model treatment_days*survived(1) =
    age
    cirrhosis
    treatment_type1 treatment_type2 treatment_type3
    /* omit hypertension + lnt_hypertension for Cox-Snell part */
    / ties=exact;
  strata cancer_stage;
  output out=coxsnell logsurv=ls; /* ls = log(S(t|x)) */
run;

/* Cox-Snell residual r = -log(S(t|x)) = -ls */
data coxsnell;
  set coxsnell;
  r = -ls;
run;

/* Nelson-Aalen estimate of cumulative hazard for r
If model fits well: H_hat(r) should be ~ r (45-degree line)
*/
ods output ProductLimitEstimates=cs_plot;
proc lifetest data=coxsnell nelson plots=none;
  time r*survived(1);
run;

proc sort data=cs_plot;
  by r;
run;

proc sgplot data=cs_plot;
  step x=r y=cumhaz;
  series x=r y=r;
  xaxis label="Cox-Snell residual r = -log(S(t|x))";
  yaxis label="Nelson-Aalen estimated cumulative hazard";
  title "Cox-Snell residual plot (time-dependent covariates excluded)";
run;
/*looks like the model fits the data well*/

/* -----
Step 7B) Generalized R^2 for the FINAL model
(including time-dependent covariates)
----- */

/* Generalized R-squared */
ods output FitStatistics=fit;

proc phreg data=lungdata2;
  model treatment_days*survived(1) =
    /* continuous */
    age
    /* binary */
    cirrhosis
    /* multi-category coded as dummies (reference = 0) */
    treatment_type1 treatment_type2 treatment_type3
    /*time dependent*/
    hypertension lnt_hypertension
    / ties=exact;
  lnt_hypertension = hypertension * log(treatment_days); /*set hypertension as time depeident covariate*/
  strata cancer_stage;
run;

data r2; set fit;
where criterion='-2 LOG L';
LRT=WithoutCovariates-WithCovariates;
R2=1-exp(-LRT/500);

```

```
run;
```

```
proc print data=r2;
```

```
var R2;
```

```
run;
```

```
/* -2LL without covariates = 2921.789, with covariates = 2902.128 */
```

```
/* Sample size in each strata: stage0: 120, stage1: 110, stage2: 138, stage3:132 */
```

```
/* Generalized R2:
```

```
LRT = 2921.789 - 2902.128 = 19.661
```

```
R2 = 1-exp(-LRT/n)= 1-exp(-19.661/500) = 0.0385
```

```
Thus, the covariates are very weakly associated with lung cancer */
```