Part 3. 生存時間解析を使った可視化具体例 Data

1 Package

本パートでは,解析(R)の Part 3:と同様に causaldata package に格納されている National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS) の nhefs データセットを用います.データの取得やライブラリ指定などは(SAS)Part 2 を参照してください.

2 Data source

前パートでは欠測データの可視化をするために,nhefs データセットを使いましたが,本パートでは解析をシンプルにするために,欠測データのある症例を取り除き,変数を seqn, death, yrdth, modth, dadth, sex, age, race, education, exercise, smokeintensity, smokeyrs に絞った,nhefs_complete データセットを作成して利用します.(以下の変数に欠測のあるものを取り除いているようだったので,条件はそれに倣いました.➡ qsmk, sex, race, age, school, smokeintensity, smokeyrs, exercise, active, wt71, wt82, and wt82_71)

```
data nhefs_complete;
 set lib1.nhefs;
 where nmiss(qsmk, sex, race, age, school, smokeintensity, smokeyrs, exercise, active, wt71, wt82, wt82_71)=0;
 keep seqn death yrdth modth dadth sex age race education exercise smokeintensity smokeyrs pack_years_n;
run;
```

また,nhefs データ中のカテゴリカルデータは数値として格納されています.データセット仕様書見れば内容は把握できるのですが,確認しやすくするためにフォーマットを作成しておきます.

```
proc format;
value sexf
        0     = "Male"
        1     = "Female"
;
value racef
        0     = "White"
        1     = "Black or other"
;
value educationf
```

```
        1     = "8th grade or less"
        2     = "HS dropout"
        3     = "HS"
        4     = "College dropout"
        5     = "College or more"
;
value exercisef
        0     = "Much exercise"
        1     = "Moderate exercise"
        2     = "LIttle or no excercise"
;
value pack_years
        0="Low"
        1="High"
        99="All "

;
run;
```

3. Question

本パートでは,喫煙の程度を示す Pack-years が死亡に影響するか評価します.
これは「Pack-year が死亡という結果の原因になっているのか」という問いであり,問いの型は因果推論になります.

4. Data

4.1 Outcome

本パートのアウトカム変数は死亡です. ただし,生存・死亡という二値変数ではなく,観察を開始してからイベントを起こした時間を解析に用います. このようなデータを Time-to-event データといい,時間とイベントを組み合わせたデータで定義します.NHEFS の研究では,1983 年 1 月 1 日がコホートの観察開始日で,1992 年 12 月 31 日が観察終了日です.これらの情報と

death: 1992 年までに死亡したかどうかを示す変数
yrdth: 死亡した年
modth: 死亡した月
dadth: 死亡した日
以上の情報を組み合わせて,Time-to-event データを作ります

```
data nhefs_complete1;
 set nhefs_complete;
/*   死亡日または打ち切り日作成*/
   if death = 1 then event_date =MDY(modth, dadth, yrdth);
```

```
   else if    death = 0 then event_date =input("1992-12-31",yymmdd10.) ;
/*観察開始日からの日数を計算*/
   survtime = event_date -input("1983-01-01",yymmdd10.);
 format event_date yymmdd10. sex sexf. race racef. education educationf. exercise exercisef.;
run;
```

4.2 Exposure

本パートの曝露変数は Pack-years です.Pack-years は喫煙の程度を示す国際的な指標で,次式で定義されます.

$$\text{Pack-years} = \frac{\text{1日の喫煙本数}}{\text{20本}} \times \text{喫煙年数}$$

この値は連続尺度データです.（再び）,説明を簡単にするために，二値変数に変換します.本パートでは 20（1 年間に 20 箱）をカットオフ値とします

```
data nhefs_complete2;
set nhefs_complete1;
 pack_years_n = (smokeintensity / 20) * smokeyrs;
 if . < pack_years_n < 20 then pack_years= 0;
 else if 20 <= pack_years_n then    pack_years= 1;
 format pack_years pack_yearsf. ;
run;
```

4.3 練習のためのデータ変更

さて,nhefs_complete データはすべての人が研究期間終了までフォローできているデータです.しかし実際,みなさんが扱うデータは研究期間終了まで研究対象者をフォローすることは途中で転院等の事象が発生するため困難です.このとき,研究期間中の打ち切りが生じます.このような打ち切りがある場合,後述する生存時間解析で作成する図表には工夫が必要になります.練習のために,研究期間中の打ち切りを発生させたデータセットを作り,今後使用します.　R のコードの処理にあわせています.しかし SAS の場合,call streaminit()で乱数シードを指定していますが,乱数で値を加工している部分はプログラム言語によって同じ結果にならないことが通常なので,以降の処理は R の結果と数値一致はしません.結果の方向性が変わらないことのみ確認しています.

```
data nhefs_complete3;
set nhefs_complete2;
    call streaminit(1234);
    death1=rand("binominal",0.6,1);
```

```
    death0=rand("binominal",0.4,1);
    death2 = ifn(pack_years =1, death1, death0);
    censor=rand("binominal",0.2,1);
    r_time=ifn((censor=1 and death=0) or (death2=1 and death=0)
                    ,rand("INTEGER",0,3652),0 );
  death = ifn(death2 = 1, 1, death);
  survtime = survtime - r_time;
  survtime_y = survtime / 365.25;
label survtime_y="Years";
run;
```

4.5 本パートで仮定した因果構造

前節で定義した,アウトカムを死亡（death）,曝露を喫煙の程度（pack_years）とします.喫煙の程度が死亡に影響するか評価するときに,単純に二値化した Pack-years ごとの死亡の程度を比較することができますが,これでは影響の評価として十分ではありません.Pack-years の高い群と低い群で背景情報が異なり,交絡が生じる可能性があるからです.

交絡を軽減するためには,交絡を生み出す要因である交絡因子を調整する必要があります.交絡因子は,曝露にもアウトカムにも影響を与える要因です.変数間の関係,つまり因果構造を頭の中で考えることもできますが,一定のルールのもと可視化するのが便利です.この因果構造を可視化するツールを DAG (Directed Acyclic Graph: 有向非巡回グラフ)といいます.DAG は,ある変数が別の変数に影響を与えるかどうかを矢印で表します.矢印は常に一方通行です.

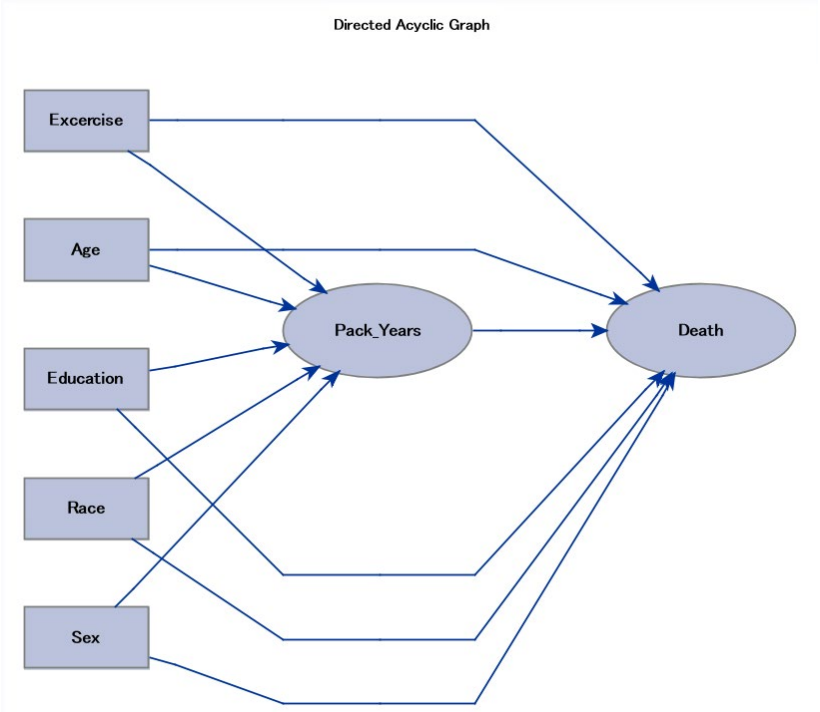
紙にざっと書いてしまっても良いのですが,せっかくなんで SAS で描いてみましょう！！…　といいたいところなのですが,SAS で R の ggdag パッケージのように因果構造と,ちょっとしたオプション指定で,いい感じに DAG を書いてくれる機能は,通常の SAS BASE/STAT の範囲では難しいです.CAUSALGRAPH プロシジャなど,いかにも描けそうな名前ですがそういった機能ではありません（変数間の因果構造から,興味のある因果効果を識別したり,推定するための変数集合を導出したりする）

座標軸を適当に線やマーカーで無理やり SGPLOT で一筆書きでグラフを描いてしまってもいいのですが,それも本末転倒な気がします.

構造方程式モデリングを扱う CALIS プロシジャで大まかの構造指定プロットデザインを作って,それを ODS Graphic Editer という SAS のグラフを手で調整可能な機能で,調整して,それらしいものを作ってみます

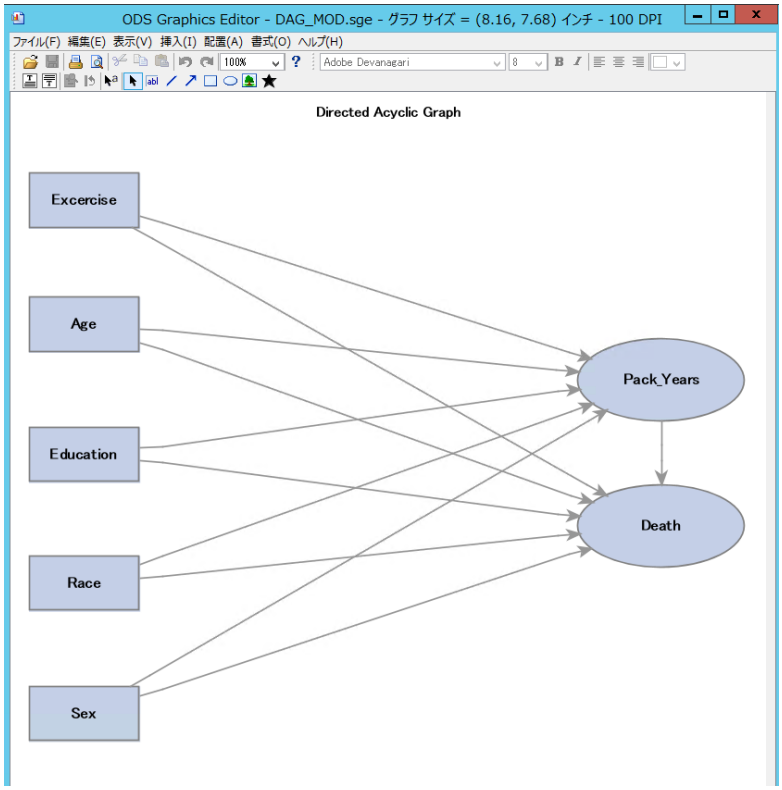
データセットの中で交絡因子になる得る変数は,sex, age, race, education, exercise です.これらが,death と pack_years にどのように影響するか可視化してみましょう

```
data dummy;
 array ar{*} Sex Age Race Excercise Education Excercise ;
 do i=1 to 8;
  do j= 1 to dim(ar);
    ar[j]=rand("uniform");
  end;
  output;
 end;
run;

ods listing gpath="出力フォルダのパス" sge=on;
ods graphics/reset   width=8.5in height=8in imagename="DAG" noborder;
proc calis data=dummy;
  path
    Sex Age Race Excercise Education   ---> Pack_Years,
    Sex Age Race Excercise Education   ---> Death,
    Pack_Years ---> Death;
pathdiagram method=flow nofittable noflag noestim novariance noerrvar title="Directed Acyclic Graph";
run;
```



位置関係の調整が難しく,見にくい配置になることがあるので

SAS ODS Graphics Editer という機能を使ってグラフ要素を手で編集可能な SGE 形式で作成し,必要に応じて編集します.



5 統計解析

5.1 背景情報の要約

研究対象集団の情報（特徴）を要約することは重要です.この情報をもとに,研究者は自身が想定している集団であるか,読者は目の前の患者さん等に結果を適用できるか検討できます.本パートでは,Pack-years の区分ごとに sex, age, race, education, exercise の情報を要約します.論文では Table 1 として載せることが多いです

Characteristic	Category Statics	Overall, N = 1566	Low, N = 744	High, N = 822	p-value
Age	N	1566	744	822	<.0001
	Mean	43.7	38.2	48.6	
	Std	12.0	11.5	10.1	
	Min	25	25	25	
	Median	43.0	34.0	48.0	
	Max	74	74	74	
Education	8th grade or less	291 (18.6%)	105 (14.1%)	186 (22.6%)	<.0001
	HS dropout	340 (21.7%)	133 (17.9%)	207 (25.2%)	
	HS	637 (40.7%)	331 (44.5%)	306 (37.2%)	
	College dropout	121 (7.7%)	70 (9.4%)	51 (6.2%)	
	College or more	177 (11.3%)	105 (14.1%)	72 (8.8%)	
Exercise	Much exercise	300 (19.2%)	160 (21.5%)	140 (17%)	0.0755
	Moderate exercise	661 (42.2%)	308 (41.4%)	353 (42.9%)	
	LIttle or no excercise	605 (38.6%)	276 (37.1%)	329 (40%)	
Race	White	1360 (86.8%)	602 (80.9%)	758 (92.2%)	<.0001
	Black or other	206 (13.2%)	142 (19.1%)	64 (7.8%)	
Sex	Male	762 (48.7%)	282 (37.9%)	480 (58.4%)	<.0001
	Female	804 (51.3%)	462 (62.1%)	342 (41.6%)	

集計コードは少し長くなりますが,簡単に書き下ろしました

```
data nhefs_complete4;
set nhefs_complete3
        nhefs_complete3(in=ina);
if ina then pack_years=99;
run;

proc sql noprint;
  select count(*) into: bign1 from nhefs_complete4 where    pack_years=0;
  select count(*) into: bign2 from nhefs_complete4 where    pack_years=1;
  select count(*) into: bign99 from nhefs_complete4 where    pack_years=99;
quit;

data classds;
format pack_years    pack_yearsf. sex sexf. race racef. education educationf.    exercise exercisef.    pack_years pack_yearsf.;
  do    pack_years=0, 1 , 99;
                do sex =0 to 1;
                output;
    end;
    call missing(of sex);
                do race =0 to 1;
                output;
    end;
    call missing(of race);
    do education = 1 to 5;
                output;
    end;
    call missing(of education);
    do exercise = 0 to 2;
                output;
    end;
    call missing(of exercise);
```

```
    do pack_years= 0,1,99;
                output;
   end;
  end;
run;

%macro frq(var=);
proc summary data=nhefs_complete4 classdata=classds nway;
  class pack_years &var.;
  output out=out_&var.;
run;
proc sort data=out_&var.;
  by &var.;
run;
proc transpose data=out_&var. out=t_out_&var.;
  var _FREQ_;
  by &var.;
  id    pack_years;
run;
ods output chisq=chisq_&var.(where=(Statistic="カイ 2 乗値"));
proc freq data=out_&var.;
where pack_years in (0,1);
  tables pack_years * &var./chisq;
  weight _FREQ_;
run;

data fix_t_out_&var.;
length label $50.;
  set t_out_&var.;
  if _N_=1 then do;
                set chisq_&var.;
                label="&var";
                p=put(Prob,pvalue6.4 -L);
  end;
  stat=vvalue(&var);
  out_LOW=catx (" ",LOW,cats("(",round(divide(LOW,&bign1)*100,0.1),"%)"));
  out_HIGH=catx (" ",HIGH,cats("(",round(divide(HIGH,&bign2)*100,0.1),"%)"));
  out_ALL=catx (" ",ALL,cats("(",round(divide(ALL,&bign99)*100,0.1),"%)"));
  keep out_: p label    stat;
run;
%mend;
%frq(var=Sex)
%frq(var=Race)
%frq(var=Education)
%frq(var=Exercise)


%macro summary(var=Age);
proc summary data=nhefs_complete4 classdata=classds nway;
  class pack_years;
  var &var;
  output out=out_&var    n=N mean=Mean std=Std min=Min median=Median max=Max ;
run;
proc transpose data=out_&var out=t_out_&var;
  var n -- max;
  id    pack_years;
run;
ods output WilcoxonTest=WilcoxonTest_&var;
proc npar1way data=nhefs_complete4 wilcoxon ;
```

```
where pack_years in (0,1);
class pack_years;
  var age;
run;

data fix_t_out_&var;
length label $50.;
  set t_out_&var;
  if _N_=1 then do;
                set WilcoxonTest_&var;
                label="&var";
                p=put(tProb2,pvalue6.4 -L);
  end;
  stat=_NAME_;
  if _NAME_ in ("Mean","Std","Median") then do;
  out_LOW=put(round(Low,0.1),8.1 -L);
  out_HIGH=put(round(HIGH,0.1),8.1 -L);
  out_ALL=put(round(ALL,0.1),8.1 -L);
end;
else do;
  out_LOW=cats(Low);
  out_HIGH=cats(HIGH);
  out_ALL=cats(ALL);
end;
keep out_: p label stat ;
run;
%mend;
%summary(var=Age)

data output;
length label stat out_LOW out_HIGH out_all p $200.;
set fix_t_out_: ;
run;

proc odstable data= output;
      column label stat OUT_ALL out_LOW out_HIGH    p ;

** header;
   define header header1; start=label; end=label; vjust=center; just=left; text "Characteristic"; end;
   define header header2; start=stat; end=stat; vjust=center; just=left ; split="#";text "Category#Statics"; end;
   define header header3; start=out_all; end=out_all; vjust=center; just=center; split="#"; text "Overall,#N =
&bign99.";end;
   define header header4; start=out_low; end=out_low; vjust=center; just=center; split="#"; text "Low,#N = &bign1.";end;
   define header header5; start=out_high end=out_high; vjust=center; just=center; split="#"; text "High,#N =
&bign2.";end;
   define header header6; start=p; end=p; vjust=center; just=center; text "p-value"; end;

** column;
   define label; print_headers=off; just=left; style={cellwidth=80}; end;
   define stat; print_headers=off; just=left; style={cellwidth=200}; end;
   define out_all; print_headers=off; just=center; style={cellwidth=80}; end;
   define out_low; print_headers=off; just=center; style={cellwidth=80}; end;
   define out_high; print_headers=off; just=center; style={cellwidth=80}; end;
   define p; print_headers=off; just=center; style={cellwidth=80}; end;

run;
```

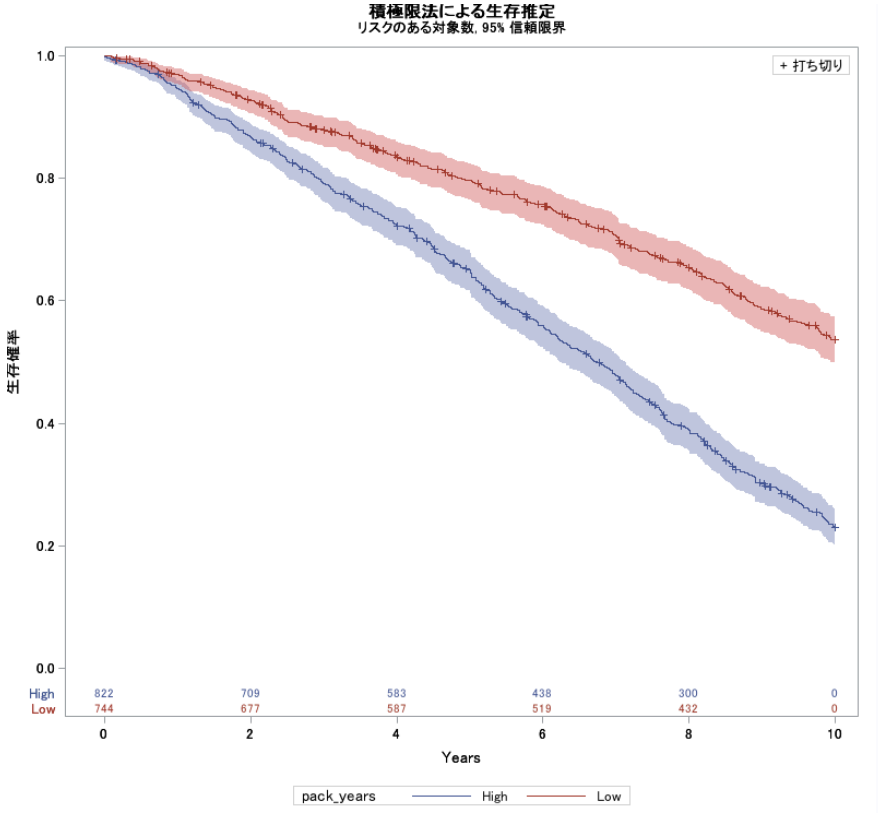
5.2 アウトカムの分布の確認

アウトカムである死亡までの時間が群間でどのように分布するか要約します.Time-to-event データの場合,Kaplan-Meier 曲線を使って時間の経過にともなう生存確率の変化を可視化します

SAS の LIFETEST プロシジャは生存時間解析を行うプロシジャで time ステートメントでイベント又は打ち切りまでの時間 * イベントまたは打ち切りを示す変数を指定し,括弧の中に打ち切りを表す数値を示します（複数記述可）.
strata で層別変数を指定します.

plot ステートメントで survival を指定すると,ods graphics という機能の働きで,特に細かい指定をしなくても,Kaplan-Meier 曲線が描画されます.

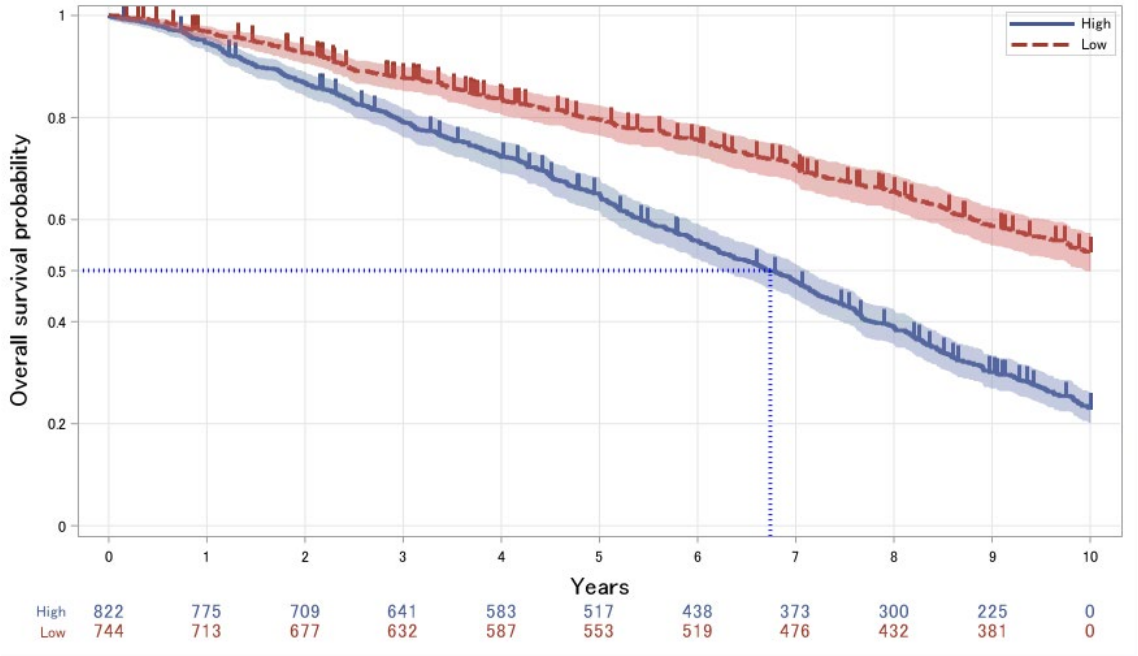
```
proc lifetest data=nhefs_complete3   plots=survival(atrisk cl) ;
        strata pack_years   /   test=logrank;
        time survtime_y * death(0);
run;
```



LIFETEST プロシジャ自体の指定で,プロットを調整することも可能ですが,

一般的に SAS の ods output 機能を使って,Kaplan-Meier 曲線を生成するための元データをデータセット出力し,それを加工して sgplot で自由に加工する方法が,柔軟性が高いです.

```
ods output Survivalplot=SurvivalPlotData;
ods output Quartiles=Quartiles;
proc lifetest data=nhefs_complete3    plots=survival(atrisk=0 to 10 by 1 cl) ;
                strata pack_years    /    test=logrank;
        time survtime_y * death(0);
run;

data SurvivalPlotData_1;
   set SurvivalPlotData ;
/*髭を散布図で描くためにシンボルを空白にする*/
   markerchar=" ";
/*＋-STD のエラーバーラインを引く機能を使って髭を表現する*/
   if ^missing(Censored) then hige=Censored+0.03;
run;

/*中央値をマクロ変数に*/
proc sql noprint;
 select Estimate into:p50_1 from Quartiles
 where Percent=50 and stratum=1;

 select Estimate into:p50_2 from Quartiles
 where Percent=50 and stratum=2;
quit;


ods graphics / reset noborder noscale    width=880 px height=510 px attrpriority=none;
proc sgplot data=SurvivalPlotData_1;
/*階段プロット部分*/
   step x=time y=survival / group=stratum name='s' lineattrs=(thickness=3);
/*ひげ部分*/
   scatter x=time y=censored /NOERRORCAPS yerrorupper=hige errorbarattrs=(pattern=1 thickness=3)
markerchar=markerchar GROUP=stratum;
/*信頼区間*/
            band    x=time upper=SDF_UCL lower=SDF_LCL/ group=stratum type=step transparency=0.6;
/*50%ドロップライン*/
 dropline x=&p50_1.    y=0.5 /dropto=both lineattrs=(color=blue    pattern=dot thickness=3);

   /*Number at risk*/
   xaxistable atrisk / x=tatrisk class=stratum location=outside colorgroup=stratum valueattrs=(size=11);
   keylegend 's' / location=inside position=topright across=1 valueattrs=(size=9) exclude="") ;
   yaxis min=0 label="Overall survival probability" grid values=(0 0.2 0.4 0.5 0.6 0.8 1) labelattrs=(size=13);
   xaxis    label="Years"    min=0    offsetmin=0.03    offsetmax=0.03    values=(0    to    10    by    1)    grid
labelattrs=(size=13);
run;
```



5.3 喫煙の程度の死亡との関連評価

次に単純な推定により,曝露とアウトカムとの関連を評価します.

5.3.1 生存時間中央値（MST: Median Survival Time）

生存時間中央値は生存確率が 50%になる時点のことです.

LIFETEST プロシジャを実行するとアウトプットに四分位点推定として表示されます

四分位点推定				
		95% 信頼区間		
パーセント	点推定	変換	[下限	上限)
75	9.82615	LOGLOG	9.37440	.
50	6.74059	LOGLOG	6.22861	7.12663
25	3.66598	LOGLOG	3.12115	4.10130

平均	標準誤差
6.31448	0.11132

四分位点推定				
		95% 信頼区間		
パーセント	点推定	変換	[下限	上限)
75	.	LOGLOG	.	.
50	.	LOGLOG	9.94661	.
25	6.13005	LOGLOG	5.21561	6.81725

平均	標準誤差
7.84613	0.11162

5.3.2 生存確率の群間比較

群間の生存確率が異なるかどうかは,例えば Log-rank 検定で推定できます.有意水準は 5%とします.

strata ステートメントの test オプションでその他の検定方法も選択できます. アウトプットの同等性の検定の箇所に表示されます

層の survtime_y に対する生存曲線の同等性の検定

順位統計量	
pack_years	ログランク
High	180.24
Low	-180.24

ログランク検定の共分散行列		
pack_years	High	Low
High	226.979	-226.979
Low	-226.979	226.979

層に対しての同等性の検定			
検定	カイ 2 乗値	自由度	Pr > Chi-Square
ログランク	143.1198	1	<.0001

p 値が有意水準未満であることから,喫煙の程度と死亡との間には関連が認められることがわかります.

このように Kaplan-Meier 曲線で生存確率を可視化したり,単純な解析で傾向を把握することは,思わぬミス（群のコーディングを逆にしていた.生存と死亡が逆だった等）を防いだり,その後の解析から得られる結果の洞察を深めるためにも重要です.

ここまでの解析で喫煙の程度と死亡には関連が認められることがわかりましたが,これから喫煙の程度が死亡に影響するか評価していきます.

5.4 喫煙の程度の死亡への影響評価

ここから,喫煙の程度が死亡に影響するか評価していきます.ポイントはバイアス,特に交絡の軽減です.バイアスを調整する方法は大きく 2 つに分かれます.

1.アウトカムモデル：アウトカムに対する回帰
2. PS モデル：曝露に対する回帰

1 は Cox 比例ハザードモデルの説明変数に曝露変数と交絡因子を入れたモデルにより,曝露の影響を評価します.アウトカムに対して直接モデルを構築するため「アウトカムモデル」といいます.

2 はまず傾向スコア（PS: Propensity score）を推定し,その後傾向スコアを利用して群間の

バランスを調整し,曝露の影響を評価します.傾向スコアは個人の背景情報から推定される曝露する確率です.曝露に対してモデルを構築するため「PS モデル」といいます.本パートでは,PS モデルを用いて解析を進めます.

バイアスを調整することができれば,喫煙の程度と死亡との間の関連（association）を定量化した推定値（本パートではハザード比）は,効果（effect）と見なすことができます.

5.4.1 推定目標（Estimand）

まず傾向スコアを算出し,それを利用して群間のバランスをとることでバイアスを調整します.傾向スコアを利用した解析方法はいくつかあり,どの方法でもバイアスを調整することはできます.しかし推定目標（Estimand）が異なります.下表に傾向スコアの利用方法と推定目標をまとめました.

傾向スコアを利用した解析方法	推定目標
層別解析	ATE（層を併合）、Conditional effect（層ごと）
回帰	Conditional effect
マッチング	ATT、ATM
逆確率重み付け	ATE

ATE は平均処置効果（Average Treatment Effect）です.これは研究対象集団が仮に全員曝露した場合と,仮に全員曝露しなかった場合のアウトカムの平均値,リスクや率を比較します.今回のテーマでは研究対象集団全員が喫煙量を減らすと予後はどうなるのか,ということに関心があるので推定目標は ATE になります.

ATT は処置群における平均処置効果（ATE in the Treatment group）.Conditional effect は条件付き効果といい, PS や層が同じ集団における ATE.ATM はマッチングした集団における平均処置効果（ATE in the matched population）です.これらの説明は割愛します.

本テーマの推定目標は ATE なので,逆確率重み付けを用いて解析を進めていきます.

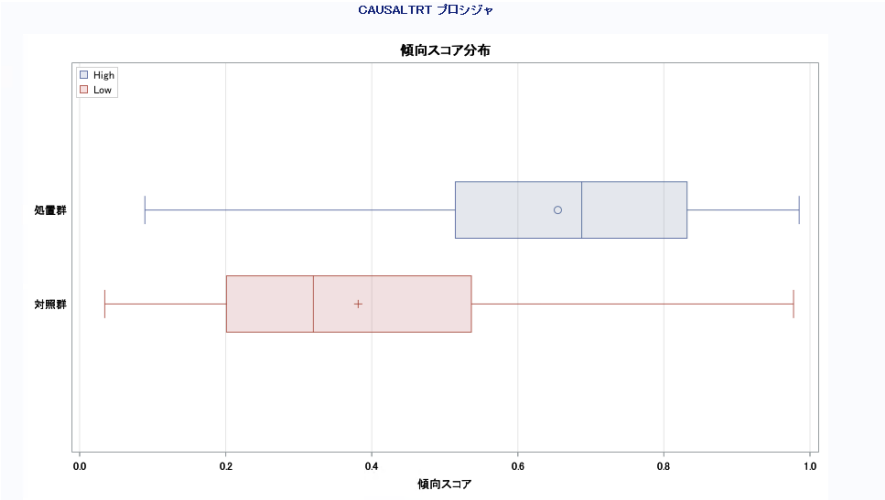
5.4.2 逆確率重み付けを用いた群間バランスの調整

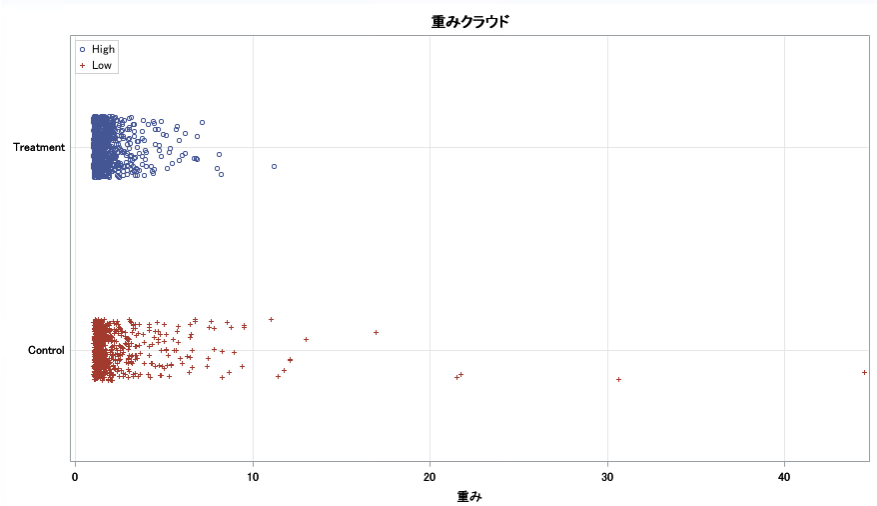
逆確率重み付け法（IPW or IPTW: Inverse Probability Treatment Weighting）は,個人ごとに推定した傾向スコアの逆数により,群間のバランスを調整する方法です.詳細な説明は割愛します

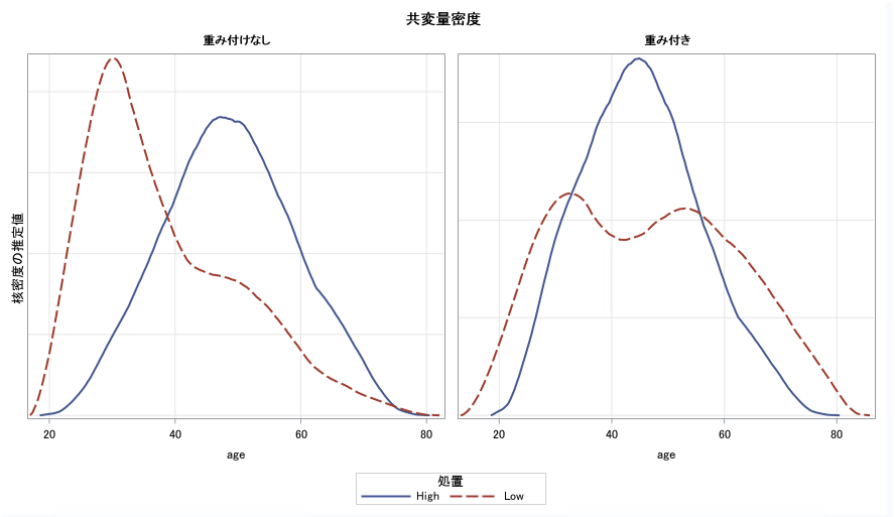
傾向スコアとその逆確率を計算する方法は複数ありますが,ここでは causaltrt プロシジャを使った方法を紹介します.

```
ods output PSCovDiff=PSCovDiff;
proc causaltrt data=nhefs_complete3 covdiffps method=IPW ;
  class pack_years sex race education exercise;;
  psmodel pack_years = sex age race education exercise/plots=all;
  model death;
  output out=causal_output ps=ps ipw=ipw;
run;
```

plots=all をつけることで作図可能なプロットがすべて表示され,チェックすることができます

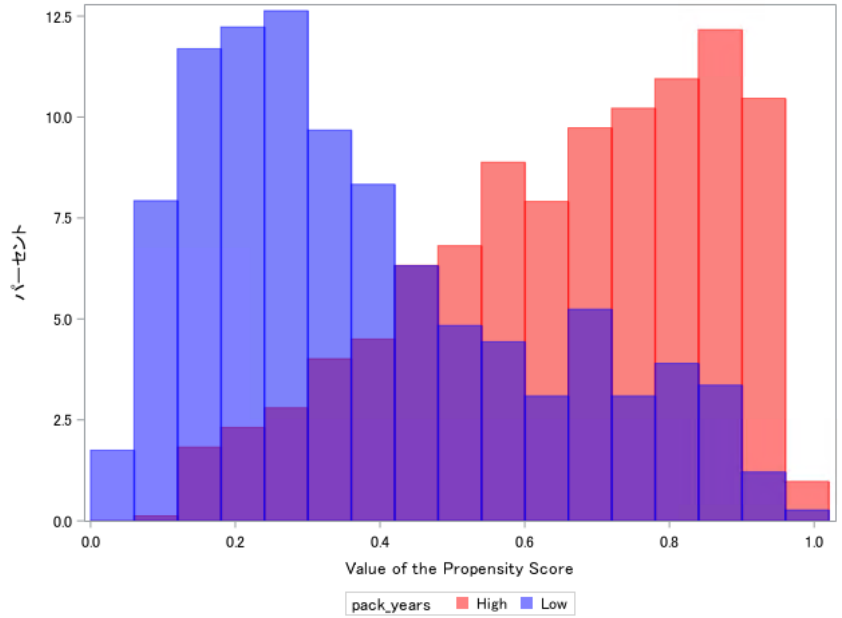




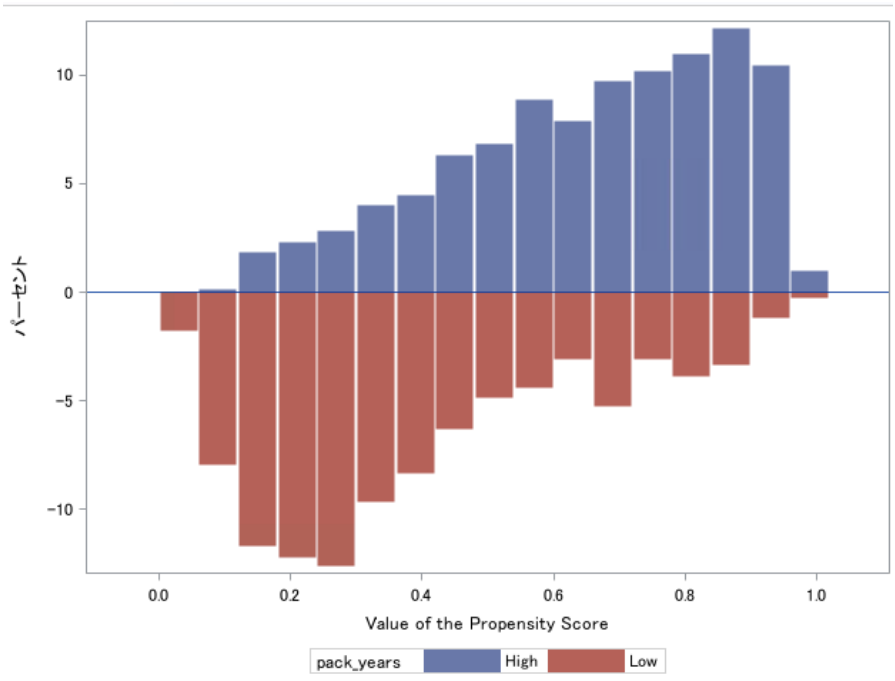


output ステートメントを使うことで,傾向スコアや逆確率の重みなどを含めて元データを出力することができるので,それを使って自分の好みのプロットも作成できます

```
ods graphics / reset    attrpriority=none;
ods output sgplot=histo;
proc sgplot data=causal_output;
styleattrs
datacolors=(red blue)
datacontrastcolors=(red blue);
  histogram ps /transparency=0.5    group=pack_years ;
run;
```



```
data histo1;
set histo;
 if  BIN_PS_GROUP_PACK_YEARS___GP  =  0  then  BIN_PS_GROUP_PACK_YEARS____Y  =
BIN_PS_GROUP_PACK_YEARS____Y * -1;
 run;

ods graphics / reset   attrpriority=none;
proc sgplot data=histo1;
where ^missing(BIN_PS_GROUP_PACK_YEARS____Y);
styleattrs
datacolors=(     red blue);
needle          x=BIN_PS_GROUP_PACK_YEARS____X          y=BIN_PS_GROUP_PACK_YEARS____Y
/transparency=0.2   baseline=0 group=BIN_PS_GROUP_PACK_YEARS___GP
              lineattrs=(thickness=26px pattern=solid);
run;
```



逆確率重み付け法では,重みの大きさを確認することも重要です.

```
proc means data=causal_output;
 class pack_years;
 var ipw;
run;
```

MEANS プロシジャ

分析変数：ipw Value of the Weight						
pack_years	Obs 数	N	平均	標準偏差	最小値	最大値
Low	744	744	2.2832840	2.7645880	1.0356297	44.4838693
High	822	822	1.8201285	1.0779642	1.0150424	11.1896814

最も大きい重みは,pack_years 低い群の 44.48 です.この値程度では問題ないのですが,極端に大きな値の場合注意が必要です.

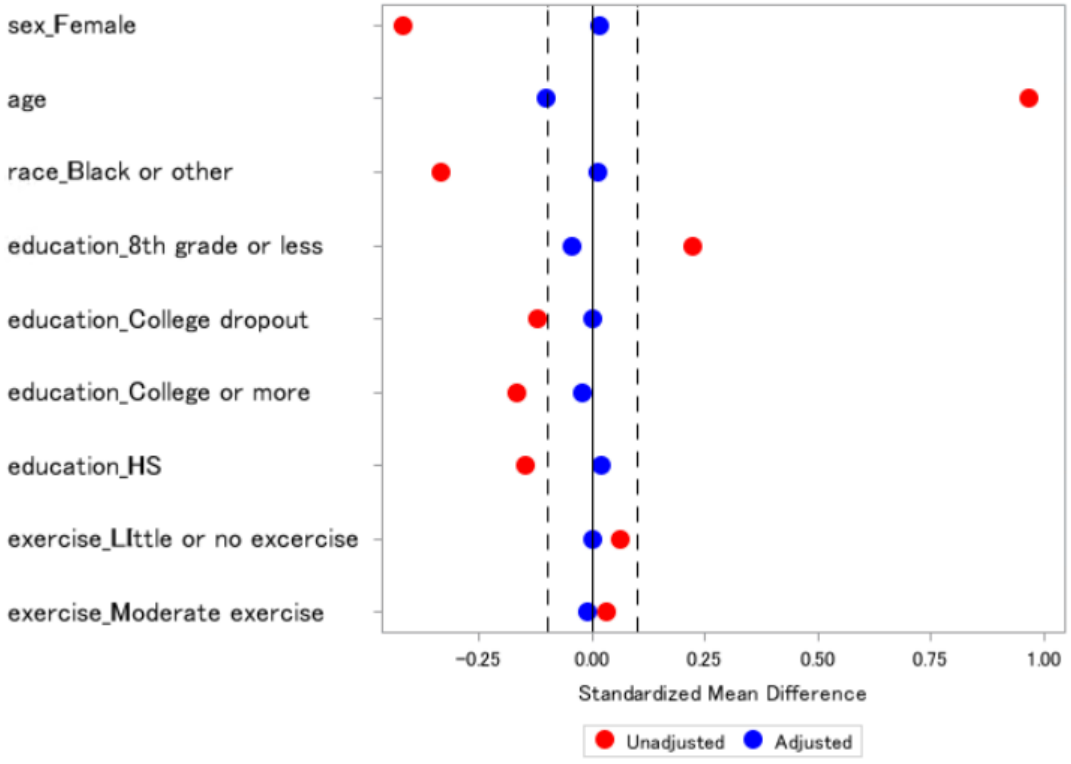
次に交絡因子ごとに群間のバランスを確認します.従来は群間の p 値で確認していましたが,最近は標準化差 (Standardized differences) で確認することが多いです.標準化差の絶対値が 0.1 以下のときに,その変数は群間でバランスがとれていると経験的に考えます
最初の causaltrt プロシジャに covdiffps をつけているため
以下の出力がアウトプットに含まれます

傾向スコアモデルの共変量の差					
		標準化された差		分散比	
パラメータ		重み付けなし	重み付き	重み付けなし	重み付き
sex	Female	-0.4190	0.0177	1.0322	0.9985
sex	Male				
age		0.9647	-0.1038	0.7719	0.4988
race	Black or other	-0.3360	0.0109	0.4649	1.0242
race	White				
education	8th grade or less	0.2212	-0.0469	1.4444	0.9386
education	College dropout	-0.1197	0.0009	0.6828	1.0030
education	College or more	-0.1688	-0.0225	0.6593	0.9412
education	HS	-0.1482	0.0216	0.9462	1.0091
education	HS dropout				
exercise	LIttle or no excercise	0.0602	-0.0002	1.0287	0.9999
exercise	Moderate exercise	0.0313	-0.0109	1.0100	0.9968
exercise	Much exercise				

数値だとわかりづらいので,可視化します

```
data PSCovDiff1;
set PSCovDiff;
where ^missing(StdDiff);
level_name=catx("_",Parameter,Level1);
y=_N_;
y2=y;
```

```
label y="Unadjusted" y2="Adjusted";
run;

proc sgplot data=PSCovDiff1;
 scatter y=y x= StdDiff /markerattrs=(symbol=circlefilled size=9pt color=pink) name="name1";
 scatter y=y2 x= StdDiffWeighted/markerattrs=(symbol=circlefilled size=9pt color=lightblue)
name="name2";
 yaxistable level_name/label=" " position=left; ;
 refline 0 / axis=x lineattrs=(color=black);
 refline -0.1 0.1 / axis=x lineattrs=(color=black pattern=dash);
yaxis reverse display=(nolabel novalues ) values=(1 to 9);
keylegend "name1" "name2";
run;
```



図より,概ね調整後はバランスとれているが,age のみ標準化差が 0.1 のラインにかかり,やや大きいことがわかります.本来ならば,バランスをとれるように交絡因子を検討し直したり,PS モデルに交互作用や高次の項を含めたり試行錯誤をおこなった後に ATE の推定をおこないますが,今回は練習なので次に進みます

5.4.3 推定

いよいよ ATE の推定です.SAS の phreg プロシジャの weight ステートメントで先に推定した重みを指定します.また逆確率重み付け法では交絡を調整するために対象集団を擬似的に膨らましているため,通常の推定では 95%信頼区間が過剰に狭くなってしまうため. NORMALIZE オプションを使ってそれを防いでいます.

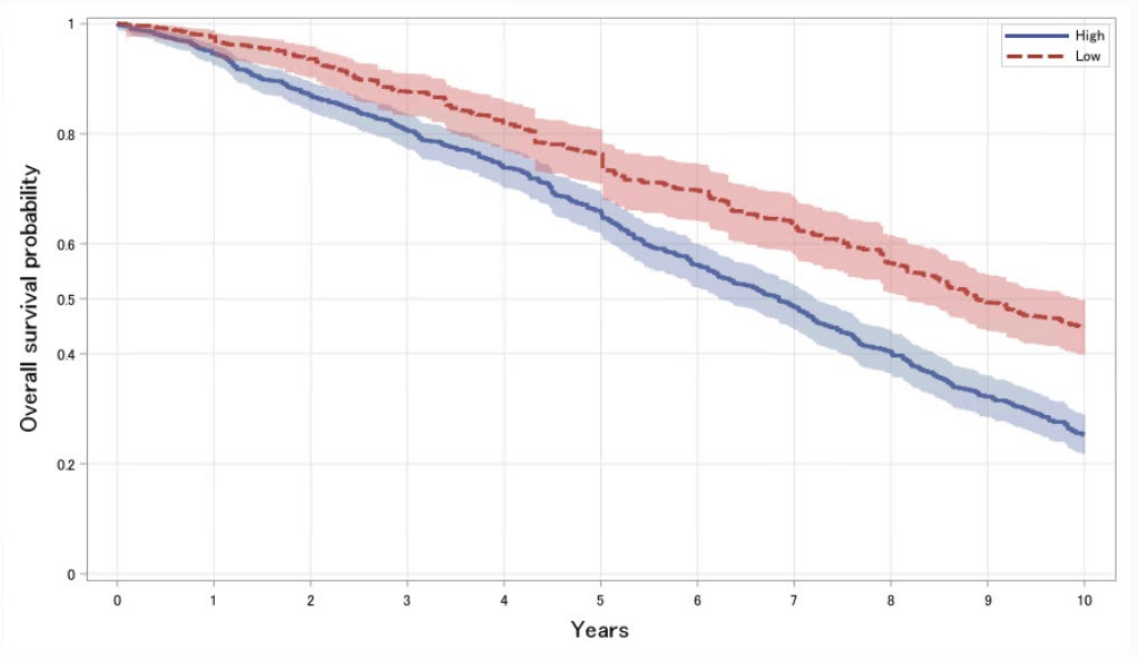
```
proc phreg data=causal_output;
 model survtime_y * death(0) = pack_years / rl;
 weight ipw / normalize;
run;
```

最尤推定値の分析

パラメータ	自由度	パラメータ推定値	標準誤差	カイ 2 乗値	Pr > ChiSq	ハザード比	95% ハザード比信頼限界	
pack_years	1	0.50695	0.06533	60.2101	<.0001	1.660	1.461	1.887

推定した結果,喫煙量が低い群に比べて高い群のハザード比は 1.66（95%信頼区間, 1.46 to 1.89, P < 0.001）であり,交絡を調整しても喫煙の程度と死亡には正の関連が認められることがわかります.交絡が十分に調整できていると考えるならば,「喫煙の程度は死亡に影響する」と解釈できます.

最後に,逆確率重み付け法で交絡を調整した生存確率を可視化します.



調整前の Kaplan-Meier 曲線と形状はほぼ同じですが,全体的に生存確率が低くなっている

ことがわかります.

6 まとめ

生存時間解析の流れを通して,可視化がどのように使われるか見てきました.可視化しなくても推定や検定はおこなえます.しかし可視化することで,思わぬミスを防いだり,よりわかりやすく結果を伝えることができます.
これからも魅力ある可視化をしていきましょう！

[森岡から]
今回は佐藤俊太朗先生と藤井亮輔様が作成された R のビジュアライゼーションについての素晴らしい資料について,もし SAS で実装したらどうなるかという企画をいただきました.データを受け取り,可視化して構造や傾向を把握し,さらに段階ごとに解析結果を可視化して確認し,わかりやすく伝える.プログラム言語が違っても,統計解析の本質は変わらないなと改めて実感しました.そして,R と見比べながら,グラフ描いたり,解析をするという経験はとても刺激になり,私にとっても学ぶことが多かったです.一点ご注意というか言い訳なのは,ここで示したのはあくまで,描き方,解析のアプローチの一例であり,他にもよりよい描き方はあると思いますし,解析部分については厳密には内部での解析処理が R と一致しているかなどまでは精査しきれておりません.あくまで,流れをおって,SAS ではこうやってくのかぐらいで見ていただけると有難いです.
最後に,SAS はビジュアライゼーションが駄目だと,R 使い・Python 使いにイジメられるのですが,ここ 10 年ぐらいの機能でいうと必ずしもそうでもないよと主張したい(笑)
自分で指定したり・構築しなければいけない要素は多いのですが,それを組み立てている過程でデータへの理解が促進される作用もあるように思うのです.まあ,でも R,簡単な指定で,きれいな出力がでていいなぁとも思いましたけど.多言語で解析すると気づきが多くて楽しいですね.今回の資料は,私が誰よりも楽しかった,私が勝手に楽しんだ資料になってしまっているのですが,ほんの少しでも皆様の役に立って,R/Stata/SAS 間のコミュニケーションの助けにもなれば僥倖です.

作成者：森岡　裕(イーピーエス株式会社)
Mail: morioka.yutaka038@eps.co.jp