Part 3 推測統計の可視化

2 Data source

別途、Complete データを csv ファイルから読み込み、必要な変数のみにまとめたり、値ラベルを定義したりします（この Complete データの csv ファイルは、R から呼び出したファイルです)。

part3.do

```
* Data source
import delimited "nhefs_complete.csv", clear

// 必要変数のみ残す
keep seqn death yrdth modth dadth sex age race education exercise ///
     smokeintensity smokeyrs

// label
label define sex 0 "Male" 1 "Female"
label values sex sex

label define race 0 "White" 1 "Back or other"
label values race race

label define edu 1 "8th grade or less" 2 "HS dropout" 3 "HS" 4 "College dropout"
5 "College or more"
label values education edu

label define exe 0 "Much exercise" 1 "Moderate exercise" 2 "Little or no exercise"
label values exercise exe
```

4 Data

4.1 Outcome

R から出力した csv を Stata で読み込むと、欠損値（NA）が文字列として扱われるため、注意が必要です。ここでは、それを逆手に取って、文字列型のまま死亡日（tmp1）を作成しています。その tmp1 を date 関数で SIF 形式に変換しています。

part3.do

```
* 死亡日・打切り日作成
gen tmp1 = "19" + yrdth + "/" + modth + "/" + dadth if death==1
gen tmp2 = date(tmp, "YMD")
gen event_date = cond(death==0, td(31Dec1992), tmp2)
```

```
drop tmp*
format %td event_date

* 観察開始日からの日数を計算
gen survtime = datediff(td(01Jan1983), event_date, "day")
```

4.2 Exposure

gen コマンドで変数を作成し、**label** コマンドで値ラベルを定義し、貼り付けています。

part3.do

```
gen pack_years_n = (smokeintensity / 20) * smokeyrs
gen pack_years = cond(pack_years >= 20, 1, 0)
label define pack_years 0 "Low" 1 "High"
label values pack_years pack_years
```

4.3 練習のためのデータ変更

R での操作をトレースしていますが、乱数の出目が異なるため、同一データにはなっていません。

part3.do

```
* 4.3 練習のためのデータ変更
set seed 1234

gen     death1     = rbinomial(1, 0.6)
gen     death0     = rbinomial(1, 0.4)
gen     death2     = cond(pack_years == 1, death1, death0)
gen     censor     = rbinomial(1, 0.2)
gen     r_time     = ///
      cond((censor == 1 & death == 0) | (death2 == 1 & death == 0), ///
      round(runiform(0, 3652)),0)
replace death      = cond(death2 == 1, 1, death)
replace survtime   = survtime - r_time
gen     survtime_y = survtime/365.25
```

4.4 本パートで仮定した因果構造

R では DAG を描画していますが、Stata では（調べた限り）DAG を描画するためのコマンドがありませんでした。DAG をベースにした解析を行なう外部コマンド **dag** を作成したチームが下記のコメントをしています。

We are working on A Stata command drawing a high-quality DAG.
https://medical-statistics.dk/MSDS/epi/dag/dag.html

5 統計解析

5.1 背景情報の要約

外部コマンド **table1** を用いて背景情報をまとめることができます。**table1** は **vars** オプション内で記述する変数とその型をしていしています。**by** オプションで指定した変数で分けて記述統計量を算出します。

part3.do

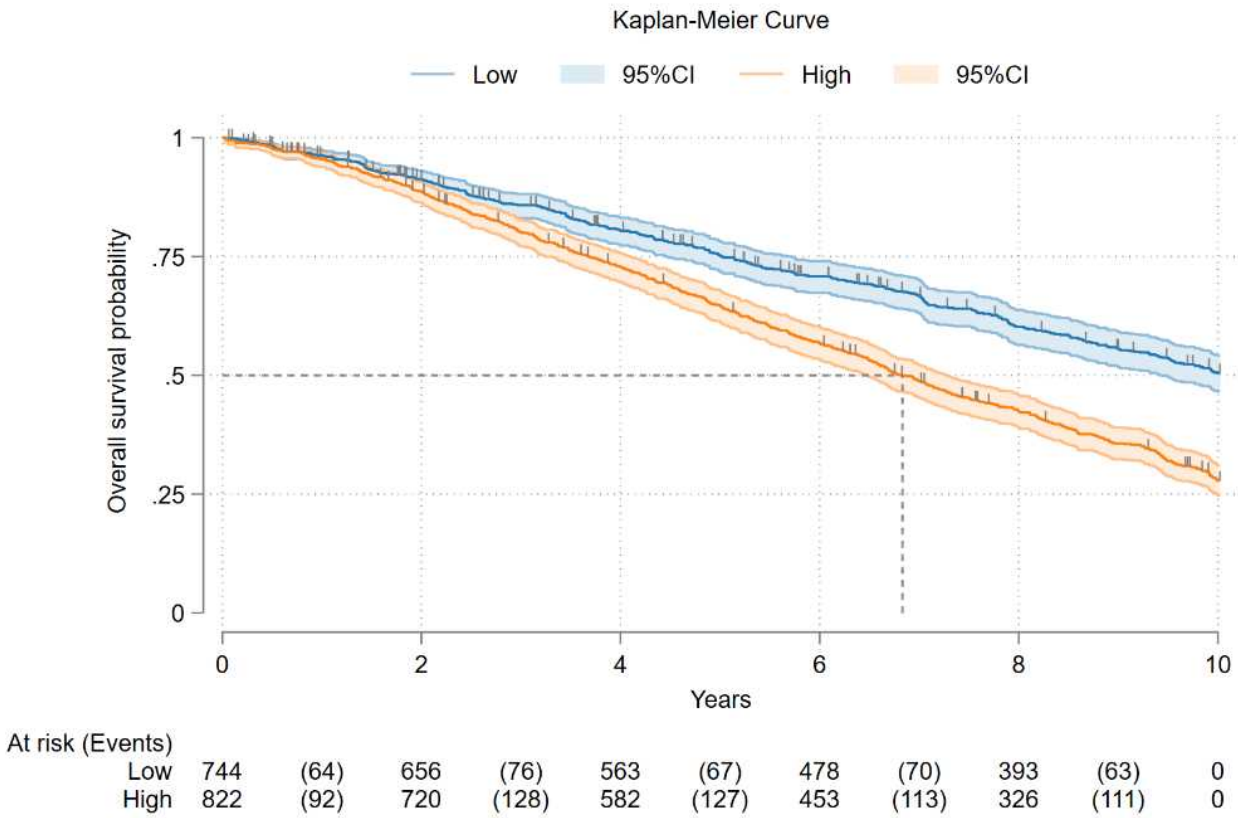
```
table1, vars(sex bin \ age conts \ race bin \ education cat \ exercise cat) by(pack_years) test
```

5.2 アウトカムの分布の確認

Kaplan-Meier 曲線を描いています。Stata では **sts graph** で生存時間分析におけるグラフ描画が行えます。リスク表を作成するために **risktable** オプション、MST を図示するための **addplot** オプションのために、かなり長大になっています。

part3.do

```
// Survival analysis 実行用にデータ設定
stset survtime_y, fail(death==1)

// High における MST を先に算出してグラフで利用
qui stsum if pack_years==1
local high_p50 = `r(p50)'

// グラフ描画
sts graph, ci scheme(white_tableau) by(pack_years)  ///
     risktable(0(2)10, rowtitle(Low)  failevents ///
            size(small) group(#1) title("At risk (Events)", size(small))) ///
     risktable(0(2)10, rowtitle(High) failevents size(small) group(#2)) ///
     xtitle("Years") ytitle("Overall survival probability") ///
     title("Kaplan-Meier Curve", size(small)) /// 表題など
     censored(single) /// 打切りの印を付ける
     legend(position(12)) /// legend を 12 時方向に
     addplot(function y=`high_p50', ///
            horizontal range(0 0.5) lcolor(gs8) lpattern(shortdash) || ///
            function y=0.5, ///
            range(0 `high_p50') lcolor(gs8) lpattern(shortdash) legend(row(1) ///
     order(2 "Low" 1 "95%CI" 4 "High" 3 "95%CI"))) ///
     name(KMC_unwt, replace)
```



リスクテーブルでは、カッコ内の数字が、その区間の Event 発症者数になります。また、打切り者数をリスクテーブルに表示する方法は、（調べた限り）ありませんでした。グラフ中に打切り者数を表示させることは可能ですが、数字が見にくくなってしまうため、採用していません。

5.3 喫煙の程度と死亡との関連評価

5.3.1 生存時間中央値

下記は Stata では、**stset** 後でのみ有効なコマンドです。今回はグラフ描画の最初で **stset** を実行していますので、有効です。オプションを変更する事で、中央値以外の N パーセンタイルを算出することもできます。

part3.do

```
stsum, by(pack_years) // 点推定値など
stci , by(pack_years) // MST とその信頼区間
```

5.3.2 群の時点ごとの生存確率と 95%信頼区間

part3.do

```
sts list, risktable(0/9) by(pack_years) enter
```

R での実行結果とは異なり、時点ゼロでの At risk 人数がゼロ人になっています。代わりに、**enter** オプションを指定しているので Enter という列があり、時点ゼロで 744 人（822 人）が追跡コホートに入っ

たことを示しています。

5.3.3 生存確率の群間比較

Log-rank 検定を実行します。

part3.do

```
sts test pack_years
```

5.4 喫煙の程度の死亡への影響評価

5.4.2 逆確率重み付けを用いた群間バランスの調整

Stata では、個人ごとの曝露確率（pack_years=1 になる確率）を logistic model で算出します。通常であれば、**xi** プレフィクスコマンドは使わなくても問題ありません。この **xi** プレフィクスコマンドを付けることによって、一時的に作成されたダミー変数を自動削除せずに残しておくことができます（今回は後で使うので残しています)。

part3.do

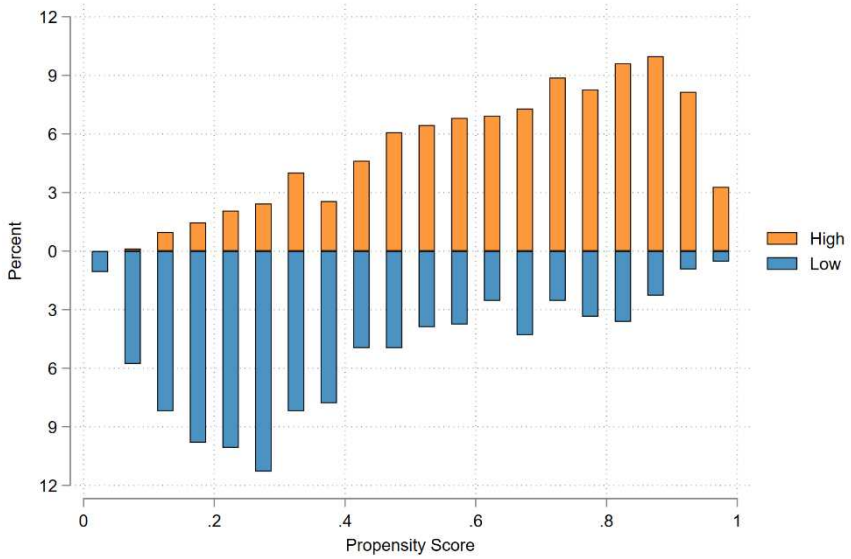
```
xi:logit pack_years i.sex age i.race i.education i.exercise
predict ps, pr

gen iptw = cond(pack_years==1, 1/ps, 1/(1-ps))
```

次に傾向スコアの分布を確認する貯めに、ヒストグラムを描画しますが、Stata の標準では、R で作成されたような上下に分かれるヒストグラムは作成できません。今回は外部コマンド **bihist** を利用して、作図します。

まず、重みを付けていない場合のヒストグラムを描画します。**bihist** コマンドはやや特殊で、描画の外見に関するオプションを **tw** オプションの中に入れて記載する必要があります。

part3.do

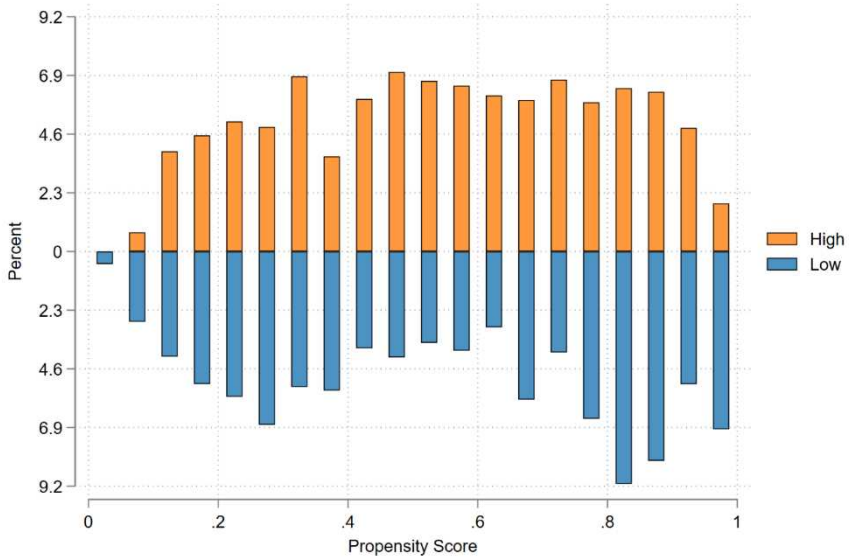
```
bihist ps, by(pack_years) percent width(0.05) start(0) ///
     tw(scheme(white_tableau)  xtitle(Propensity Score) ytitle(Percent) ///
            legend(order(2 1))) ///
     name(unwt_hist, replace)
```



次に、重みを付けてたグラフを描画します。重みを指定するための記述（**[pw=iptw]**）を追記します。

part3.do

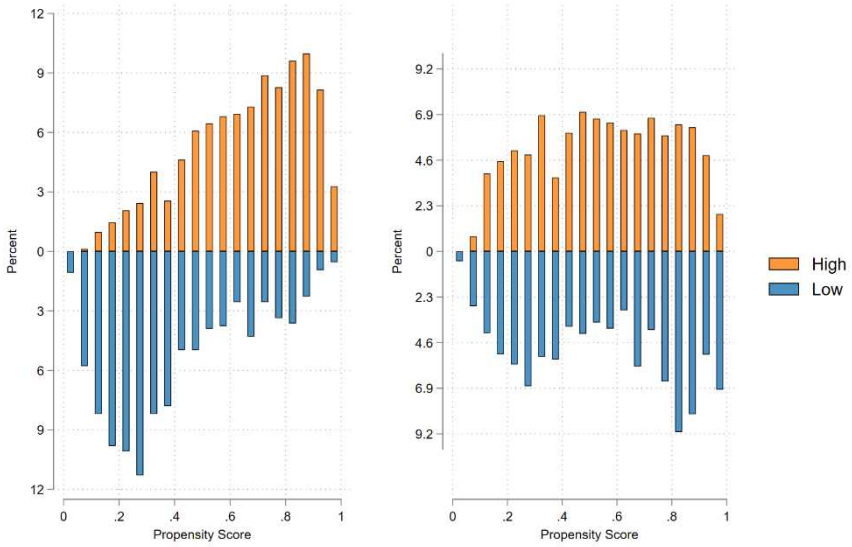
```
bihist ps [pw=iptw], by(pack_years) percent width(0.05) start(0) ///
     tw(xtitle(Propensity Score) ytitle(Percent) ///
            legend(order(2 1))) ///
     name(iptw_hist, replace)
```



これら 2 つのグラフを 1 つにまとめることで、比較しやすいグラフを作成しようと思います。標準コマンドであれば、**graph combine** で結合するところですが、Legend が重複してしまうという欠点があります。一般には、結合する片方で Legend を付けないという対策で回避しますが、ここでは、外部コマンド **grc1leg** を利用します（Graph Combine 1 Legend の略）。

オプション **ycommon** は y 軸を 2 つのグラフで同じスケールにするという指定で、**positon(3)**は Legend を 3 時方向に入れるという指定です。

part3.do
```
grc1leg unwt_hist iptw_hist,  ycommon position(3) name(comb_hist, replace)
```



重みの大きさの確認を行います。これは、**bysort** プレフィックスを付け、**summarize** コマンドで重み変数 **iptw** を表示させています。なお、**bysort** は **bys**、**summarize** は **su** で略しています。

part3.do
```
bys pack_years: su iptw, detail
```

次に、交絡要因ごとに群間のバランスを確認します。ここでは外部コマンドの **covbal** を用います。なお、この結果が R で行われた結果が一部異なっています。この原因・理由については検証できていません。

part3.do
```
covbal pack_years ps _Isex_1 age _Irace_1 _Ieducation_2 _Ieducation_3 ///
       _Ieducation_4 _Ieducation_5 _Iexercise_1 _Iexercise_2, ///
       saving(bal_unwt, replace)

covbal pack_years ps _Isex_1 age _Irace_1 _Ieducation_2 _Ieducation_3 ///
       _Ieducation_4 _Ieducation_5 _Iexercise_1 _Iexercise_2, ///
       wt(iptw) saving(bal_iptw, replace)
```

ここで、傾向スコア算出の際に、**xi** をプレフィックスにつけたことが生きてきます。**_I** から始まる変数は、**xi** によって生成された変数です。**_Ieducation_1** は変数 **education=1** のダミー変数（つまり、8th grade or less）を示しています。他も同様です。

3 カテゴリ以上ありカテゴリ変数において標準化差を算出するには、2 つ方法があり、1 つはダミー変

数化したものを二値変数としてそれぞれについて標準化差を計算うする方法で、もう 1 つは専用の公式を用いる方法です。Rの **love.plot** 関数では 1 つめの選択をとっていますので、ここでもそれに合わせました。

1 つめの **covbal** コマンドでは、重み付け前のバランスを標準化差と分散比で表示させます。2 つめの **covbal** では、3 行目に **wt(iptw)**というオプションが入っていますが、これが変数 **iptw** で重み付けた後のバランス評価を行うという指示になります。いずれの **covbal** にも **saving** オプションを指定しています。これは、次に可視化するときに役に立つので、得られた結果表をデータとしてセーブするためのものです。

変数の群間バランスを図示するための標準コマンドおよび外部コマンドは調べて限りありませんでした。しかし、**covbal** コマンドで出力したデータを元にすれば、Stata でも散布図の応用で作図可能です。

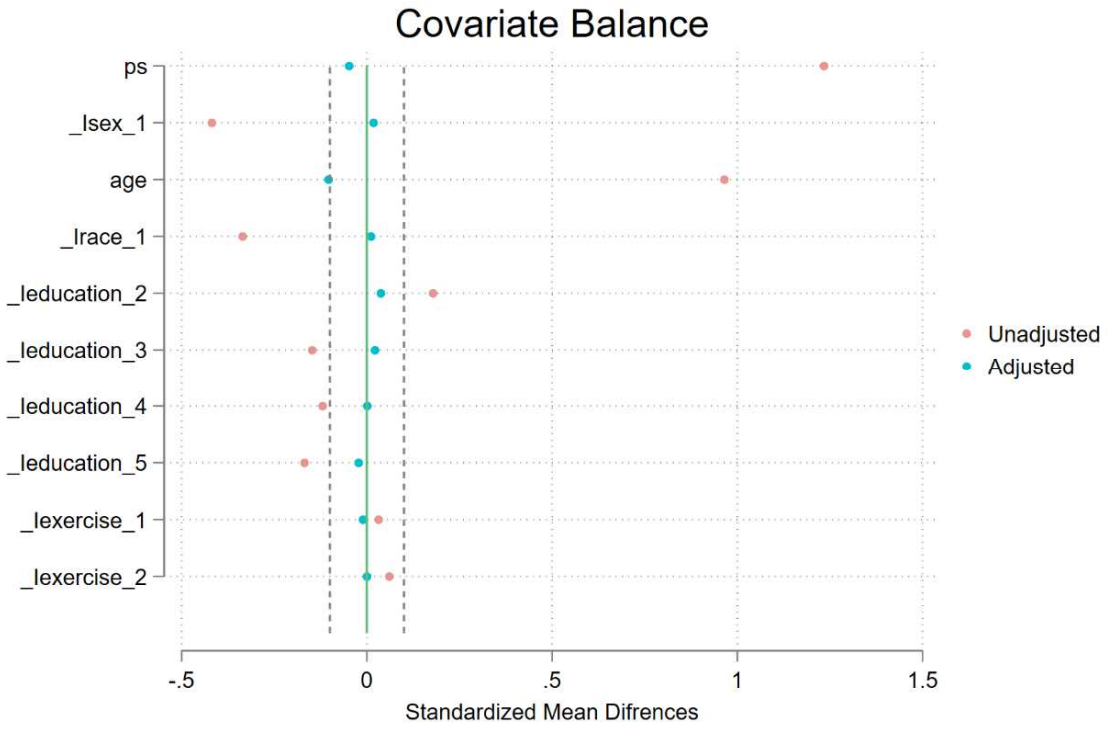
part3.do

```
capture frame drop covbal
frame create covbal
frame covbal {
	* データセット加工
	use bal_unwt, clear
	gen odr = _N - _n + 1
	keep odr varname stdiff varratio
	rename stdiff stdiff_unwt
	rename varratio varratio_unwt

	merge 1:1 varname using bal_iptw
	assert _merge == 3 // _merge==3 以外があれば、何かおかしいので assert で止める
	drop _merge tr_mean tr_var tr_skew con_mean con_var con_skew
	rename stdiff stdiff_iptw
	rename varratio varratio_iptw
	sort odr

	labmask odr, value(varname)

	* カラーパレット設定
	colorpalette hcl, select(1 6 9) nograph
	local unadj `r(p1)'
	local adj `r(p3)'
	local zero `r(p2)'
```

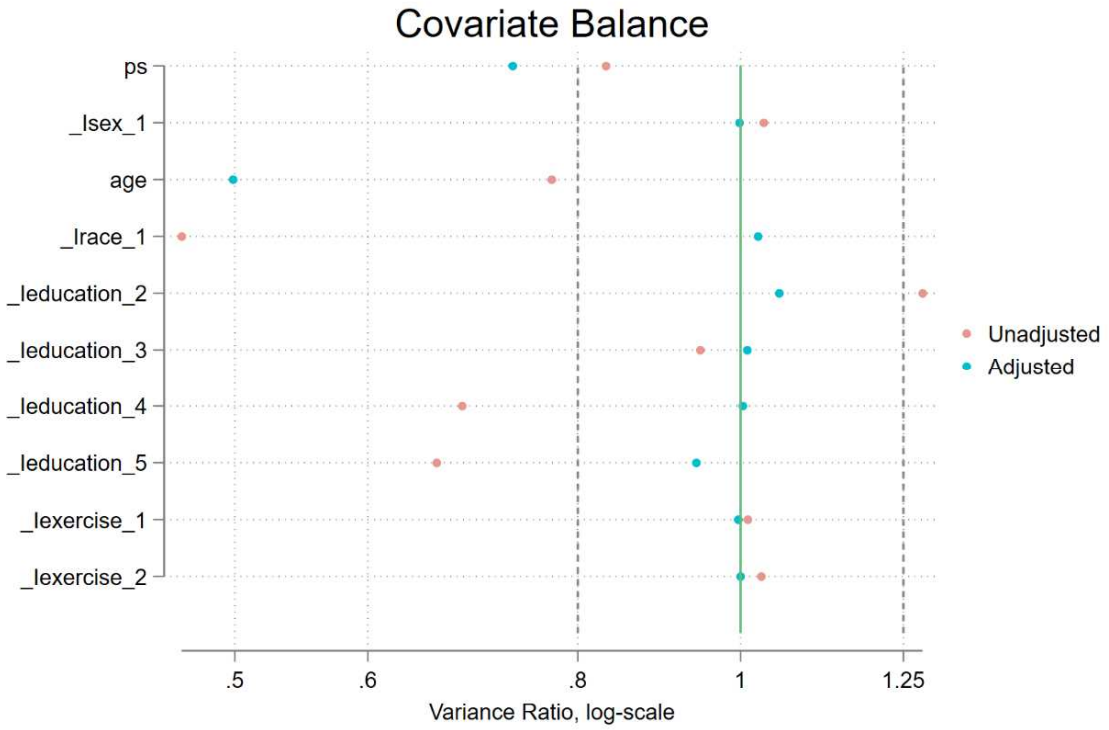
```
        * 標準化差のグラフ
        twoway ///
        catter odr stdiff_unwt, ylabel(10(1)1, valuelabel) mcolor("`unadj'") || ///
        scatter odr stdiff_iptw, ylabel(10(1)1, valuelabel) mcolor("`adj'") || ///
        function y= 0.1, horizontal range(0 10) lcolor(gs8) lpattern(shortdash) || ///
        function y=-0.1, horizontal range(0 10) lcolor(gs8) lpattern(shortdash) || ///
        function y=0   , horizontal range(0 10) lcolor("`zero'")  ///
        legend(order(1 "Unadjusted" 2 "Adjusted")) ///
        xtitle("Standardized Mean Difrences") title(Covariate Balance) ///
        name(bal_smd, replace)

        * 分散比のグラフ
        twoway ///
        scatter odr varratio_unwt, ylabel(10(1)1, valuelabel) mcolor("`unadj'") || ///
        scatter odr varratio_iptw, ylabel(10(1)1, valuelabel) mcolor("`adj'")|| ///
        function y=1.25, horizontal range(0 10) lcolor(gs8) lpattern(shortdash) || ///
        function y= 0.8, horizontal range(0 10) lcolor(gs8) lpattern(shortdash) || ///
        function y=   1, horizontal range(0 10) lcolor("`zero'")  ///
        legend(order(1 "Unadjusted" 2 "Adjusted")) ///
        xscale(log) xlabel(0.5 0.6(0.2)1.0 1.25) xscale(extend) ///
        xtitle("Variance Ratio, log-scale") title(Covariate Balance) ///
        name(bal_vr, replace)
}
```

作図は **twoway** で 5 つのレイヤーを重ねています。1 つめが重み付け無しの場合のプロット、2 つめが重み付けた時のプロット、3 つめが閾値の上限の縦ライン、4 つめが閾値の下限の縦ライン、5 つめが中央の縦ラインです。閾値ラインの設定は **function y=0.1** などの数値を変えることで変更できます。

標準化差のプロットは下記の様になります。



分散比のプロットは下記の様になります。横軸を対数軸にしていることに注意して下さい。



5.4.3 推定

ATE を推定します。Stata で生存時間分析を実施する際には、生存時間分析のための設定を先に指定しておく必要があります。そのための設定が **stset** コマンドです。時間を表す変数、重み、イベントを示す変数を設定しました。次に **stcox** コマンドで Cox 比例ハザードモデルによる解析が行うことができます。

part3.do

```
// Survival analysis 実行用にデータ設定（iptw で重みを付ける）
stset survtime_y [pw=iptw], fail(death==1)

// Cox 比例ハザードモデル
stcox pack_years
```

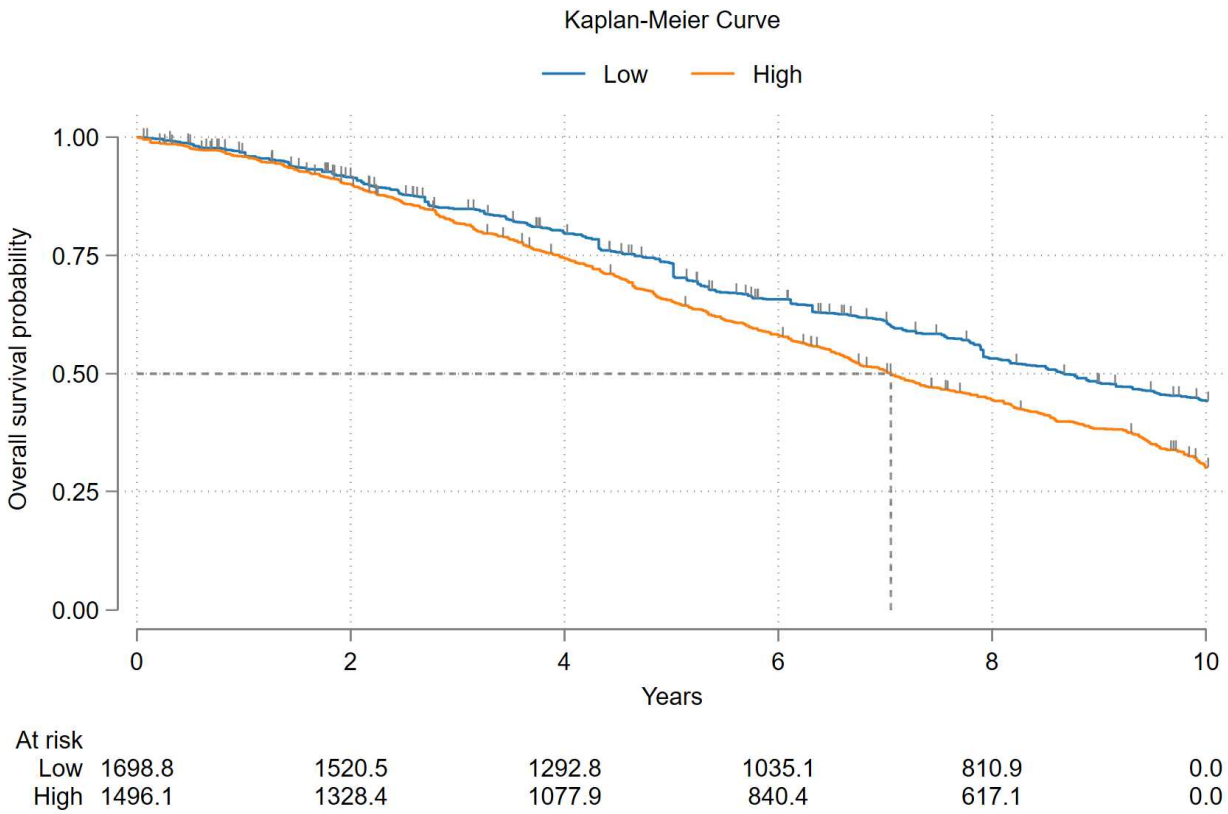
乱数の出目が R とはことなったり、原因不明ですが IPTW による共変量バランスの調整具合が異なったりしていますので、ハザード比についても、R での結果（点推定値 1.46）とは異なり 1.40（95%CI: 1.17, 1.68）となりました。

この重み付けによって交絡を調整した生存確率について、可視化します。基本的に重み付けを行う前と同じ操作を行っていますが、**ci** オプションを削除しています。これは、Stata による生存確率のグラフでは、stset で重み付けを指定していると信頼区間の描画がでないためです。また、イベント数を表示することは出来なくはありませんが、重み付けであるため、イベント数には小数点下の数値もあります。いろいろと工夫はしたのですが、小数点下の桁数の表記が上手くいかず、諦めました。

part3.do

```
// High における MST を先に算出してグラフで利用
qui stsum if pack_years==1
local high_p50 = `r(p50)'

// グラフ描画本体
sts graph, scheme(white_tableau) by(pack_years)  ///
       risktable(0(2)10, format(%9.1f) rowtitle(Low) ///
              failevents format(%9.1f) size(small) group(#1) ///
              title("At risk", size(small))) ///
       risktable(0(2)10, format(%9.1f) rowtitle(High) ///
              failevents size(small) group(#2)) ///
       xtitle("Years") ytitle("Overall survival probability") ///
       title("Kaplan-Meier Curve", size(small)) ///
       censored(single) ///
       legend(position(12)) ///
       addplot(function y=`high_p50', horizontal range(0 0.5) ///
```

```
        lcolor(gs8) lpattern(shortdash) || ///
        function y=0.5, range(0 `high_p50') lcolor(gs8) ///
        lpattern(shortdash) legend(row(1) ///
    order(1 "Low" 2 "High"))) name(KMC_iptw, replace)
```



At risk						
Low	1698.8	1520.5	1292.8	1035.1	810.9	0.0
High	1496.1	1328.4	1077.9	840.4	617.1	0.0