（2017.7.6 改訂）

LightStone

Stata で簡単に試せる ERM

ERM の概要と eregress の実施例

Stata15 に新搭載された ERM を紹介します．本文はすべてマニュアル『[ERM] Stata Extended Regression Models Reference Manual』からの「intro 1」から「intro 8」までの内容を基に加筆・修正を行ったものです．

ERM の概要

ERM（Extended Regression Model，拡張回帰モデル）は Stata が独自に名付けるモデルで，

(a) 内生共変量

(b) 内生性のあるサンプルセレクション

(c) 内生性のある処置

の 3 つの要素が同時に混在する場合でも一致性のある推定値が得られます．(a)〜(c) のうちいずれかのみが存在するケースにももちろん対応できます．ERM は連続，バイナリ，カウントの各アウトカムを従属変数とした場合に対応しています．

既存コマンドとの対応

各要素が単独のケースについては Stata14 でも分析が可能でした．ERM でも同様に推定が実施でき内容が重複する部分があります．表 1 は，ERM コマンド（`eregress`, `eintreg`, `eprobit`, `eoprobit`）と既存コマンドの主な対応です．

表 1. 既存コマンドと ERM コマンドの対応

既存コマンド	ERM コマンド
内生変数を含む線形回帰	
`ivregress liml y1 x (y2 = z)`	`eregress y1 x, endogenous(y2 = z x)`
外生処置を含む線形回帰	
`teffects ra (y x1 x2) (t1)`	`eregress y x1 x2, extreat(t1) vce(robust)` `estat teffects`
内生処置を含む線形回帰	
`etregress y x, treat(t1 = x z)`	`eregress y x, entreat(t1 = x z, nointeract)`
サンプルセレクションを含む線形回帰	
`heckman y x, select(s1 x z)`	`eregress y x, select(s1 = x z)`

単純な考え方

ERM が持つ 3 要素について，概略と解決できることを整理します.

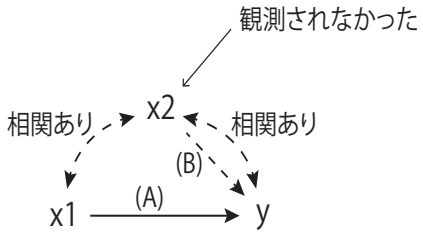
内生共変量

例として次のモデルを考えます。

$$\mathtt{y} = \beta_0 + \beta_1 \mathtt{x1} + \beta_2 \mathtt{x2} + e.\mathtt{y} \tag{1}$$

- y は連続アウトカム
- x1，x2 は説明変数（または共変量）
- $\beta_0, \beta_1, \beta_2$ は回帰係数
- e.y は誤差項

ここで，x2 にはアウトカム y と説明変数 x1 の両方との間に相関があり，この x2 が何らかの理由で観測できなかったとします．このとき，x2 はモデルに組み込むことができず，当然 (B) が推定できないだけでなく，x1 の y への影響 (A) が正しく推定されないという見過ごせない問題も生じます.



$$\mathtt{y} = \beta_0 + \beta_1 \mathtt{x1} + e.\mathtt{y} \tag{2}$$

上式 (2) で，前述の式 (1) から x2 が欠落したことにより，本来ならば x1 が e.y と相関するはずですが，式 (2) のみで推定をおこなうと相関が表れません．このことからも β_1 が正しく推定できないことが理解できると思います．このとき式 (2) のモデルにおいて x1 には内生性があると言われます.

内生共変量の生じる原因

x1 に内生性が生じる要因にはこのほか下記 2,3 などもあります.

1. x1 および y と相関のある変数が欠落した（観測できなかった）
2. x1 の測定に誤差が含まれた

3. x1 から y への因果関係に加えて，y から x1 への逆向きの因果関係があった

上記が一つでも該当するとき，x1 には内生性があり，内生共変量です.

内生共変量の解決策

内生共変量がある場合の解決法には,「アウトカムへの影響が内生共変量を介してのみ表れるような変数（操作変数)」を見つけ出しモデルに組み込む，という方法があります．具体的には，x1 との相関はあるが $e.\mathtt{y}$ との相関はないと考えられる変数（操作変数，仮に z1 とします）を見つけ出し，

$$\mathtt{x1} = \gamma_0 + \gamma_1 \mathtt{z1} + e.\mathtt{x1} \tag{3}$$

という新たな式をモデルに組み込みます．数式の解法としては，直感的な方法として，下記のように式 (3) の回帰係数 $\widehat{\gamma_0}$，$\widehat{\gamma_1}$ と回帰値 $\widehat{\mathtt{x1}}$ を求めて，メインのモデル式に代入する 2 段階最小二乗法があります.

$$\widehat{\mathtt{x1}} = \widehat{\gamma_0} + \widehat{\gamma_1}\mathtt{z1}$$

$$\mathtt{y} = \beta_0 + \beta_1 \widehat{\mathtt{x1}} + e.\mathtt{y} \tag{4}$$

上記で，z1 と $e.\mathtt{y}$ との相関が想定されない状況であれば β_1 は一致性を失うことなく推定できます．既に Stata15 以前のバージョンで搭載された ivregress で推定ができます.

```
. ivregress liml y (x1 = z1)
```

ERM コマンドでもほぼ同様の推定が行えます．ただし，2 段階最小二乗法とは異なる手法を用いるため，推定値が若干異なりますが，一致性が失われることはありません．詳細は，[ERM] マニュアル内の eregress の Methods and formulas セクションを参照ください.

```
. eregress y x1, endogenous(x1 = z1, nomain)
```

内生性のあるサンプルセレクション

内生サンプルセレクションの原因と解決策

前ページの例を引き続き使用します.

アウトカム y には欠損値が含まれたとします。y の欠損値がランダムでなく，観測できない何らかの要因の影響を受けたとしたら，モデルの誤差項について追加の考慮が必要かもしれません.

アウトカムが欠損値となった観測も含めて分析を行う方法の一つにサンプルセレクションモデルがあります.

新たな変数 selected を作成し，y が欠損値のときは selected=1，そうでないときは selected=0 という値を持たせます.

```
. generate selected = !missing(y)
```

この selected をたとえば z1 で説明できると考えられるとき，次のモデルを構築できます.

$$\texttt{selected} = \delta_0 + \delta_1 \texttt{z1} + e.\texttt{selected} \tag{5}$$

y が欠損値となるかどうかが z1 に左右された，というモデルです．そして，元のモデルを以下とします.

$$\texttt{y} = \beta_0 + \beta_1 \texttt{x1} + e.\texttt{y} \tag{6}$$

上式 (5) の e.selected が元のモデル式（6）の e.y と相関がないとき，サンプルセレクションは外生性となり，実は式 (5) をモデルに組み込まなくとも一致性を失わずに推定が行えます．問題は e.selected と e.y に相関があった場合です．このときサンプルセレクションは内生性となり，上式 (5) も含めたモデルで推定します.

ERM コマンドを用いると，内生サンプルセレクションを考慮した推定が次で実行できます.

```
. eregress y x1, select(selected = z1)
```

上記は既に Stata15 以前のバージョンで搭載された heckman でも同様の推定ができます.

```
. heckman y x1, select(selected z1)
```

内生サンプルセレクションの例

療養施設における健康増進キャンペーンを挙げます．ウォーキングプログラムを実施し，プログラム参加者には 6 か月のウォーキング実施後に健康状態の変化 y を 10 段階で自己申告により調査したとします．プログラムに参加しなかった人の y は欠損値とします．プログラムの参加／非参加は体重の測定値 (z1) など基礎項目が考慮されたとします.

完全に基礎項目のみでプログラム参加者を決定した場合，サンプルセレクションは外生となります．ただ，医師による問診で家族構成などを聞いて総合判断を下した場合など，観測が難しいところでの判断が働いた場合，サンプルセレクションは内生となり得ます．医師による判断は，ウォーキングプログラムで健康状態が改善しそうかどうかが目安なったことが十分考えられますから，プログラムの参加者はウォーキングの効果が出やすい人が集まった傾向があると考えられます．サンプルセレクションモデルを用いればそうした点を考慮した効果推定が行えます.

ERM コマンドでは，推定を次で実行できます.

```
. eregress y, select(selected = z1)
```

内生サンプルセレクションと内生共変量を共に含む例

たとえば体重の測定結果 z1 に大きな誤りがあったことが後で分かったとします．このとき，z1 の操作変数 iz1 が見つかると一致性を失うことなく推定が行えます．

ERM コマンドでは，次を実行することで，内生サンプルセレクションに内生共変量を組み合わせた推定ができます．

```
. eregress y, select(selected = z1) endogenous(z1 = iz1, nomain)
```

上記のように数式で書き表したときのような直観的なコマンド構文で実施できます．上記では内生共変量がサンプルセレクションモデルに含まれましたが，メインのモデルに含まれた場合にも ERM コマンドを用いることができます．既に Stata15 以前のバージョンで搭載された heckman や ivregress ではここまで簡単には実施できません．

処置効果の推定

シンプルな処置効果モデルを考えます．血圧 y を抑える処置をランダムに患者に割り付けた場合を考えます．$\texttt{x1}_i$ は患者 i の年齢とします．treated は 1 が処置あり，0 が処置なしとします．

$$\texttt{y}_i = \beta_0 + \beta_1\texttt{x1}_i + \beta_2\texttt{treated}_i + \beta_3\texttt{treated}_i\texttt{x1}_i + e.\texttt{y}_i \tag{7}$$

患者 i が処置なしであった場合の血圧 $\texttt{y}_i$，あるいは患者 i が処置ありであった場合に仮に処置なしであった場合の仮想の血圧 $\texttt{y}_i$ は以下のように書けます．

$$\texttt{y}_i = \beta_0 + \beta_1\texttt{x1}_i + e.\texttt{y}_i$$

患者 i が処置ありであった場合の血圧 $\texttt{y}_i$，あるいは患者 i が処置なしであった場合に仮に処置ありであった場合の仮想の血圧 $\texttt{y}_i$ は以下のように書けます．

$$\texttt{y}_i = \beta_0 + \beta_1\texttt{x1}_i + (\beta_2 + \beta_3\texttt{x1}_i) + e.\texttt{y}_i$$

処置効果は各患者における実際の測定値と反実仮想（counterfactual）の測定値の差となり，$\hat{\beta}_2 + \hat{\beta}_3\texttt{x1}_i$ となります．ATE はその平均となります．

ERM コマンドを用いると，推定が次で実行できます．

```
. eregress y x1, extreat(treated)
```

上記は既に Stata15 以前のバージョンで搭載された teffects を使用した次のコマンドでも同様の推定ができます．

```
. teffects ra (y x1) (treated)
```

内生性のある処置

サンプルセレクションの場合と同様，処置のモデル化を考えます．

$$\texttt{treated}_i = \eta_0 + \eta_1 \texttt{z2}_i + e.\texttt{treated}_i \tag{8}$$

元のモデル（アウトカムモデル）は先ほどと同じで式 (7) とします．サンプルセレクションの場合と同様，$e.\texttt{treated}_i$ が $e.\texttt{y}_i$ と相関しない場合，処置は外生性とされます．式 (8) をモデルに組み込む必要はありません．この場合に該当するケースとしては，処置がランダムに割り付けられたときのほか，処置モデル (式 (8)) で，処置の割り付けが完全に記述できる場合も含まれます．

ERM コマンドでは，外生処置のためのオプション `extreat()` 使用した次のようなコマンドで推定を実行できます．

```
. eregress y x1, extreat(treated)
```

$e.\texttt{treated}_i$ が $e.\texttt{y}_i$ と相関する場合，処置は内生性とされ，その相関を考慮しないと一致性のある処置効果は推定できません．式 (8) をモデルに組み込む必要があります．実際の場面では，処置をランダム割り付けできない場合がほとんどで，処置モデルを完全に記述できない事態も予測され，$e.\texttt{treated}_i$ と $e.\texttt{y}_i$ との相関を考慮したモデルのほうが広い応用範囲で包含的な推定が可能となります．

ERM コマンドでは，内生処置のためのオプション `entreat()` 使用した次のようなコマンドで推定を実行できます．

```
. eregress y x1, entreat(treated = z2)
```

内生サンプルセレクションの場合と同様，内生処置の場合も内生共変量が含まれた場合の推定が ERM コマンドで推定できます．

上記は既に Stata15 以前のバージョンで搭載された `etregress` でも同様の推定ができます．

```
. etregress y c.x1#i.treated, treat(treated = z2)
```

内生サンプルセレクションと同様，内生処置にも内生共変量が混在するケースも考えられ，これを ERM コマンドを使用して比較的簡単に推定を行うことができます．内生処置，内生共変量に加えて，さらに内生サンプルセレクションが混在するケースも ERM コマンドを使用して比較的簡単に推定を行うことができます．Stata15 以前のバージョンで搭載された `etregress`，`ivregress`，`heckman` では簡単には推定を実施できません。

実施例　－ eregress －

ERM コマンドのうち連続アウトカムを取扱う `eregress` について，架空のデータ[1]を用いたコマンド操作の結果とその解釈を掲載します．

[1]マニュアル『[ERM] Stata Extended Regression Models Reference Manual』の intro 8 に掲載されているデータセットです．

- データは架空の大企業 A で実施した減量プログラムの効果測定に関するものです.
- 減量プログラムは参加者自由とし，年度初めに参加者を集って 1 年度を通して様々なアクティビティを行いました．年度終わりに参加者自由の体重測定を行い，1 年間の体重の減少量（weightloss（単位は kg））を算出しました.
- 従業員は全員，年度初めに行われる健康診断で体重 (weight) を測定することになっており，年齢（age），性別（sex），健康志向（health）とともにプログラムの効果を推定しました.
- 健康志向（health）とは健康促進のアクティビティへの参加意欲を測定した結果とし，誤差なく測定できているとします.
- 減量プログラムの参加・不参加は wellpgm（参加は 1(yes)，不参加は 0(no)）が示すとします.
- 減量プログラム参加者ばかりでなく不参加者も含めて全員が体重測定を行ったとした場合の測定結果を weightloss0 とし，weightloss と区別します.

はじめに次のコマンドをコマンドウィンドウで実行して，架空データの読み込みを行います.

```
. use http://www.stata-press.com/data/r15/wellness
```

話を簡単にするために，次のように処置（wellpgm）と説明変数の交互作用を考えない簡素化したモデルで減量プログラムの効果を推定するとします.

$$\texttt{weightloss0}_i = \beta_0 + \beta_1 \texttt{age}_i + \beta_2 \texttt{sex}_i + \beta_3 \texttt{weight}_i + \beta_4 \texttt{health}_i + \beta_5 \texttt{wellpgm}_i + u_i$$

通常は処置効果が年齢や性別など個人の属性で変わるとされ，モデル式には交互作用項が入ることになりますが，今回はそれを考えません．$\texttt{wellpgm}_i$ の係数 β_5 が減量プログラムの効果となります.

```
. regress weightloss0 age i.sex weight health i.wellpgm
```

Source	SS	df	MS	Number of obs	=	545
				F(5, 539)	=	589.61
Model	2417.76071	5	483.552141	Prob > F	=	0.0000
Residual	442.044242	539	.820119187	R-squared	=	0.8454
				Adj R-squared	=	0.8440
Total	2859.80495	544	5.25699439	Root MSE	=	.9056

weightloss0	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0991644	.0038045	-26.06	0.000	-.1066378	-.0916909
sex						
male	-1.481883	.0937504	-15.81	0.000	-1.666044	-1.297722
weight	.1359547	.0054405	24.99	0.000	.1252676	.1466419
health	.4814308	.0255931	18.81	0.000	.4311564	.5317053
wellpgm						
yes	1.254928	.1076792	11.65	0.000	1.043406	1.46645
_cons	-3.754726	.4432054	-8.47	0.000	-4.625348	-2.884105

後の比較のために推定結果を保存します.

```
. estimates store true
```

結果から減量プログラムの効果は 1.25 と分かります．今回のモデルは簡素なので処置効果平均（ATE）も 1.25 です[2]．処置群の処置効果平均（ATET）も 1.25 です．減量プログラムに参加した従業員が仮に参加しなかった場合，年度終わりの体重が平均で 1.25 kg 重かったことが期待されます.

内生共変量

前述のように内生共変量が存在する原因は，観測されなかった要因がアウトカムと説明変数に影響を与えたことにあります．ここでは前述の (1) に該当する事案，すなわちアウトカム weightloss0 といくつかの説明変数との間に相関のある変数が観測されなかった状況をつくりだすため，先ほどのモデルで用いた health を故意に除外して考えます.

健康志向（health）がその他の説明変数，とりわけ減量プログラムへの参加／不参加（wellpgm）に対するある程度の影響があった可能性は十分考えられます．true で保存した先ほどのモデルから health を落として推定し、減量プログラムの効果の推定値が大きく変化するか確認します.

```
. regress weightloss0 age i.sex weight i.wellpgm
```

Source	SS	df	MS	Number of obs	=	545
				F(4, 540)	=	392.25
Model	2127.55956	4	531.88989	Prob > F	=	0.0000
Residual	732.24539	540	1.35600998	R-squared	=	0.7440
				Adj R-squared	=	0.7421
Total	2859.80495	544	5.25699439	Root MSE	=	1.1645

weightloss0	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0846678	.0047906	-17.67	0.000	-.0940783	-.0752572
sex						
male	-.866975	.1129843	-7.67	0.000	-1.088918	-.6450323
weight	.0746476	.0056015	13.33	0.000	.0636441	.085651
wellpgm						
yes	1.963876	.1297023	15.14	0.000	1.709093	2.218659
_cons	.0731151	.5062801	0.14	0.885	-.9214046	1.067635

```
. estimates store base
```

予想した通り減量プログラムの効果は 1.96 となり先ほどからは大きく離れた値を推定しました．先ほどのモデルが真だとすれば，今回の推定値は 95% 信頼区間に真の値 1.25 を含まない結果となりました．また，ほかの説明変数の係数推定値も変化しました.

[2]交互作用を含めた場合，ATE を推定するには eregress の実行後に estat teffects を実行します.

まず，年度初めの体重（weight）について，操作変数を用いることにより真の値を推定できないか考えます．体重（weight）が性別（sex）と減量プログラム開始以前の体育館の利用回数（月平均）（gym）という新たな変数の関数であると考え，内生共変量を含むモデルで推定することを考えます．weight は内生共変量，sex と gym は操作変数となります．sex，gym ともに年度間の体重減少量とは相関がないと考えています．とりわけ gym については，減量プログラムが行われる以前の体育館の利用回数であり，時間的なずれのあることを相関なしと考える大きな理由としています．推定には eregress を用います．

```
. eregress weightloss0 age i.sex i.wellpgm, endogenous(weight = i.sex gym)

Iteration 0:   log likelihood = -2754.3967
Iteration 1:   log likelihood = -2754.3958
Iteration 2:   log likelihood = -2754.3958
Extended linear regression                      Number of obs     =        545
                                                Wald chi2(4)      =    1567.81
Log likelihood = -2754.3958                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weightloss0						
age	-.085052	.0046453	-18.31	0.000	-.0941567	-.0759473
sex						
male	-1.485033	.1816514	-8.18	0.000	-1.841063	-1.129003
wellpgm						
yes	1.939382	.1258347	15.41	0.000	1.69275	2.186013
weight	.1402866	.0152444	9.20	0.000	.1104082	.1701651
_cons	-4.98968	1.205458	-4.14	0.000	-7.352335	-2.627025
weight						
sex						
male	9.540271	.6963573	13.70	0.000	8.175436	10.90511
gym	-.8164216	.0778396	-10.49	0.000	-.9689844	-.6638588
_cons	80.07947	.5413232	147.93	0.000	79.01849	81.14044
var(e.weig~0)	1.685296	.1887178			1.353192	2.098906
var(e.weight)	65.98361	3.997172			58.59652	74.30198
corr(e.wei~t, e.weightlo~0)	-.493591	.0790984	-6.24	0.000	-.6326146	-.3237501

```
. estimates store endog
```

weight の効果を操作変数を用いて推定した結果 0.140 となり，base で保存した先ほどのモデルの結果の 0.0746 よりも true で保存したモデルの結果 0.136 に近づきました．

結果の表を上から 4 つの部分に分けたときに 2 つ目にあたる weight からはじまる部分は，endogenous() で指定したモデルの推定結果です．性別（sex）が男性 (male) であると年度初めの体重（weight）が 9.54 kg 高いことの分析結果が読み取れます．また，減量プログラム前の月平均体育館利用回数（gym）が 1 回増えると逆に 0.816 kg 低いとの分析結果が読み取れます．これら 2 つの係数推定値は p 値（P>|z| 列の値）が 0.000（0.001 以下）で有意であることが分かり，2 変数の変化により weight の値が変化していると考えてよいことが分かります．

4 つ目にある corr(e.wei~t, e.weightlo~0) は e.weight と e.weightloss0 との共分散です．共分散を標準化すると相関係数となります．共分散の推定値が -0.494 であり，p 値が 0.000（0.001 以下）であることから，共分散が 0 から有意に離れていることが分かり，weight には内生性があった，すなわち非観測な要因が weight と weightloss0 とに影響を与えていたことが分かります．もちろんこれは true モデルで重要だった説明変数 health を除外した影響が含まれているはずなので，当然の結果と言えます．

内生処置

処置効果は，アウトカムを処置群（処置ありのグループ）と対照群（処置なしのグループ）との平均差から求めるのではなく，反実仮想（counterfactual）を導入した前述の枠組みで推定します．

処置の割り付けに非観測な要因が影響し，同時にメインモデルの誤差項とも相関を持つような場合，処置には内生性があり，それを考慮した上で一致性のある効果推定を行う必要があることは前述の通りです．ここでは処置モデルとして，減量プログラムへの参加（wellpgm）が年齢（age）と喫煙者／非喫煙者（smoke，1 が喫煙者）の関数だというモデルを構築します．年齢（age）が減量プログラムへの参加（wellpgm）の傾向を左右する，また喫煙者（smoke）かどうかも減量プログラムへの参加傾向を左右すると考えました．

先ほどの eregress コマンドに対し entreat() オプションで処置モデルを組み込みます．モデルは処置に関して簡素化してあり，アウトカムモデルにおいて処置と共変量を考えないので、entreat(... , nointeract) としてオプションを用います．コマンドをウィンドウに打ち込むときは *Enter* キーで改行せず自動で行われる右への折り返しに改行を委ねます．

```
. eregress weightloss0 age i.sex i.wellpgm, endogenous(weight = i.sex gym)
  entreat(wellpgm = age i.smoke, nointeract)

Iteration 0:   log likelihood = -2948.2131
Iteration 1:   log likelihood =      -2948
Iteration 2:   log likelihood = -2947.9973
Iteration 3:   log likelihood = -2947.9973
Extended linear regression                      Number of obs     =        545
                                                Wald chi2(4)      =    1264.15
Log likelihood = -2947.9973                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weightloss0						
age	-.1027296	.0067867	-15.14	0.000	-.1160312	-.089428
sex						
male	-1.476728	.1772287	-8.33	0.000	-1.82409	-1.129367
wellpgm						
yes	1.154828	.249413	4.63	0.000	.6659874	1.643669
weight	.1386038	.0148817	9.31	0.000	.1094361	.1677715
_cons	-3.735122	1.21943	-3.06	0.002	-6.125161	-1.345083
wellpgm						

```
                 age |  -.0943764   .0073377   -12.86   0.000    -.1087581   -.0799948

               smoke |
                 yes |  -1.532311   .1770475    -8.65   0.000    -1.879317   -1.185304
               _cons |    4.27193   .3393742    12.59   0.000     3.606769    4.937092
---------------------+------------------------------------------------------------------
weight               |
                 sex |
                male |   9.539291     .69535    13.72   0.000      8.17643    10.90215
                 gym |  -.8206946   .0778126   -10.55   0.000    -.9732045   -.6681846
               _cons |   80.09352   .5409731   148.05   0.000     79.03323    81.15381
---------------------+------------------------------------------------------------------
var(e.weig~0)        |   1.762869   .1944185                      1.420185     2.18824
var(e.weight)        |   65.98398   3.997216                      58.5968     74.30244
---------------------+------------------------------------------------------------------
corr(e.wel~m,        |
e.weightlo~0)        |   .4679942   .1101397     4.25   0.000     .2270628    .6549438
corr(e.wei~t,        |
e.weightlo~0)        |  -.4745619   .0781291    -6.07   0.000    -.6129184   -.3079701
corr(e.wei~t,        |
   e.wellpgm)        |   -.087583   .0690664    -1.27   0.205    -.2205384    .0485677
------------------------------------------------------------------------------
```

```
. estimates store entrt
```

処置効果は 1.15 であり，base モデルの値 1.96 よりも true モデルの値 1.25 に近い結果となりました．内生共変量の議論と同様，true モデルで重要だった説明変数 health を除外した影響が含まれているはずなので，当然の結果と言えます．

内生サンプルセレクション

減量プログラム後の体重測定は従業員全員が行ったのでなく，自主的に測定に出向いた従業員のみで行われました．体重減少量を示す変数 weightloss には欠損値が含まれ，測定を行わなかった従業員のデータであることを表しています．データセットには，仮に全員が測定を行ったとした場合の体重減少量 weightloss0 とは別に存在し，欠損値以外については両者の値は一致しています．まず，先ほどのコマンドでアウトカムのみを変更し，体重測定を行った従業員のデータのみで推定をやり直します．

```
. eregress weightloss age i.sex i.wellpgm, endogenous(weight = i.sex gym)
 entreat(wellpgm = age i.smoke, nointeract)

Iteration 0:   log likelihood = -1822.7501
Iteration 1:   log likelihood = -1822.5792
Iteration 2:   log likelihood = -1822.5784
Iteration 3:   log likelihood = -1822.5784
Extended linear regression                      Number of obs     =        337
                                                Wald chi2(4)      =     647.44
Log likelihood = -1822.5784                     Prob > chi2       =     0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
weightloss   |
         age |  -.1043032   .0084798   -12.30   0.000    -.1209232   -.0876832
```

sex						
male	-1.599589	.2078047	-7.70	0.000	-2.006878	-1.192299
wellpgm						
yes	.8355675	.2870002	2.91	0.004	.2730575	1.398078
weight	.1415195	.0171767	8.24	0.000	.1078538	.1751853
_cons	-3.488841	1.441393	-2.42	0.016	-6.31392	-.6637625
wellpgm						
age	-.0881761	.0086047	-10.25	0.000	-.1050411	-.0713111
smoke						
yes	-1.463798	.2045569	-7.16	0.000	-1.864722	-1.062874
_cons	4.1675	.3927992	10.61	0.000	3.397628	4.937372
weight						
sex						
male	9.631966	.8492097	11.34	0.000	7.967545	11.29639
gym	-.8661298	.0921124	-9.40	0.000	-1.046667	-.6855928
_cons	79.85692	.6620915	120.61	0.000	78.55924	81.15459
var(e.weig~s)	1.874868	.2381447			1.461678	2.404861
var(e.weight)	60.85603	4.688359			52.32713	70.77508
corr(e.wel~m, e.weightloss)	.6209647	.1078541	5.76	0.000	.3648902	.7896991
corr(e.wei~t, e.weightloss)	-.4535506	.089826	-5.05	0.000	-.6111862	-.2613065
corr(e.wei~t, e.wellpgm)	-.1387765	.0844177	-1.64	0.100	-.298965	.0290187

```
. estimates store samplebase
```

entrt として保存した直前の推定との間で結果を比較してみます.

```
. estimates table entrt samplebase, stats(N) equations(1) keep(#1:)
```

Variable	entrt	samplebase
age	-.10272957	-.10430319
sex		
male	-1.4767285	-1.5995888
wellpgm		
yes	1.154828	.83556752
weight	.13860382	.14151952
_cons	-3.7351221	-3.488841
N	545	337

減量プログラムの効果（wellpgm の係数）を見ると，1.15 から 0.836 へ変化しており，true モデルの 1.25 からは点推定値が遠ざかりました．最終行の N（観測数）の値から，推定に使用した観測の数が確かに 545 から 337 に変化しており，目的通りの推定を行えたことが分かります.

次にこの 337 の観測における推定から，母集団における推定へ対象を変化させることを考えます．欠損値が完全にランダムでない場合，samplebase で保存した 337 の観測における推定結果は母集団における推定とは異なる可能性があります．weightloss が欠損値かどうかは別の変数 completed（0 が欠損値）でも知ることができます．completed を用いてサンプルセレクションのモデルを次のように構築します．

```
completed = i.wellpgm experience i.salaried
```

新出の変数について，experience は勤続年数を，salaried はフルタイム労働かパートタイム労働か（1 がフルタイム）を示します．減量プログラムへの参加，勤続年数の長さ，それにフルタイム労働かどうかの 3 要因が年度終わりの体重測定の参加傾向を左右する，と考えました．また，3 要因以外にも観測されない要因が体重測定参加と体重減少量の両方に影響を与えていると考え，内生サンプルセレクションモデルを組み込んで推定を行います．

```
. eregress weightloss age i.sex i.wellpgm, endogenous(weight = i.sex gym)
 entreat(wellpgm = age i.smoke, nointeract)
 select(completed = i.wellpgm experience i.salaried)

Iteration 0:   log likelihood = -2830.5178
Iteration 1:   log likelihood = -2807.3314
Iteration 2:   log likelihood = -2802.0802
Iteration 3:   log likelihood = -2800.8729
Iteration 4:   log likelihood = -2800.8319
Iteration 5:   log likelihood = -2800.8318
Extended linear regression                      Number of obs     =        545
                                                      Selected    =        337
                                                      Nonselected =        208
                                                Wald chi2(4)      =     749.04
Log likelihood = -2800.8318                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weightloss						
age	-.1114998	.0083531	-13.35	0.000	-.1278715	-.0951281
sex						
male	-1.560765	.2062746	-7.57	0.000	-1.965056	-1.156474
wellpgm						
yes	.9246275	.2750269	3.36	0.001	.3855848	1.46367
weight	.14354	.0175073	8.20	0.000	.1092263	.1778537
_cons	-3.679888	1.464123	-2.51	0.012	-6.549515	-.81026
completed						
wellpgm						
yes	.6553902	.2263862	2.90	0.004	.2116814	1.099099
experience	-.8153984	.0617977	-13.19	0.000	-.9365196	-.6942772
salaried						
yes	.4709859	.1419878	3.32	0.001	.192695	.7492768
_cons	4.902936	.3973849	12.34	0.000	4.124076	5.681796
wellpgm						
age	-.0938617	.0072734	-12.90	0.000	-.1081173	-.079606

smoke						
yes	-1.477078	.1772103	-8.34	0.000	-1.824404	-1.129752
_cons	4.228481	.337379	12.53	0.000	3.56723	4.889732
weight						
sex						
male	9.506396	.6960864	13.66	0.000	8.142091	10.8707
gym	-.8184902	.0779351	-10.50	0.000	-.9712401	-.6657402
_cons	80.10245	.5407952	148.12	0.000	79.04251	81.16239
var(e.weig~s)	2.015328	.263477			1.559777	2.603927
var(e.weight)	65.98395	3.997213			58.59678	74.30241
corr(e.com~d, e.weightloss)	.5434105	.0824836	6.59	0.000	.362338	.6849556
corr(e.wel~m, e.weightloss)	.5878321	.1054098	5.58	0.000	.3440372	.7573749
corr(e.wei~t, e.weightloss)	-.4801763	.089175	-5.38	0.000	-.6353685	-.2877017
corr(e.wel~m, e.completed)	.3753168	.1523364	2.46	0.014	.0470351	.6304273
corr(e.wei~t, e.completed)	-.0643813	.0718768	-0.90	0.370	-.2030702	.0768401
corr(e.wei~t, e.wellpgm)	-.096324	.0691411	-1.39	0.164	-.2292586	.0401382

```
.  estimates store endsel
```

推定の結果，処置効果は 0.925 となり，さきほどのの samplebase モデルの 0.836 よりは true モデルの値 1.25 に近い推定結果となりました.

株式会社　ライトストーン
2017 年 6 月