Lasso 変数選択を利用した処置効果モデル

`telasso` コマンドは、lasso を使用して潜在的な制御変数を選択しながら、観測データから平均治療効果(ATE)、平均治療効果(ATET)、および潜在的アウトカム平均(POM)を拡張逆確率重み付け(AIPW)によって推定します。

`telasso` は次の 5 段階で行われます。

1. Lasso を使用して、各処置レベルのアウトカムモデルの変数を選択します。
2. ステップ 1 で選択した変数に基づいて、各処置レベルの結果の個別の回帰モデルに適合し、各標本の処置固有の予測結果を取得します。
3. Lasso を使用して、処置モデルの変数を選択します。
4. ステップ 3 で選択した変数に基づいて、処置モデルのパラメータを推定し、逆確率の重み(IPW)を計算します。
5. 処置固有の予測結果の加重平均を計算します。ここで、加重はステップ 4 で計算された逆確率の加重です。これらの加重平均の対比により、ATE の推定値が得られます。

ステップ 1 と 3 は、それぞれアウトカムモデルと処置モデルのモデル選択を実行します。選択した変数を使用して、手順 2、4、および 5 でモーメント条件を構築し、ATE を推定します。得られた推定量は、CI、オーバーラップ、および独立した同一分布の仮定の下で一貫しています。推論は、ステップ 1 と 3 で発生する可能性のある軽度のモデル選択ミスに対してロバストです。この推定量は、ステップ 5 で使用される二重ロバストモーメント条件により、アウトカムまたは処置モデルのいずれかでのモデルの設定ミスに対しても頑健です。

Lasso 共変量選択による ATE の推定

最初に、2 種類の肺移植を比較する例で telasso を説明します。両側肺移植(BLT)は通常、手術後の短期間の死亡率が高くなりますが、片肺移植(SLT)と比較して生活の質が大幅に改善されます。 結果として、これら 2 つの治療オプションのどちらかを選択する必要がある患者にとって、生活の質に対する BLT(対 SLT)の影響を知ることは不可欠です。個人の 1 秒間の努力呼気量(FEV1)に基づいて、生活の質を測定できます。Koch、Vock、Wolfson (2018)に触発された架空のデータセット(lung.dta)があります。アウトカム(fev1p)は、手術の 1 年後に測定された FEV1%です。FEV1%は、同様の特性を持つ健康な人と比較した患者の FEV1 の割合です。治療変数(トランスタイプ)は、治療が

BLT か SLT かを示します。

サンプルデータセットをダウンロードして、内容を確認します。

```
use https://www.stata-press.com/data/r17/lung, clear
describe * , short
```

```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
agep            byte    %10.0g                Patient age (years)
bmip            double  %10.0g                Patient body mass index
diabetesp       byte    %12.0g     lbdiab     Patient diabetes status
heightp         double  %10.0g                Patient height (cm)
o2amt           double  %10.0g                Oxygen delivered
karn            byte    %8.0g      lbyes      Karnofsky score > 60
lungals         double  %10.0g                Lung allocation score
racep           byte    %8.0g      lbrace     Patient race
sexp            byte    %8.0g      lbsex      Patient gender
lifesvent       byte    %8.0g      lbyes      Life support ventilator needed
assisvent       byte    %8.0g      lbyes      Assisted ventilation needed
centervol       double  %10.0g                Center volume
walkdist        double  %10.0g                Walking distance in 6 minutes
o2rest          byte    %8.0g      lbyes      Oxygen needed at rest
aged            byte    %10.0g                Donor age (years)
raced           byte    %8.0g      lbrace     Donor race
bmid            double  %10.0g                Donor body mass index
smoked          byte    %8.0g      lbyes      Donor if has history of smoking
cmv             byte    %8.0g      lbyes      Positive cytomegalovirus test
deathcause      byte    %8.0g      lbyes      Cause of death - traumatic brain injury
diabetesd       byte    %12.0g     lbdiab     Donor diabetes status
expandd         byte    %8.0g      lbyes      Expanded donor needed
heightd         double  %10.0g                Donor height (cm)
sexd            byte    %8.0g      lbsex      Donor gender
distd           int     %10.0g                Donor to treatment center distance
lungpo2         double  %10.0g                Lung PO2
lungalloc       byte    %8.0g      lballo     Lung allocation status
hratio          double  %10.0g                Height ratio
ischemict       double  %10.0g                Ischemic time
genderm         byte    %19.0g     lbgm       Matching gender status
racem           byte    %17.0g     lbrm       Matching race status
transtype       byte    %8.0g      lbtau      Lung transplant type
fev1p           double  %10.0g                Percentage of predicted value of FEV1
```

31 の変数は、患者とドナーのいくつかの特性を測定します。制御変数を構築するには、これらの変数とそれらの間の相互作用を使用したいと考えています。連続変数とカテゴリ変数を区別するために、これらの変数名を 1 つずつ入力するのは面倒です。

まず、`vl set` を使用して、変数を連続変数とカテゴリ変数に自動的に分割します。グローバルマクロ `$vlcategorical` にはすべてのカテゴリ変数名が含まれ、`$vlcontinuous` にはすべての連続変数名が含まれます。

```
vl set
```

Macro	Macro's contents	
	# Vars	Description
System		
$vlcategorical	18	categorical variables
$vlcontinuous	13	continuous variables
$vluncertain	2	perhaps continuous, perhaps categorical variables
$vlother	0	all missing or constant variables

Notes

1. Review contents of **vlcategorical** and **vlcontinuous** to ensure they are correct. Type **vl list vlcategorical** and type **vl list vlcontinuous**.
2. If there are any variables in **vluncertain**, you can reallocate them to **vlcategorical**, **vlcontinuous**, or **vlother**. Type **vl list vluncertain**.
3. Use **vl move** to move variables among classifications. For example, type **vl move (x50 x80) vlcontinuous** to move variables **x50** and **x80** to the continuous classification.
4. *vlnames* are global macros. Type the *vlname* without the leading dollar sign (**$**) when using **vl** commands. Example: **vlcategorical** *not* **$vlcategorical**. Type the dollar sign with other Stata commands to get a *varlist*.

次に、vl create を使用して、カスタマイズされた変数リストを作成します。具体的には、$cvars にはアウトカム(fev1p)を除くすべての連続変数が含まれ、$fvars は処置(transtype)を除くすべてのカテゴリ変数で構成されます。最後に、vl sub は、グローバルマクロ$allvars を、$cvars の連続変数と$fvars のカテゴリ変数の間の完全な 2 次相互作用に置き換えます。アウトカムモデルと処置モデルの両方の制御変数として $allvars を使用します。

```
vl create cvars = vlcontinuous - (fev1p)
vl create fvars = vlcategorical - (transtype)
vl sub allvars = c.cvars i.fvars c.cvars#i.fvars
```

これで、telasso を使用して ATE を推定する準備が整いました。telasso のデフォルトである線形結果モデルとロジット処理モデルを想定しています。

```
telasso (fev1p $allvars) (transtype $allvars)
```

```
Estimating lasso for outcome fev1p if transtype = 0 using plugin method ...
Estimating lasso for outcome fev1p if transtype = 1 using plugin method ...
Estimating lasso for treatment transtype using plugin method ...
Estimating ATE ...

Treatment-effects lasso estimation    Number of observations      =        937
Outcome model:   linear               Number of controls          =        454
Treatment model: logit                Number of selected controls =          8
```

fev1p	Coefficient	Robust std. err.	z	P>|z|	[95% conf. interval]	
ATE						
transtype						
(BLT vs SLT)	37.51841	.1606703	233.51	0.000	37.20351	37.83332
POmean						
transtype						
SLT	46.4938	.2021582	229.99	0.000	46.09757	46.89002

すべての患者が BLT を選択した場合の fev1p は、すべての患者が SLT を選択した場

合に予想される平均 46%よりも 38%高いと予想されます。454 の制御変数のうち、telasso は 8 つだけを選択します。アウトカムモデルと処置モデルの両方のモデル選択を要約するには、lassoinfo を使用できます。

```
lassoinfo
```

```
    Estimate: active
     Command: telasso
```

Variable	Model	Selection method	lambda	No. of selected variables
fev1p				
transt~e ~0	linear	plugin	.2239121	5
transt~e ~1	linear	plugin	.1986153	6
transtype	logit	plugin	.0748279	3

アウトカム、処置変数などの lasso によって、どの変数が選択されているかを確認したい場合は、lassocoef を使用できます。fev1p のアウトカムに関連する lasso が 2 つあることに注意してください。1 つは処置の transtype が 0 の場合のアウトカム fev1p 用で、もう 1 つは transtype が 1 の場合の fev1p 用です。対照的に、transtype の処置では、lasso は 1 つしかありません。したがって、オプション for()を使用して、変数の lasso を指定するだけで済みます。

```
lassocoef (., for(fev1p) tlevel(0)) (., for(fev1p) tlevel(1)) (., for(transtype))
```

	fev1p(0)	fev1p(1)	transtype
heightp	x	x	
centervol	x	x	
walkdist	x	x	x
lungpo2	x	x	x
diabetesd#c.lungpo2			
0	x		
diabetesp#c.walkdist			
0		x	
assisvent#c.walkdist			
0		x	
ischemict			x
_cons	x	x	x

```
Legend:
  b - base level
  e - empty cell
  o - omitted
  x - estimated
```

3 つすべてのモデルにおいて、lungpo2 が選択されています。heightp, centervol,

0.diabetesd#c.lungpo2, 0.diabetesp#c.walkdist, 0.assisvent#c.walkdist および ischemict は 1 つまたは 2 つのモデルでのみ選択されています。

チューニングパラメータの変更

デフォルトでは、telasso はプラグイン法を使用して、lasso でチューニングパラメータ λ を選択します。BIC、交差検証、または適応型 lasso を使用して、最適なλを選択することもできます。ここでは、selection(bic)オプションを追加して、BIC を最小化するλを選択させます。

```
telasso (fev1p $allvars) (transtype $allvars), selection(bic)
```

```
Estimating lasso for outcome fev1p if transtype = 0 using BIC ...
Estimating lasso for outcome fev1p if transtype = 1 using BIC ...
Estimating lasso for treatment transtype using BIC ...
Estimating ATE ...

Treatment-effects lasso estimation    Number of observations      =        937
Outcome model:   linear               Number of controls          =        454
Treatment model: logit                Number of selected controls =         18
```

fev1p	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
transtype (BLT vs SLT)	37.54872	.2222001	168.99	0.000	37.11322	37.98423
POmean						
transtype SLT	46.44739	.2282797	203.47	0.000	45.99997	46.89481

結果の解釈は、前述と同じように行うことができます。ただし、454 の調整変数のうち、BIC を使用した telasso は 18 を選択します。これはプラグイン法より大きくなっています。

ATET の推定

ATET を推定する場合は、atet オプションを追加するだけです。次のように入力します。この結果からは、BLT を受けた患者では SLT に比べて約 36%アウトカムが高くなります。

```
telasso (fev1p $allvars) (transtype $allvars), atet
```

```
Treatment-effects lasso estimation    Number of observations      =        937
Outcome model:   linear               Number of controls         =        454
Treatment model: logit                Number of selected controls =         8
```

fev1p	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
ATET transtype (BLT vs SLT)	35.78157	.1831478	195.37	0.000	35.42261	36.14053
POmean transtype SLT	43.35214	1.268976	34.16	0.000	40.86499	45.83929

高次セミパラメトリックモデル

交絡因子を制御するためにいくつかの変数が不可欠であることを理論が示唆することもありますが、モデルの関数形式については言及されていません。次の例では、telassoを使用して高次元セミパラメトリックモデルでATEを推定する方法を説明します。Chernozhukov and Hansen (2004)によって報告されたデータを使用して、純金融資産(資産)に対する401(k)適格性(e401k)の影響を推定したいと考えています。これらのデータは、1990 年の収入とプログラムへの参加に関する調査の世帯のサンプルからのものです。データセットは家長の収入(income)、年齢(age)、教育年数(educ)、年金受給者かどうか(pension)、婚姻状況(married)、およびIRAに参加しているかどうか(ira)が情報として含まれています。

金融資産に対する 401(k)資格の影響を判断する際の懸念事項の 1 つは、401(k)プランを提供する会社で働くことを選択することが無作為に割り当てられるわけではないことです。この問題を克服するために、Poterba, Venti and Wise (1994, 1995)は、収入を条件付けした後、401(k)を提供する会社で働くことを外生性と見なすことができると提案しています。

収入、年齢、教育を共変量としてモデルに含めたいと考えていますが、それらが線形的にモデルに入ると仮定したくはありません。より柔軟なモデルが必要なため、mksplineを使用して、収入、年齢、および教育のデータの等間隔またはパーセンタイルでノットを持つ線形スプラインを作成します。生成された変数は、セミパラメトリックモデルを形成するために使用されます。これらの変数には、後で簡単に参照できるように共通の接頭辞「_b」があります。

```
use https://www.stata-press.com/data/r17/assets, clear
mkspline _bincome 5 = income
mkspline _bage 5 = age
mkspline _beduc 5 = educ
```

生成された線形スプライン変数を記述することで見ることができます。収入、年齢、教育の各変数に対して、5 つの変数が生成されていることがわかります。

```
describe _b*
```

```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
_bincome1       float   %9.0g                 income: (.,48424.8)
_bincome2       float   %9.0g                 income: (48424.8,96849.6)
_bincome3       float   %9.0g                 income: (96849.60000000001,145274.4)
_bincome4       float   %9.0g                 income: (145274.4,193699.2)
_bincome5       float   %9.0g                 income: (193699.2,.)
_bage1          float   %9.0g                 age: (.,32.8)
_bage2          float   %9.0g                 age: (32.8,40.6)
_bage3          float   %9.0g                 age: (40.59999999999999,48.4)
_bage4          float   %9.0g                 age: (48.39999999999998,56.2)
_bage5          float   %9.0g                 age: (56.19999999999997,.)
_beduc1         float   %9.0g                 educ: (.,4.4)
_beduc2         float   %9.0g                 educ: (4.4,7.8)
_beduc3         float   %9.0g                 educ: (7.800000000000001,11.2)
_beduc4         float   %9.0g                 educ: (11.2,14.6)
_beduc5         float   %9.0g                 educ: (14.6,.)
```

次に、カテゴリ変数(pension, married, および ira)と生成されたスプライン変数(_b*)を相互作用させて、コントロール変数を定義します。グローバルマクロ$controlsには、定義された制御変数が含まれており、アウトカムと処置モデルの両方で使用します。

```
global vars c.(_b*) i.(pension married ira)
global controls $vars (($vars)#($vars)
```

これで、telassoを使用して、純金融資産に対する 401(k)適格性の ATE を推定する準備が整いました。線形ノンパラメトリック系列を使用して、アウトカムモデルを近似します。さらに、ロジットノンパラメトリック系列を使用して処置モデルを近似します。グローバル マクロ$controlsは、ノンパラメトリック系列を定義します。

```
telasso (assets $controls) (e401 $controls)
```

```
Estimating lasso for outcome assets if e401k = 0 using plugin method ...
Estimating lasso for outcome assets if e401k = 1 using plugin method ...
Estimating lasso for treatment e401k using plugin method ...
Estimating ATE ...

Treatment-effects lasso estimation    Number of observations      =      9,913
Outcome model:   linear               Number of controls          =        243
Treatment model: logit                Number of selected controls =         26
```

assets	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
e401k (Eligible vs Not eligible)	7850.285	1092.952	7.18	0.000	5708.139	9992.431
POmean						
e401k Not eligible	13893.49	785.1038	17.70	0.000	12354.72	15432.27

すべての労働者が 401(k)プランのある企業で働いている場合の純金融資産は、すべての労働者が 401(k)プランのない企業で働いている場合に予想される平均の 13,839 ドルよりも 7,850 ドル多いと予想されます。

診断

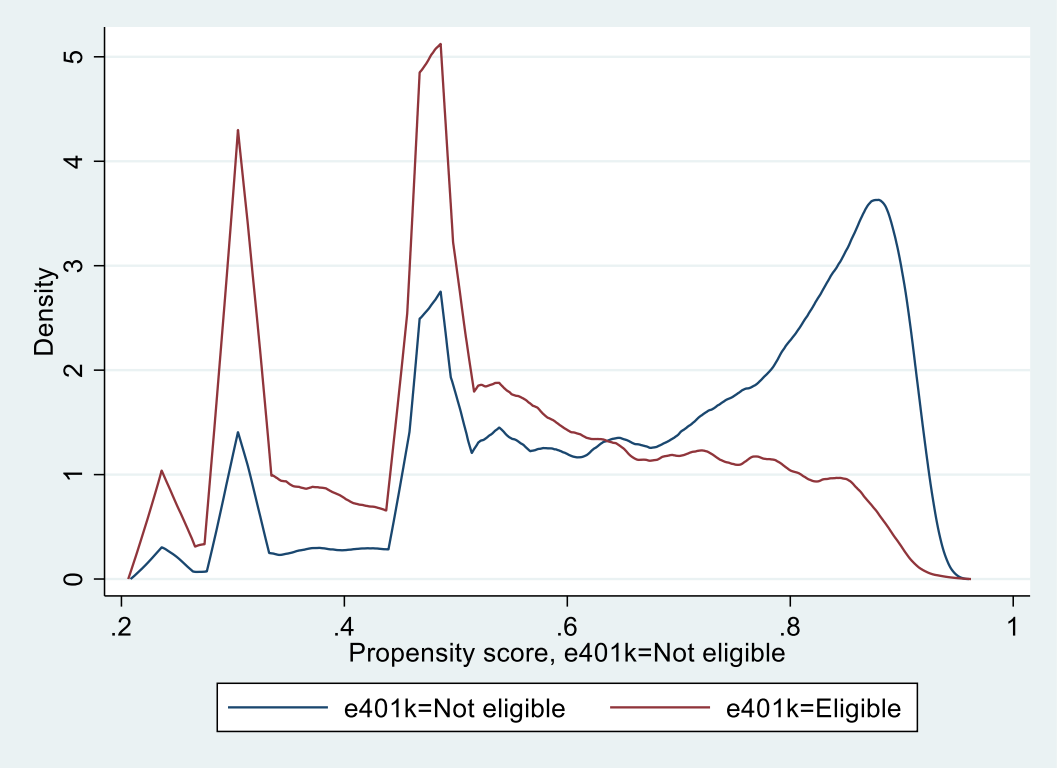
Lasso を利用した処置効果モデルは、名前の通り lasso 変数選択と従来の処置効果モデルを組み合わせたものですので、通常の処置効果モデルおよび lasso 推定と同様の推定後のモデル診断機能が利用できます。

処置効果モデルの診断機能

オーバーラップの仮定

オーバーラップ条件を確認するには、処置群と統制群でグラフを作成し、オーバーラップの仮定に違反しているかどうかを確認できます。共変量の各組み合わせで、対照群と処置群の両方で観測値が見られる可能性がある場合、オーバーラップ仮定は満たされます。`telasso` を実行した後、次を実行します。

```
teoverlap
```



得られたグラフでは、両群のピークの位置は、およそ重なっているように見えます。

バランスチェック

ATE を推定した後、マッチしたデータに偏りがないかを確認します。`tebalance summarize` は、推定後に処置群全体の共変量のバランスをチェックするために使用されるレポートを表示します。

```
tebalance summarize
```

Covariate balance summary

	Raw	Weighted
Number of obs =	9,913	9,913.0
Treated obs =	3,682	4,887.2
Control obs =	6,231	5,025.8

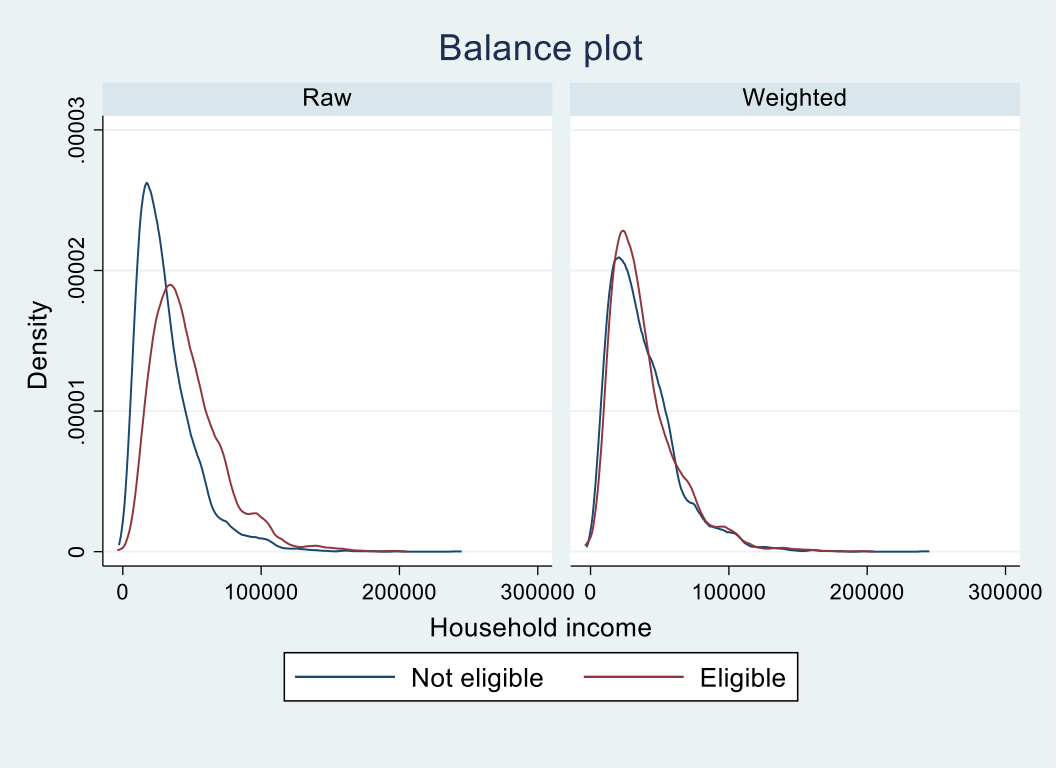
	Standardized differences		Variance ratio	
	Raw	Weighted	Raw	Weighted
_bincome1	.750948	.0482659	.7243709	.8794933
pension				
No pension	-.5207834	-.0116358	1.610504	1.011534
_bincome1# _bage1	.7577255	.0468209	.7530602	.8881179
_bincome1# _beduc3	.7181144	.0477066	.7563269	.8856304
pension# married				
No pension#N~d	-.4875712	-.0248925	.5221977	.9721957
Receives pen~d	.2649014	.0040213	1.729845	1.008399
pension# ira				
No pension#No	-.5387541	-.0201817	1.070546	1.004432

マッチした標本の結果は、共変量のバランスが取れていることを示しています。Standardized difference 列に表示される両群の差はすべてゼロに近く、Variance ratio 列の分散日はすべて 1 に近くなっていることがわかります。

共変量ごとの密度

tebalance density コマンドは推定後に共変量のバランスをチェックするために使用されるカーネル密度プロットを生成します。変数 income を調査するには次のように入力します。

```
tebalance density income
```



一致したサンプルの密度プロットはほとんど見分けがつきません。これは、推定された AIPW でのマッチングが共変量のバランスをとっていることを意味しています。

Lasso の診断機能

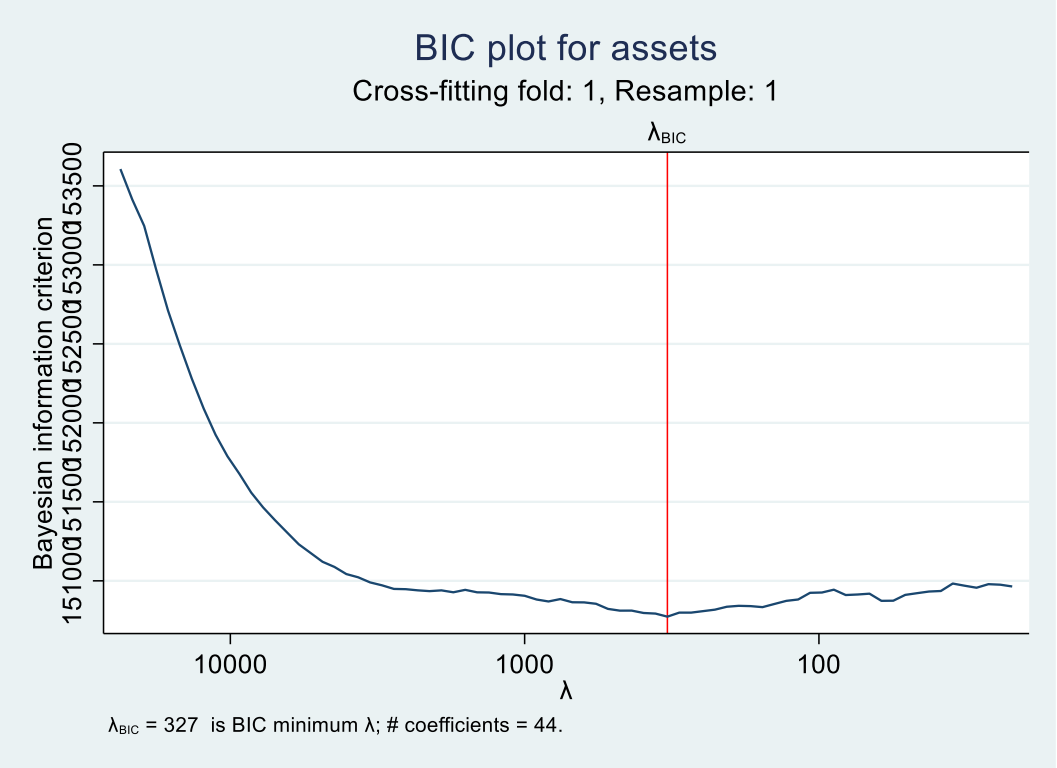
ここで紹介する CV/BIC 関数の変化グラフと係数のパスプロットは、罰則パラメータλの選択方法に相互検証、適用型 lasso またはベイズ情報規準(BIC)のいずれかを指定した際にのみ有効です。`telasso` コマンドでは、デフォルトでプラグイン推定によりλを選択することに注意してください。ここでは、先の推定コマンドに `selection(bic)`オプションを追加して、BIC を基準にλを選択させます。推定結果は省略します。

```
telasso (assets $controls) (e401 $controls), selection(bic)
```

ハイパーパラメータの変化のグラフ

罰則パラメータλの変化に応じて CV/BIC 関数がどのように変化するかについてに cvplot, `bicplot` コマンドを使用できます。ここでは、処置変数 `e401k` が 1 の場合のアウトカム `assets` に対する `lasso` の BIC プロットを示します。`bicplot` では、オプション `for()`および `tlevel()`を使用して、指定された処置レベルでアウトカム変数の lasso を参照します。

```
bicplot, for(assets ) tlevel(0)
```

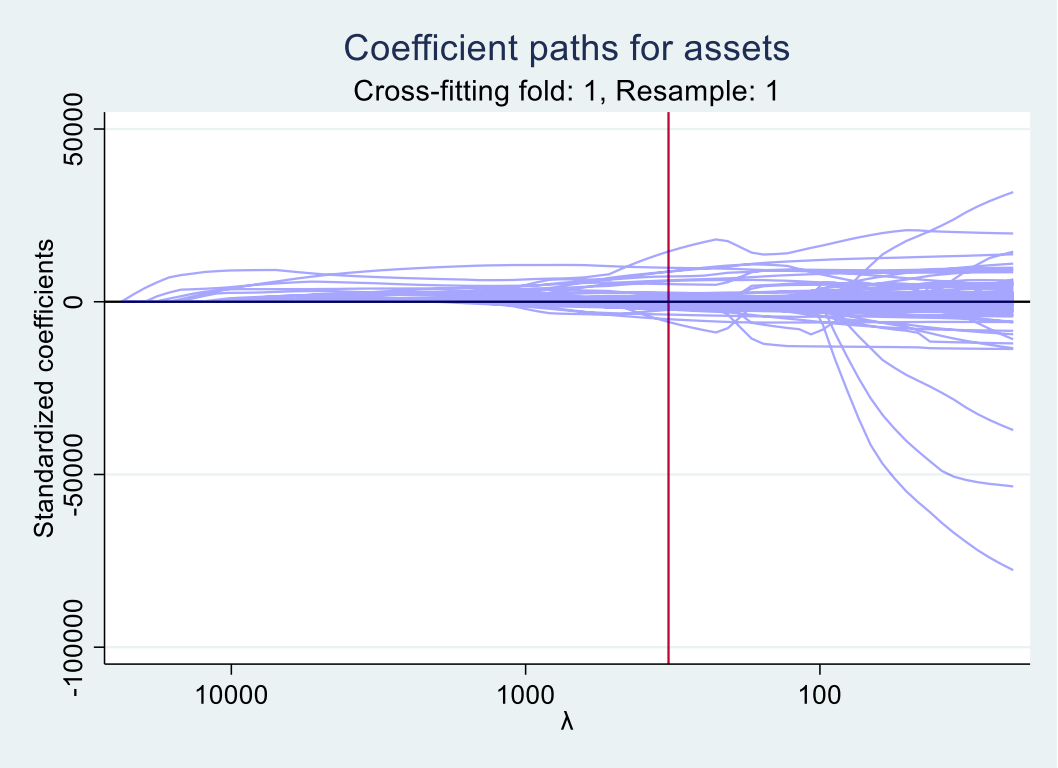


グラフから$\lambda = 327$で BIC が最小化され、44 の共変量が選定されていることがわかります。

係数のパスのグラフ

係数パスのプロットは、lasso の罰則係数の探索グリッドにおける、各係数のパスを示します。`coefpath` コマンドで係数のパスをプロットします。先述の `lassoinfo, lassocoef` コマンドと同様に for()でアウトカムまたは処置変数を選択します。アウトカム変数を指定する場合、さらに処置レベルを指定しなければなりません。ここでは、探索されているλの値が大きいため、`xunits(rlnlambda)`オプションで x 軸を対数軸にし、反転させます。さらに xline()オプションで、選択されたλに基準線を引きます。

```
coefpath, for(assets ) tlevel(0) xunits(rlnlambda) xline(327)
```



罰則係数が小さくなるほど、モデルに入る係数が大きくなることがわかります。このパスプロットは候補となる共変量の数が比較的少ない時にグリッド探索時の変数ごとの挙動を確認するのに有効ですが、多くなると上記の様に個々の係数の変化がわかりにくくなるという問題があります。

参考文献

StataCorp. 2021. *Stata Treatment-effect Reference Manual: Potential outcomes/counterfactual outcomes release 17*. College Station, TX: Stata Press.

Chernozhukov, V., and C. B. Hansen. 2004. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *Review of Economics and Statistics* 86: 735-751.

Koch, B., D. M. Vock, and J. Wolfson. 2018. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* 74: 8-17.

Poterba, J. M., S. F. Venti, and D. A. Wise. 1994. 401(k) plans and tax-deffered saving. In *Studies in the Economics of Aging*, ed. D. A. Wise, 105-142. New York: National Bureau of Economic Research.

Poterba, J. M., S. F. Venti, and D. A. Wise. 1995. Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics* 58: 1-32.