第二回　スプライン補間でのノットの選択

スプライン補間を行う際，ノットの数や位置の選択は分析者の手に委ねられます．研究の目的から自ずとノットの数・位置が決まる場合もありますが，ここではそれ以外の，ノットの数・位置の決定に選択の余地が残される場合について，いくつかの指針を列挙したいと思います．

精度と簡易度のトレードオフ

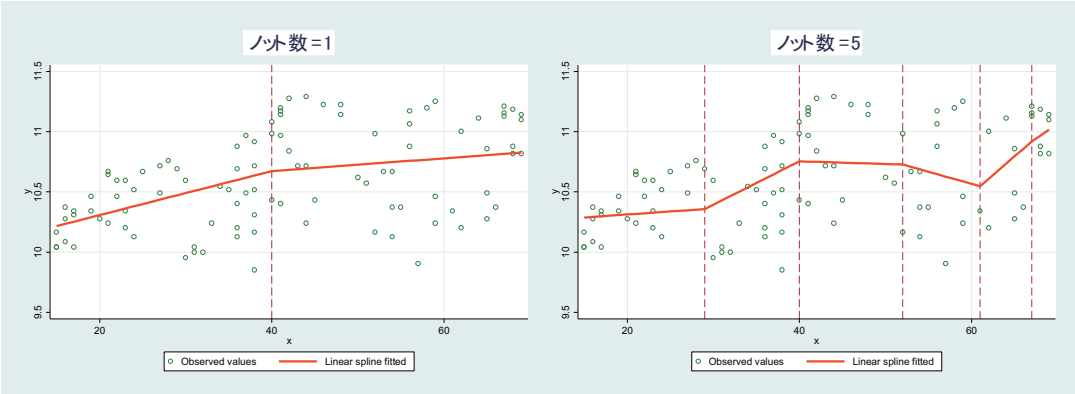


図 1: ノット数=1 とノット数=5 の 2 種類の線形スプラインフィット。ノット数=5 のノット位置は，ある程度の間隔を保ちつつ任意に選んだ 5 点。

図 1 は説明変数 x_i と従属変数 y_i の組 (x_i, y_i)　$(i = 1, 2, ...)$ 計 100 組からなる観測データに対して、ノット数=1 とノット数=5 の 2 種類の線形スプラインでフィットしたときの様子です．フィット値の当てはまりの良さを示す量の代表的なものとして，決定係数 R^2 があります．図 1 のそれぞれの場合の値は，ノット数 1 で $R^2 = 0.237$，ノット数 5 で $R^2 = 0.286$ であり，ノット数 5 の場合のほうが観測データにより近い，当てはまりの良い線を描けていると言えます．一方で，どちらが分布の特徴をより分かりやすく表しているかと問うと，ノット数 1 の場合のほうを選ぶ人が大半ではないかと考えられます．1 の場合では，x_i が 40 を越えると y_i の増加率が小さくなるという，粗くても単純明快な解釈を与えることができます．

モデルの評価

モデルの良さを示す量には，さまざまなものがあります．先述の決定係数 R^2 はモデルの説明変数を増やすほど良い値となり，説明変数のより多い複雑なモデルほど高く評価される傾向にあります．

そこで一般的には，変数の増加をペナルティとして加味した全体的なモデル評価量である自由度調整済み決定係数，AIC，BIC などが評価基準として用いられます．スプライン補間では，ノット数が一つ増えることで，見た目には説明変数が一つ増えます．スプライン補間を用いたモデルの良さを示す量としては，AIC，BIC などのほか，評価基準を自ら定義して用いることも可能と考えられます．なお，AIC に関しては，書籍 [2] にて，スプライン補間を用いたモデルの評価基準として挙げられています．

AIC は，モデルの尤度を L，パラメータ数（ここではノットで結ばれる各区間の傾きと定数項なので，ノット数+2）を k とすると，次式で計算されます．

$$AIC = -2\ln L + 2k$$

AIC が小さいほどモデルは高く評価されます．図 1 の二つの場合について計算すると，ノット数が 1 の場合 $AIC = 74.18$，5 の場合 $AIC = 75.53$ となり，この基準では前者のほうがわずかに良いモデルということになります．ノット数以外にもノットの位置に選択の余地があるので，この結果からノット数が 1 の場合のほうが 5 の場合よりも必ず AIC が良いと一概に言うことはできません．

Stata で AIC や BIC を確認するには，`regress` などの推定コマンドの実行後に、`estat ic` を実行します．たとえば図 1 のノット数 5 の場合の AIC を確認するコマンドの流れは，以下のようになります．

```
webuse mksp1, clear
mkspline age1 29 age2 40 age3 52 age4 61 age5 67 age6 = age
regress lninc age1-age6
estat ic
```

Akaike´s information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	100	-47.61224	-30.76644	7	75.53289	93.76908

Note: N=Obs used in calculating BIC; see [R] BIC note.

残差診断

スプライン補間による回帰分析は，重回帰分析の一種であり，回帰分析が仮定とする残差の分散均一性，残差の独立性が満たされるモデルが適切なモデルとなります．もしスプライン補間による回帰分析で，フィット値に不偏性を持たせたい場合，上記の仮定を満たすことが必要になります．

Stata で残差診断を行うには，通常の重回帰分析のように残差対フィット値プロット，正規 Q-Q プロットなどを描いて判断します．残差対フィット値プロットを描画するには，`regress` などの推定コマンドの実行後に `rvfplot` を実行します．先ほどの AIC の確認の例を用いることにすると，プロットの表示は，引き続いて以下を実行することで実現できます．

```
rvfplot
```

さらに正規 Q-Q プロットを描画するには，やはり推定コマンドの実行後に `predict r, residual` などで任意の名前の変数（ここでは `r`）に残差を格納したのち，`qnorm r` を実行します．

```
predict r, residual
qnorm r
```

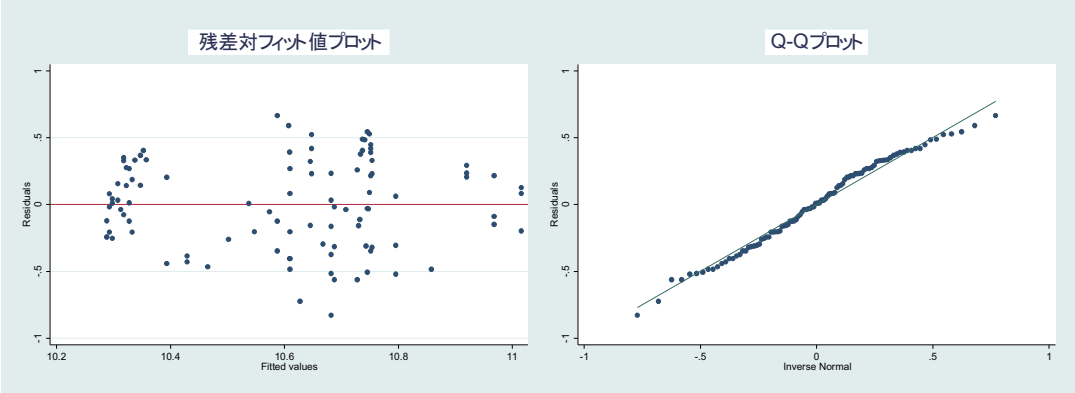


図 2: ノット数=5 の場合の残差分析プロット

パラメータの有意性

AIC，*BIC* の値は，パラメータの有意性に関わりなく決まります．このため，*AIC* の値が良い場合でも，パラメータの推定値すなわち図 1 の例であれば直線の傾きや切片が有意でないこともあります．

図 1 のフィットでのパラメータの推定値と有意性

	ノット数=1	ノット数=5
切片	9.9***	10***
区間 1 での傾き	0.018***	0.0049
区間 2 での傾き	0.0053	0.036*
区間 3 での傾き		−0.0022
区間 4 での傾き		−0.020
区間 5 での傾き		0.062
区間 6 での傾き		0.048
左端からノットを基準に区間 1，区間 2，...	$^{*}p < 0.05$;** $p < 0.01$;*** $p < 0.001$	

傾きの有意差検定を目的とするのであれば，対象となる傾きの推定値に一定の有意性が求められます．

Stata で，restricted cubic spline（制限 3 次スプライン）の場合のほか，linear spline（線形スプライン）の場合でも `mkspline` コマンドを `marginal` オプション付きで実行したケースでは，区間 i のパラメータ（線形スプラインの場合は傾き）が区間 $i-1$ のパラメータとの差分となります．この性質を利用すると，`regress` による推定で有意性が認められなかったパラメータがあれば，それらの中で p 値が最も高いパラメータの属する区間一つをモデルから除いて，再び推定を行うといった，いわばステップワイズ法のような方法で区間／ノットを減らし，すべてのパラメータに有意性の認められるモデルを探し当てることも実施できます．

信頼区間

スプライン補間による折れ線（または曲線）グラフには，さらに信頼係数 95% などの信頼区間を描画する場合もあります．信頼区間を描画すると，モデル推定で求まるフィット値の確からしさを可視化することができます．

数式などによる証明法は未確認ですが，信頼区間の形状は AIC や BIC とそれほど関連しそうになく，信頼区間の良し悪しというものがあるとすれば，それは AIC や BIC では評価できないものと考えられます．

信頼区間の特徴の一つに，信頼区間の幅があります．幅は独立変数の値によって変わりますが，平均的な幅を考えると，それはノット数が増えるほど広くなると推察されます．これはノット数が増えることでモデルで推定するパラメータが増え，未確定な値が増えることが主な理由です．

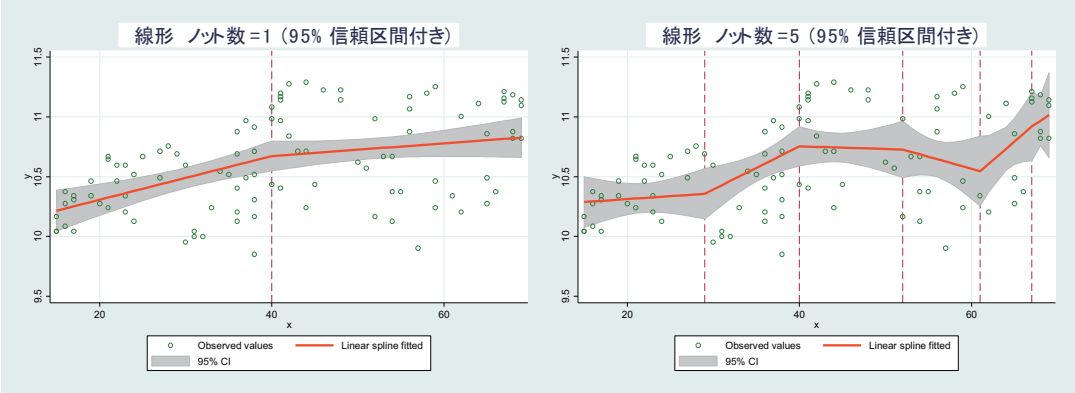


図 3: 95% 信頼区間付きプロット（線形スプライン）

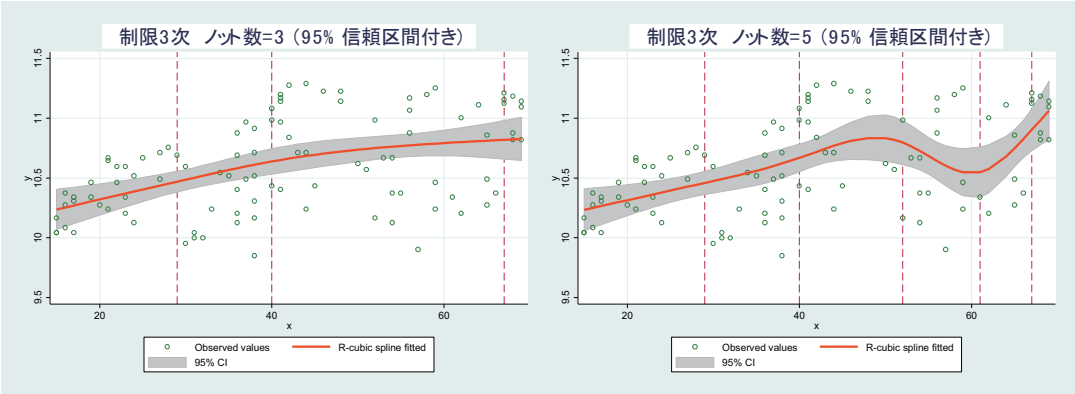


図 4: 95% 信頼区間付きプロット（制限 3 次スプライン）

信頼区間は従属変数 y_i の推定値 $\hat{y}_i$ が従う確率分布から求まります．線形回帰では，誤差 ε_i が平均 0，分散 σ^2 の正規分布に従うと仮定するため，分散の真の値 σ^2 が既知ならば，推定値 $\hat{y}_i$ は平均 $\hat{y}_i$，分散が独立変数 x_i の関数で表される式 $g(x_i)$ と σ^2 との積（$g(x_i)\sigma^2 = V(\hat{y}_i)$ と表すことにします）である正規分布に従います．しかし，実際の線形回帰では分散 σ^2 は未知なので，データから推定した値 $\hat{\sigma}^2$ を求めて使用し，結果として推定値 $\hat{y}_i$ は t 分布に従うことになります．95% 信頼区間は $\hat{y}_i \pm t_{0.025}^{(\mathrm{df})}\sqrt{\hat{V}(\hat{y}_i)}$ で求められ，符号が $+$ のときの値が上限，$-$ のときの値が下限となります．ここで，df は自由度で，(サンプルサイズ) $-$ (パラメータ数) です．

Stata で 95% 信頼区間を描画するには，`regress` などの推定コマンド実行後に，標準誤差 $\sqrt{\hat{V}(\hat{y}_i)}$，上限，下限を以下のように計算し，`rarea` を `twoway` コマンドのなかで用います．こちらも，先ほど実行したコマンド（`qnorm r`）に引き続いて実行するときの例です．

```
predict yhat, xb
predict se, stdp
generate ub = yhat + invttail(e(N)-7, 0.025)*se
generate lb = yhat - invttail(e(N)-7, 0.025)*se
sort age
twoway (rarea lb ub age) || (line yhat age) || (scatter lninc age)
```

上記で，`yhat` は従属変数 y_i の推定値 $\hat{y}_i$ を収める変数の名前として任意に選んだ変数名です．同様に，`se` は標準誤差 $\sqrt{\hat{V}(\hat{y}_i)}$ を，`ub` は上限を，`lb` は下限を，それぞれ収める変数の名前として任意に選んだ変数名です．`invttail(`df`,`p`)` は，自由度 df の t 分布において $t = \infty$ の側から累積したとき，確率 p をもたらす t を逆算する関数です．t 分布は自由度が十分大きければ正規分布と大差ないので，サンプルサイズが十分大きければ `invttail()` は 1.96 になります．

ちなみに，図 3 は上記コマンドにさらにオプションを指定して作成しています．具体的なコマンドは以下です．改行後も引き続き同じコマンドとなる///記号を使用しているため，do ファイルへコピー& ペーストしてご利用ください．

```
twoway ///
   ( rarea lb ub age, bcolor(gs12) lwidth(none) ) ///
 || ( scatter lninc age, msymbol(Oh) mcolor(green) ) ///
 || ( line yhat age, clcolor(orange_red) clwidth(thick) ), ///
 ytitle("ln(income)") ///
 legend(order(2 "Observed values" ///
              3 "Linear spline fitted" ///
              1 "95% CI")) ///
 xline(29 40 52 61 67, lpattern(dash)) ///
 xlabel(20 40 60, grid) ///
 title("Linear Spline with 95% CI, 5 knots ", ///
      size(huge) box bcolor(white) bmargin(small))
```

説明変数のパーセント点

ノットの数・位置を決める際の指標として，書籍 [2] では，説明変数のデータ数すなわちパーセント数を基準にする方法を示した文献を紹介しています．Stata でも制限 3 次スプライン用のオプションで利用できます．

書籍の記載では，以下の表で与えられるノットの数・位置の組み合わせにおいて，サンプルサイズ n が 30 未満の場合 $k = 3$，30 以上 100 未満の場合 $k = 4$，100 以上の場合 $k = 5$ の制限 3 次スプライン回帰が観測データの多くの分布パターンにおいてよい選択肢であったとしています．

このノット数とパーセント点の組み合わせを Stata で利用するには，以下のように `mkspline` コマンドの中で `nknot(ノット数)` というオプションを指定します．

```
mkspline a = age, cubic nknots(5)
```

k（ノット数）	パーセント点（ノット位置）
3	10, 50, 90
4	5, 35, 65, 95
5	5, 27.5, 50, 72.5, 95
6	5, 23, 41, 59, 77, 95
7	2.5, 18.33, 24.17, 50, 65.83, 81.67, 97.5

局所的な激しい変化

大抵の観測データは変化がなだらかで，ノット間の間隔が広いスプライン補間で特徴を十分に描写できると推察できますが，それに該当しないケースの一例として，局所的に激しい変化を伴う場合が考えられます．

制限 3 次スプライン補間ではノット（位置 x_j）を経るごとに，式の上では $\beta_j\,(x-x_j)^3$ 項の加算という変化が起きるのみであることから推察されるように，ノット間に存在するカーブはただ一つです．また式は高々3 次であるため，変化が急激であるほど，より多くのノットがないと相応のカーブが生まれません．複雑な形状が想定される箇所には，多くのノットを配置する必要が出てきます．

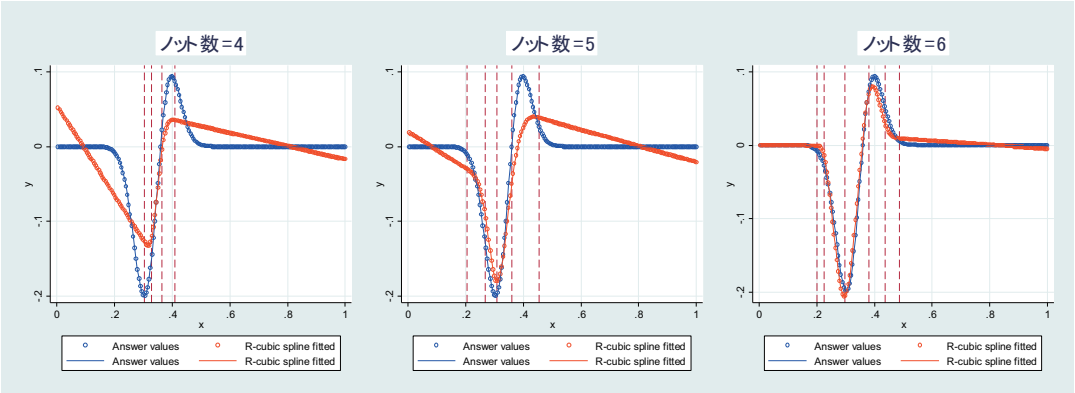


図 5: ノット数の増加に伴う局所激変事象のスプライン補間の改善

ノット数・位置の機械的選択

スプライン補間のノット数・位置の選択方法をアルゴリズムに組み込み，最適な値を自動で選択する試みも行われています．以下の表は，アルゴリズムの主要なステップを記載しています．

Step 1	ノットの数・位置の組み合わせ $(n, k_1, k_2, ..., k_n)$ を一つ選ぶ
Step 2	$(n, k_1, k_2, ..., k_n)$ でスプライン回帰を実行する
Step 3	評価基準とする量 γ を計算する
Step 4	評価基準とする量の中でこれまでの最高値 Γ と比較し，優れていれば最高値 Γ を更新する
Step 5	Step 1〜4 を繰り返す
Step 6	最終的に最高の評価基準量を与えたノットの数・位置の組み合わせを最適として採用する

ノットの数・位置の組み合わせを一つ選ぶ方法としては，総当たり的なものと，ランダムなものの大きく二種類が考えられます．評価基準とする量 γ には，AIC など場合に応じた量を使用します．

Stata で実施する一例として，図 5 のノット数=6 の場合について，ノットの位置をランダムに選ぶ方法をとるプログラムを記載します．

```
/*既知解としてのデータの作成*/
clear
set obs 200
gen x = _n / _N
gen env = normalden((x-0.34)/0.05)
gen y = env * sin(2*_pi*(x-0.36)/0.5)
*scatter y x

/*ノット数の設定*/
scalar knotnumber = 6

/*反復回数の設定*/
scalar iteration  = 2000

*set seed            1234
local  xnumber    = knotnumber - 1
scalar supaic     = 2 * knotnumber

/*最適なノット位置の選定*/
forvalues i = 1/`=iteration' {
  local knotlist ""
  forvalues j = 1/`=knotnumber' {
    local ustring=strofreal(runiform())
    local knotlist "`knotlist' `ustring'"
  }
  numlist "`knotlist'", sort
  local knotlist=r(numlist)
  *display "step `i': knots:`knotlist'"
  display "." _continue

  capture drop rx?
  capture drop yhat
  mkspline rx = x, cubic knots(`knotlist')
  quietly regress y rx1-rx`xnumber'
  quietly estat ic
  matrix A = r(S)
  scalar aic = A[1,5]
  if aic < supaic {
    scalar supaic = aic
    local supknotlist `knotlist'
  }
}

/*グラフ描画*/
capture drop rx?
capture drop yhat
mkspline rx = x, cubic knots(`supknotlist')
regress y rx1-rx`xnumber'
estat ic
predict yhat, xb

twoway ///
    ( scatter y x    , msymbol(Oh) mcolor(blue) ) ///
 || ( scatter yhat x , msymbol(Oh) mcolor(orange_red) ) ///
 || ( line y x       , clcolor(blue) clwidth(medium) ) ///
 || ( line yhat x    , clcolor(orange_red) clwidth(medium) ), ///
 ytitle("y") ///
 legend(order(1 "Answer values" ///
              2 "R-cubic spline fitted" ///
  3 "Answer values" ///
  4 "R-cubic spline fitted")) ///
 xline(`supknotlist', lpattern(dash)) ///
 xlabel(0 .2 .4 .6 .8 1, grid) ///
 title("ノット数=`=knotnumber'", size(huge) box bcolor(white) bmargin(medium))
```

参考文献

[1] Stata マニュアル『*[R] Stata Base Reference Manual Release 14*』, pp.1,523-1,529, Stata Press (http://www.stata.com/manuals14/rmkspline.pdf)

[2] Harrell, F. E., Jr. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. New York: Springer.

[3] Shuichi Kawano, Kei Hirose, Shohei Tateishi, and Sadanori Konishi 2010. *Recent Development in Regression Modeling and L_1 Type Regularization* (回帰モデリングと L_1 型正則化法の最近の展開). pp.211-242, 日本統計学会誌第 39 巻 第 2 号 (https://www.terrapub.co.jp/journals/jjssj/pdf/3902/39020211.pdf)

[4] William D. Dupont, W. Dale Plummer『*Using Stata 9 to Model Complex Nonlinear Relationships with Restricted Cubic Splines*』 (http://www.stata.com/meeting/4nasug/RCsplines.pdf)

[5] 基礎統計学 III 『自然科学の統計学』, 1992, 東京大学出版会

[6] C. J. Stone. Comment: Generalized additive models. 1986, *Statistical Sci*, 1:312-314

株式会社　ライトストーン
2016 年 8 月

研究者向けの統計解析ソフトウェア **Stata® 評価版のお申込み**

機能制限なし！30日間無料で使える評価版でまずはお試しください！

https://www.lightstone.co.jp/stata/evaluate.html

※学生の方への評価版の提供はございません。大学の講義で使用するなど単位取得に関わる形でのご利用となる場合は学生版のご購入をご検討ください。

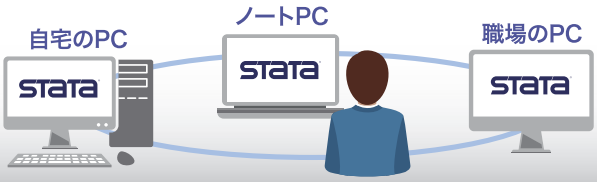
ライセンスはサブスクリプションがおすすめ！

サブスクリプションライセンスを選ぶメリット

- 常にStataの最新バージョンが利用できます
- 技術サポートの対象となります
- 毎年の経費としてStataを導入できます
- 利用者の増減に柔軟に対応可能
- 初期導入費を抑えたい場合にもおすすめ

シングルライセンス（サブスクリプション）

職場のPC・ノートPC・自宅のPC等、個人所有のPCにインストール可能です。
同時に使用できるのは1台のみとなります。



マルチユーザライセンス（サブスクリプション）　2ユーザ以上

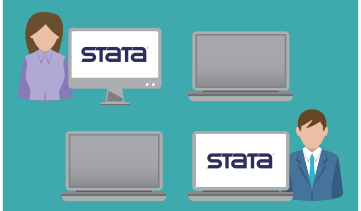
ボリュームライセンス

シングルライセンスを複数人でまとめてご購入いただく際のボリュームディスカウント商品です。

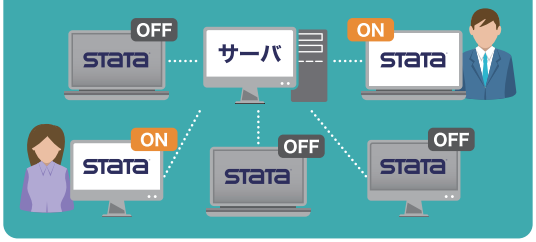
同時起動ライセンス

ご利用環境により下記3通りの運用方法からご選択いただけます。

PC毎にスタンドアロンインストールをして運用
例:2ライセンスの場合



クライアント＆サーバ環境によるネットワーク運用
例:2ライセンスの場合



1台の仮想マシンにユーザがアクセスする運用
例:2ライセンスの場合

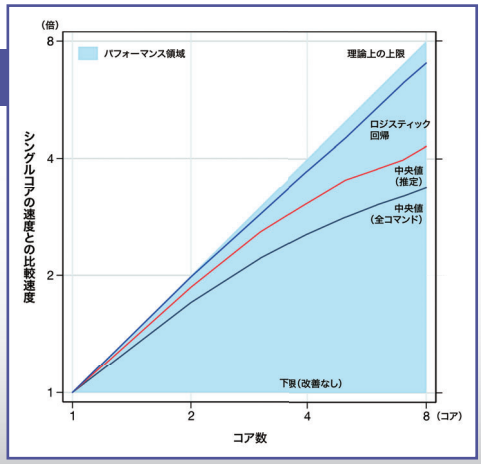
＼データ処理時間の削減ならMP／

Stata／MPの演算能力

Stata MPはPCの持つマルチコアの特性を活かして、処理を分散・並列する機能を備えます。およそ85%以上のコマンドで処理速度が向上し、コア数に応じた計算時間の短縮が期待できます。コマンドごとの処理速度の向上の度合いについては以下のページや資料をご覧ください。

コア数	全てのコマンド	推定コマンド	ロジスティック回帰
2	1.7倍	1.8倍	1.9倍
4	2.6倍	3.1倍	3.8倍
8	3.3倍	4.2倍	6.8倍

弊社Webページ内の『Stata／MP』
https://www.lightstone.co.jp/stata/statamp.html



株式会社 ライトストーン
〒101−0031
東京都千代田区東神田 2−5−12　龍角散ビル 7F
TEL：03−3864−5211　e-mail：sales@lightstone.co.jp
FAX：03−3865−0050　WEB：https://www.lightstone.co.jp/