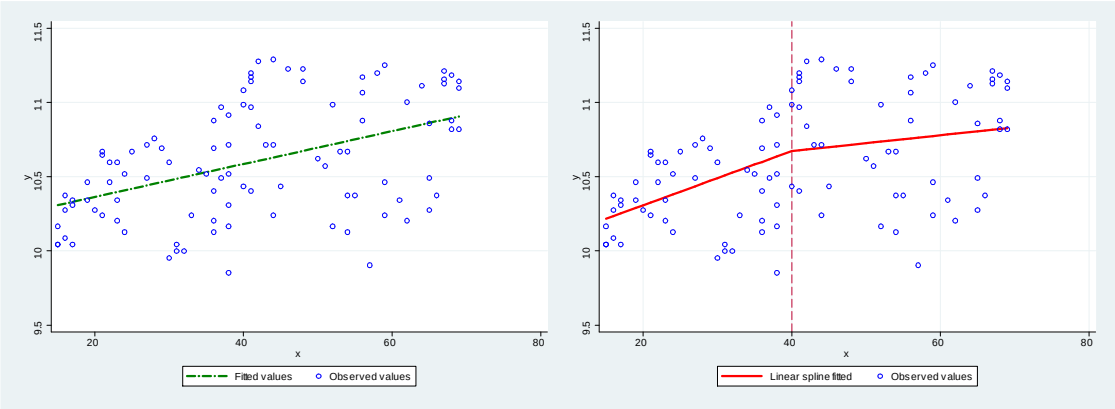
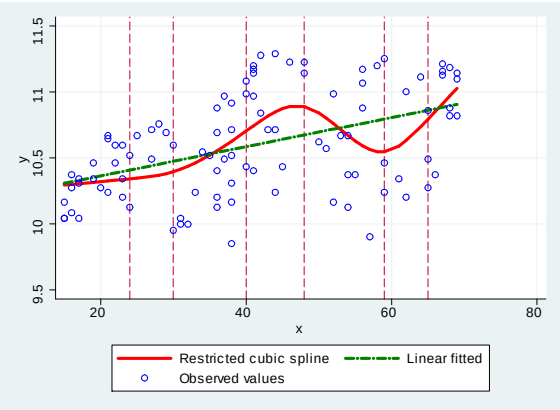
第一回　mksplineによるスプライン補間

回帰分析ではいつも単純な直線で表される線形関係を仮定したモデルで分析するとは限りません。たとえば以下の図で右にあるような，ある位置を境に傾きが変化することを仮定して、傾きの変化が有意かどうか検定したい場合があります。



あるいは，滑らかな曲線を用いたほうが，より真の分布に近い描写ができる場合もあります。



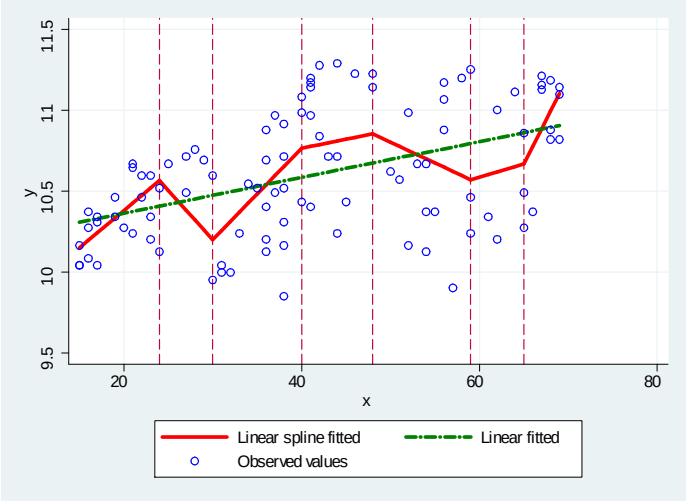
グラフが直線になる線形関数ではなく，折れ線や滑らかな曲線になるような関数でフィットを行う方法の一つに，スプライン補間があります．Stata では `mkspline` コマンドを利用して，linear spline（線形スプライン），または restricted cubic spline（制限 3 次スプライン）の 2 種類のスプラインのモデルでのフィットが行えます。各場合でどのような処理が行われるのか，具体的に見ていきます。

Linear spline

線形スプラインは，あらかじめ定めたノット（knot，結び目）という説明変数上の節点位置で折れ曲がることができる折れ線グラフで表されるモデルです。1 次スプライン，区分線形補間とも呼ばれます。

Stata で行うには，まず `mkspline` コマンドで，いわば説明変数をノットで分解する作業を行い，分解でできる新たな成分変数を作成します（コマンドの構文については後述します）。その後，作成された成分変数を説明変数として回帰推定を行い，フィット曲線を描きます。以下は，マニュアル[1]に記載のある `mkspline` の Examples のデータセット[2]について，変数 `age` を `x` に，`lninc` を `y` に変数名変更したもので行った例です。

```
mkspline x1 24 x2 30 x3 40 x4 48 x5 59 x6 65 x7 = x
regress y x1-x7
predict yhat, xb
twoway line yhat x, clcolor(red) clwidth(thick) ///
 || lfit y x, clcolor(green) clwidth(thick) ///
 || scatter y x, msymbol(Oh) mcolor(blue) ///
 ytitle("y") ///
 legend(label(1 "Linear spline fitted") ///
 label(2 "Linear fitted") ///
 label(3 "Observed values")) ///
 xline(24 30 40 48 59 65) xlabel(, grid)
```



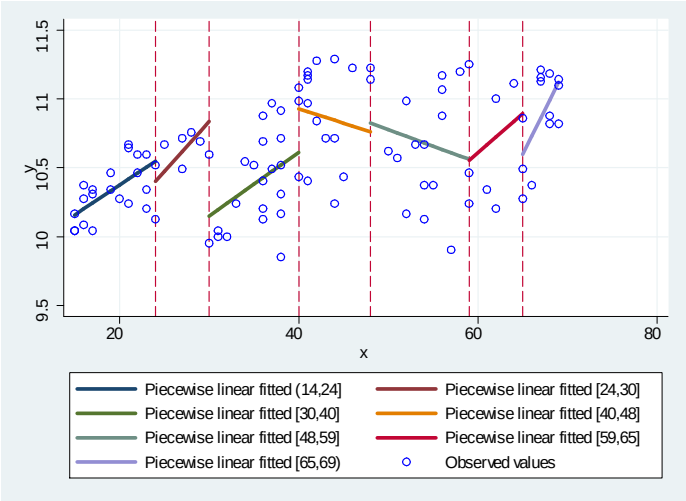
グラフで，縦の破線がノットの位置です。`x=24` にはじまる 6 つのノットを設定しました。これにより，7 つの区間が `x` 上にできます。折れ線グラフは，ノットで傾きが変化する，全体としては連続な線です。グラフがノットで連続とは，ノット上で線が接続しているということです。この連続性を必要条件とするため，線形スプラインによる推定は，単純に 7 つの区間ごとの別々の推定とは内容を異にします。

[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]`use http://www.stata-press.com/data/r14/mksp1` で読み込めます。

LightStone® 株式会社ライトストーン

```
regress y x   if x < 24
predict yhat1 if x <= 24, xb
regress y x   if 24 <= x & x < 30
predict yhat2 if 24 <= x & x <= 30, xb
( 以下、略)
```



`mkspline` で作成される成分変数を詳しく見てみます。`mkspline` で作成されるのは，(ノット数)+1 個の新たな変数です。ここでは，`x1` から `x7` までの 7 つが作成されます。

```
mkspline x1 24 x2 30 x3 40 x4 48 x5 59 x6 65 x7 = x
```

このコマンドは，以下のような構文をとり，新変数名とノットの値は所定の位置に直接入力します。

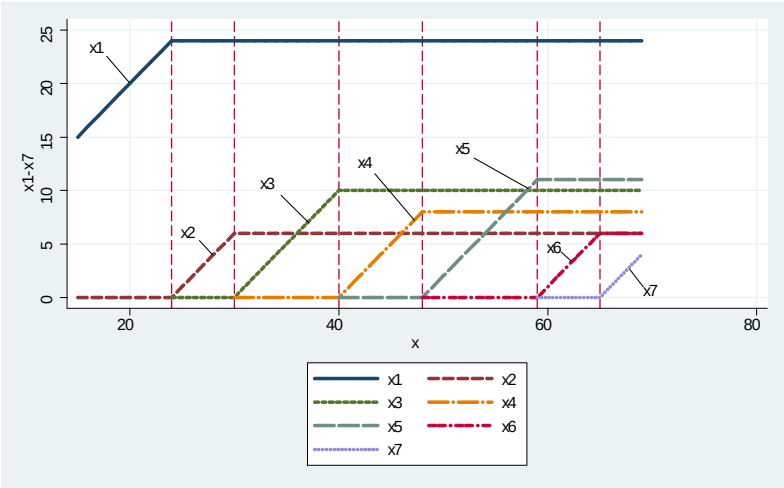
mkspline　新変数 1　ノット 1　新変数 2　ノット 2　...　新変数 7　=　既存の変数

新変数のとる値は，Stata マニュアルにある `mkspline` の記載のように，以下の式に従います。

$$
\begin{aligned}
\texttt{x1} &= \min\left(\texttt{x}, k_1\right) \\
\texttt{x2} &= \max\left(\min\left(\texttt{x}, k_2\right), k_1\right) - k_1 \\
\texttt{x3} &= \max\left(\min\left(\texttt{x}, k_3\right), k_2\right) - k_2 \\
\texttt{x4} &= \max\left(\min\left(\texttt{x}, k_4\right), k_3\right) - k_3 \\
\texttt{x5} &= \max\left(\min\left(\texttt{x}, k_5\right), k_4\right) - k_4 \\
\texttt{x6} &= \max\left(\min\left(\texttt{x}, k_6\right), k_5\right) - k_5 \\
\texttt{x7} &= \max\left(\texttt{x}, k_6\right) - k_6
\end{aligned}
$$

上式は，マニュアルの Methods and formulas 節にある式の変数を，適宜ここでの場合に置き換えたものです。k_1，k_2，... はコマンドで指定したノット 1，ノット 2，... の値です。グラフに表すと，`x` と `x1`，`x2`，... それぞれの関係は次の図のようになります。

LightStone® 株式会社ライトストーン



グラフの左端で，x1 は x と同じ値，それ以外はゼロとなります。x1 から x7 はすべて，順番に各区間で傾き 1 で上昇し，上昇後はそのままの値を保ちます。また，各区間で値を変える変数は 1 つだけです。

次に，x1 から x7 を説明変数とした回帰について見てみます。回帰モデルは，式（1）のように表せます。

$$\mathtt{y} = \alpha + \beta_1\mathtt{x1} + \beta_2\mathtt{x2} + \beta_3\mathtt{x3} + \beta_4\mathtt{x4} + \beta_5\mathtt{x5} + \beta_6\mathtt{x6} + \beta_7\mathtt{x7} + u \tag{1}$$

つまり，先ほどの x1 から x7 の線形結合であるので，先ほどの x1 から x7 までのグラフを線形合成したグラフが y の予測値のグラフとなります。区間ごとに，ゼロとなる項を省略し，整頓して記述すると，上式は次のように変形できます。

$$\mathtt{y} = \left\{ \begin{array}{ll} \{\alpha\} + \beta_1\mathtt{x1} + u & (\mathtt{x} < k_1) \\ \{\alpha + \beta_1 k_1\} + \beta_2\mathtt{x2} + u & (k_1 \leq \mathtt{x} < k_2) \\ \{\alpha + \beta_1 k_1 + \beta_2 (k_2 - k_1)\} + \beta_3\mathtt{x3} + u & (k_2 \leq \mathtt{x} < k_3) \\ \{\alpha + \beta_1 k_1 + \beta_2 (k_2 - k_1) + \beta_3 (k_3 - k_2)\} + \beta_4\mathtt{x4} + u & (k_3 \leq \mathtt{x} < k_4) \\ \vdots & \end{array} \right\}$$

上記は紙面の都合上，k_4 以上の区間については省略しました。フィットにより α，β_1，β_2，... の点推定値が確定してしまえば，各区間の式で最初にある中カッコで囲った部分は定数になります。続く x1，x2，... などを含む項が 1 次項，最後の u が誤差項です。1 次項に着目すれば，β_i は x の左から i 番目の区間の傾きになることが理解できます。また，ノットにおける連続性について，代表してノット 2（k_2）近傍だけを吟味すると，まず x を負の側から k_2 に，すなわち 2 番目の区間において x2$-(k_2 - k_1)$ を負の側からゼロに近づけていくと，

LightStone® 株式会社ライトストーン

$$\lim_{\mathtt{x2}-(k_2-k_1)\ \to\ -0} = \{\alpha + \beta_1 k_1\} + \beta_2 \mathtt{x2}$$
$$= \alpha + \beta_1 k_1 + \beta_2 (k_2 - k_1)$$

また，x を正の側から k_2 に，すなわち 3 番目の区間において x3 を正の側からゼロに近づけていくと，

$$\lim_{\mathtt{x3}\to\ +0} = \{\alpha + \beta_1 k_1 + \beta_2 (k_2 - k_1)\} + \beta_3 \mathtt{x3}$$
$$= \alpha + \beta_1 k_1 + \beta_2 (k_2 - k_1)$$

と，同じ式となり，連続性が担保されていることが分かります（式で，u は 0 としました）。regress コマンドを実行して最小二乗法（OLS）によるフィットを行うと，以下の結果になります。

```
. regress y x1-x7

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(7, 92)        =      7.01
       Model |  5.28030734         7  .754329621   Prob > F        =    0.0000
    Residual |  9.89300058        92  .107532615   R-squared       =    0.3480
-------------+----------------------------------   Adj R-squared   =    0.2984
       Total |  15.1733079        99  .153265737   Root MSE        =    .32792

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0460597   .0212775     2.16   0.033     .0038008    .0883186
          x2 |  -.0607856   .0297025    -2.05   0.044    -.1197773    -.001794
          x3 |   .0565476    .017106     3.31   0.001     .0225736    .0905216
          x4 |   .0111599   .0219786     0.51   0.613    -.0324916    .0548113
          x5 |  -.0259026   .0189202    -1.37   0.174    -.0634798    .0116746
          x6 |    .016147   .0328659     0.49   0.624    -.0491276    .0814216
          x7 |   .1087038   .0572334     1.90   0.061    -.0049667    .2223742
       _cons |   9.459758   .4226253    22.38   0.000     8.620388    10.29913
------------------------------------------------------------------------------
```

この結果で求められた α，β_1，β_2，... の推定値（Coef. 列の値）を用いて，y の推定値を計算するには，predict yhat，xb（yhat の位置には任意の新変数名を入力）を実行します。実行コマンド，およびその後の線形スプラインのグラフの描画コマンド，結果のグラフは本節の冒頭に掲載の通りです。

ノット 1 で傾きが有意に変化したかを検定するには，lincom x2 - x1 を実行して，係数の減算結果がゼロと有意に異なるかを検定します。また，mkspline コマンドを marginal オプション付きで実行することにより，regress による x2 の推定係数が，ノット 1 での傾きの変化分となるので、regress の結果だけで変化の有意性が確認できます。

Restricted cubic spline

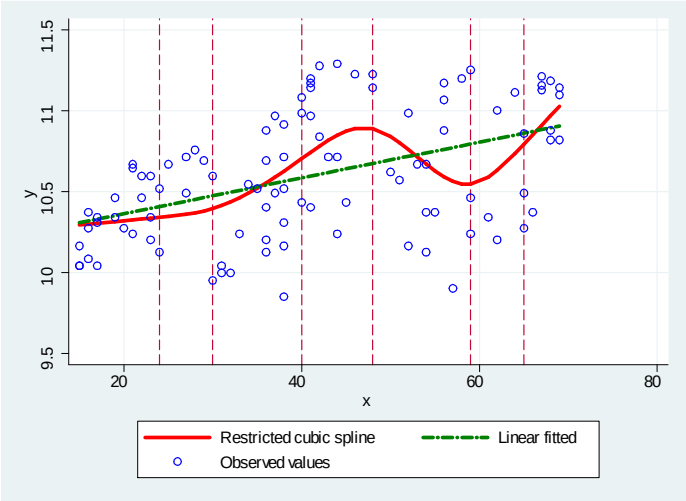
制限 3 次スプラインは，スプライン補間として代表的な 3 次のスプラインの一つです。3 次スプラインとは，1 次スプラインと同様にあらかじめノットを定め，ノットで区切られる区間ごとにひとつ

LightStone® 株式会社ライトストーン

ずつ 3 次多項式を求めて結んだ曲線です。各 3 次多項式はノットにおいて連続，さらに 1 次導関数，2 次導関数も連続となることを必要条件とします。restrict（制限）3 次スプラインとは，3 次スプラインのうち両端の区間で関数が 1 次多項式という制限を課し，これにより両端の区間での値を安定させたものです。

Stata では，やはり線形スプラインのときと同様，ノットとする x の値を指定し，成分変数を作成をして，作成された成分変数を説明変数として推定を行い，フィット曲線を描きます。以下の例では，前節の線形スプラインと同じデータセットを用いています。

```
mkspline x1 24 x2 30 x3 40 x4 48 x5 59 x6 65 x7 = x
regress y x1-x7
predict yhat, xb
twoway line yhat x, clcolor(red) clwidth(thick) ///
 || lfit y x, clcolor(green) clwidth(thick) ///
 || scatter y x, msymbol(Oh) mcolor(blue) ///
 ytitle("y") ///
 legend(label(1 "Linear spline fitted") ///
 label(2 "Linear fitted") ///
 label(3 "Observed values")) ///
 xline(24 30 40 48 59 65) xlabel(, grid)
```



mkspline で作成される成分変数を詳しく見てみます。mkspline で作成されるのは，(ノット数) − 1 個の新たな変数です。ここでは，rx1 から rx5 までの 5 つが作成されます。

```
mkspline sx = x, cubic knots(24 30 40 48 59 65)
```

このコマンドは，以下のような構文をとり，新変数名の stem（共通の文字列）とノットの値は，所定の位置に直接入力します。

mkspline　新変数の stem　=　既存の変数, cubic knots(ノット 1　ノット 2　...)

新変数のとる値は，Stata マニュアルにある mkspline の記載のように，以下の式に従います。

LightStone® 株式会社ライトストーン

$$
\begin{aligned}
\texttt{rx1} &= \texttt{x} \qquad (2)\\
\texttt{rx2} &= \frac{(\texttt{x}-k_1)_+^3 - (k_6-k_5)^{-1}\left\{(\texttt{x}-k_5)_+^3(k_6-k_1) - (\texttt{x}-k_6)_+^3(k_5-k_1)\right\}}{(k_6-k_1)^2}\\
\texttt{rx3} &= \frac{(\texttt{x}-k_2)_+^3 - (k_6-k_5)^{-1}\left\{(\texttt{x}-k_5)_+^3(k_6-k_2) - (\texttt{x}-k_6)_+^3(k_5-k_2)\right\}}{(k_6-k_1)^2}\\
\texttt{rx4} &= \frac{(\texttt{x}-k_3)_+^3 - (k_6-k_5)^{-1}\left\{(\texttt{x}-k_5)_+^3(k_6-k_3) - (\texttt{x}-k_6)_+^3(k_5-k_3)\right\}}{(k_6-k_1)^2}\\
\texttt{rx5} &= \frac{(\texttt{x}-k_4)_+^3 - (k_6-k_5)^{-1}\left\{(\texttt{x}-k_5)_+^3(k_6-k_4) - (\texttt{x}-k_6)_+^3(k_5-k_4)\right\}}{(k_6-k_1)^2}
\end{aligned}
$$

ここで $(u)_+$ のようにカッコに + 記号が添えられた記号は切断冪（べき）級数とも呼ばれ，次式で定義されます。

$$
(u)_+ = \left\{ \begin{array}{ll} u & (u \geq 0) \\ 0 & (u < 0) \end{array} \right\}
$$

上式は，マニュアルの Methods and formulas 節にある式の変数を，適宜ここでの場合に置き換えています。k_1，k_2，... はコマンドで指定したノット 1，ノット 2，... の値です。ここで，sx2 から sp5 までの式には $(u)_+$ の形をした切断冪級数が 3 つずつ見られますが，k_1 を越えない範囲ではどの式においてもいずれの切断冪級数もゼロであり，x が増加するに従って 1 つ目の冪級数だけが順次ゼロではなくなってきます。また，後ろの 2 つについては，k_5 を越えない範囲ではいずれもゼロのままです。したがって k_5 を越えない範囲では，どの式も $\frac{(\texttt{x}-k_i)^3}{(k_6-k_1)^2}$ となり比較的単純な 3 次多項式で表されます[3]。さらに，k_6 以降の最後の区間ではいずれの式も 1 次多項式になります。rx1 から rx5 をグラフに表すと，以下のようになります。

rx1 は x と全く同じです。それ以外は x が小さいうちはゼロですが，増加してノットを越えるに従って，次々と 3 次関数としてグラフが立ち上がります。前述のように，x が k_5 までは 1 つ目の $(u)_+$ の項で表される，係数に変化のない 3 次関数です。k_5 から k_6 までは式は複雑ですが依然として 3 次関数で，k_6 以降は傾き一定の 1 次関数です。

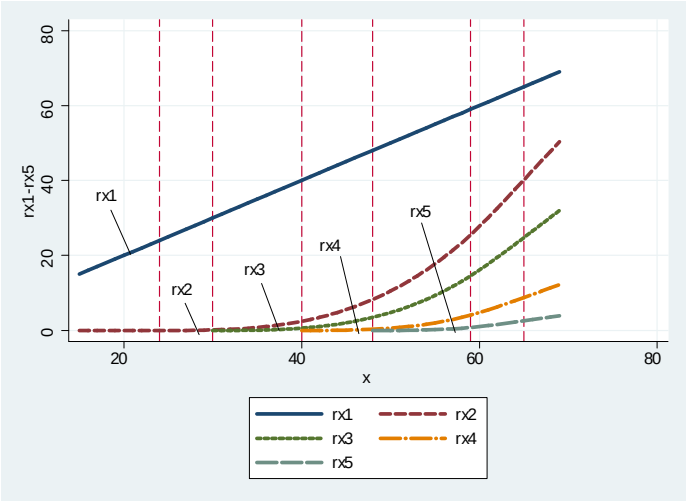
次に，rx1 から rx5 を説明変数とした回帰について見てみます。回帰モデルは次のようになります。

$$
\texttt{y} = \alpha + \beta_1\texttt{rx1} + \beta_2\texttt{rx2} + \beta_3\texttt{rx3} + \beta_4\texttt{rx4} + \beta_5\texttt{rx5} + u \qquad (3)
$$

つまり，先ほどの rx1 から rx5 の線形結合であるので，先ほどの rx1 から rx5 までのグラフを線形合成したグラフが y の予測値のグラフとなります。上式は一見すると 1 次多項式のように見えま

[3]**制限のない**3 次スプラインでは，ノットを横切るごとに回帰式に $A_i\,(x-k_i)_+^3$（A_i は任意の係数）が加算され，成分変数は (ノット数) + 1 となります。しかし，**制限**3 次スプライン推定では、両端の区間で回帰式が 1 次関数という制約が課されることから，左端の区間で 2 次項と 3 次項の係数がゼロとなり，また最後から 1 番目と 2 番目の係数も，他の係数が決まれば一意に決定されます。この点が反映され，mkspline で作成される成分変数は rx1 のみが 1 次式となり，$(x-k_5)_+^3$ と $(x-k_6)_+^3$ が rx2 から rx5 へ合算され，結果として rx6 と rx7 の作成は省かれています。

LightStone® 株式会社ライトストーン



すが，式（2）を代入すると明らかなように，実質的には x の 3 次多項式です。ノットを横切るごとに定数項，および 1 次項，2 次項，3 次項の係数の値が変わります。変化分は k_1，k_2，... の選び方や β_1，β_2，... の点推定値で決まりますが，具体的に式に表すと少し複雑です。

次に，連続性について，引き続き式（3）に（2）を代入した形を考えます。前述のように k_1，k_2，... はあらかじめ決まっており，β_1，β_2，... の値もひとたび点推定値が決まれば定数となることから，これらの係数部分を B，C などの大文字 1 字で置き換えて，さらに各区間でゼロになる項の記述をしないことにすると，以下のような式が書けます。

$$\mathrm{y} = \left\{ \begin{array}{ll} \alpha + B\mathrm{x} + u & (\mathrm{x} < k_1) \\ \alpha + B\mathrm{x} + C(\mathrm{x} - k_1)^3 + u & (k_1 \leq \mathrm{x} < k_2) \\ \alpha + B\mathrm{x} + C(\mathrm{x} - k_1)^3 + D(\mathrm{x} - k_2)^3 + u & (k_2 \leq \mathrm{x} < k_3) \\ \vdots & \end{array} \right\} \tag{4}$$

紙面の都合上，k_3 以上の区間については記述を省略しました。上式から，x が増加して各ノットを経るにつれ，$C(\mathrm{x} - k_1)^3$，$D(\mathrm{x} - k_2)^3$，... の項が順次加算されるのみであることが分かります。これらの項は，x が k_1，k_2，... というノットを経たばかりのうちは極めてゼロに近いので，感覚的に連続性が保たれていることが理解できます。また，式（4）の 1 次導関数，2 次導関数についても，各ノットを経ると加算される項は，$C(\mathrm{x} - k_1)^3$，$D(\mathrm{x} - k_2)^3$，... の 1 次導関数 $3C(\mathrm{x} - k_1)^2$，$3D(\mathrm{x} - k_2)^2$，...，そして 2 次導関数 $6C(\mathrm{x} - k_1)$，$6D(\mathrm{x} - k_2)$，... であるので，いずれもノット近傍でゼロであり，それぞれの連続性が担保されていると考えてもよさそうなことが分かります。また，両端の区間の式は，計算すると一次多項式の形で表せます。

最後に `regress` コマンドで最小二乗法（OLS）によるフィットを行うと，以下の結果になります。

```
. regress y rx1-rx5

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(5, 94)        =      8.33
       Model |  4.65699513         5  .931399026   Prob > F        =    0.0000
    Residual |  10.5163128        94  .111875668   R-squared       =    0.3069
```


LightStone® 株式会社ライトストーン

				Adj R-squared	= 0.2701
Total	15.1733079	99	.153265737	Root MSE	= .33448

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rx1	.0052178	.0123723	0.42	0.674	-.0193477	.0297834
rx2	.1788471	.238566	0.75	0.455	-.2948313	.6525256
rx3	-.2650082	.554882	-0.48	0.634	-1.366739	.8367231
rx4	-.4225963	.7852023	-0.54	0.592	-1.981634	1.136441
rx5	1.299019	.8247793	1.57	0.119	-.3385991	2.936638
_cons	10.21692	.2776262	36.80	0.000	9.66569	10.76816

この結果で求められた α，β_1，β_2，... の推定値がどこにどのように効いてくるのかについては，先ほどの式での検討からも推察されるように，解釈が難しいとされています。y の推定値を計算するには，`predict ryhat, xb`（ryhat の位置には任意の新変数名を入力）を実行します。実行コマンド，およびその後の線形スプラインのグラフの描画コマンド，結果のグラフは本節の冒頭に掲載の通りです。

制限 3 次スプラインの適合度（goodness of fit）は，AIC などで評価されます。

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. 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）

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

株式会社　ライトストーン
2016 年 7 月

LightStone® 株式会社ライトストーン

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。

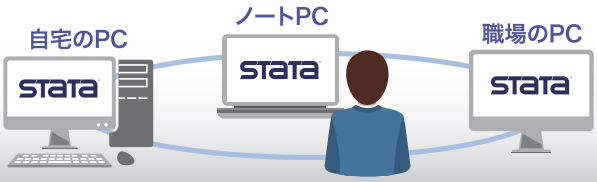
ライセンスはサブスクリプションがおすすめ！

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

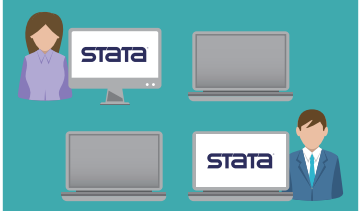
ボリュームライセンス

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。

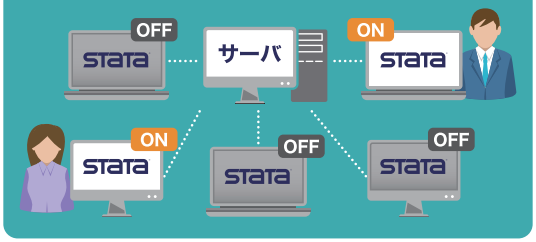
同時起動ライセンス

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

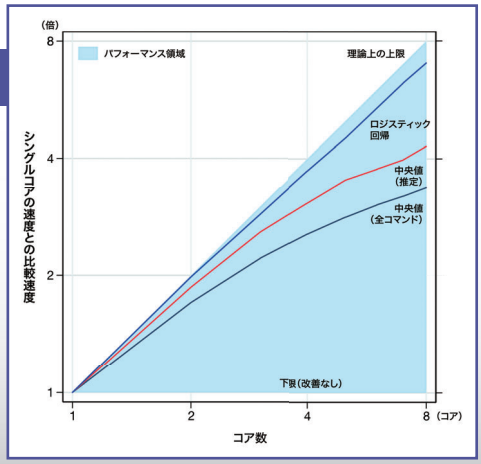
＼データ処理時間の削減ならMP／

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/