SAS/STATの最新分析機能

THE POWER TO KNOW®

SAS Institute Japan (株)
技術本部
泉水 (いずみ) 克之

本日は、

- SAS Forum **ユーザー会 学術総会** 2006 **にご参加いただき、誠にありがとうございます。**

本日のお話

- SAS9.1.3とSAS9.2の「はざ間」なので、あまり目新しいネタがありません。

でも、

- その間に、いくつかのプロシジャが提供されたので、そちらについて少しご紹介します。

具体的には、STATの

- 線形回帰モデルにおける新しい変数選択法を提供する

GLMSELECTプロシジャ

それと、

- 分位点回帰に対応した

QUANTREGプロシジャ

この2つの概略を

- ご紹介します。

はじめに、お願いです

- これからご紹介する2つの追加機能は、SAS 9.1.3 では評価版の扱いとなっています。正規版と同等にお考えにならないようお願いします。

お願いの続き

- 次期リリースSAS 9.2 では、正規版のプロシジャとなる予定ですが、その際に構文の変更などが行なわれる可能性があります。

まず1つ目のお題

- 線形回帰モデルにおける変数選択

回帰モデルでの変数選択って

- 今でも研究されており、方法は星の数ほど提案されているようですが、、、

SASでは

- 線形回帰モデルに対応した

REGプロシジャ

で、いくつかの方法が用意されています。

たとえば、総当り的な方法

- 調整寄与率
- MallowsのCp
- (やろうと思えば) AIC
- その他

たとえば、逐次選択

- 変数増加法
- 変数減少法
- 変数増減（ステップワイズ）法
- いずれもF統計量に基づく

でも、このご時世、

- ちょっと古めかしい感じもしたりしなかったり。
- 20年も前から同じ機能は存在していましたし。

総当り法だと、

- 説明変数の数が増えると、処理に時間がかかるし、、、

逐次選択については、

- そもそも、F値に基づいて逐次選択するのってどうなの？
- と、指摘している人達もいます。

そこで、GLMSELECT

- 米国SAS Institute Inc.では、新たにGLMSELECTプロシジャを開発しています。
- SAS9.1.3において評価版として公開しています。

GLMSELECTは、

- 今のところ、32bit Windows環境でのみ利用可能なモジュールが提供されています。

GLMSELECTを使うには

- 以下のURLあたりから、EXEファイルをダウンロードしてインストールしてください。

http://support.sas.com/rnd/app/da/glmselect.html

それよりも、

- 日本のGoogleかYahoo!でGLMSELECTで検索を。
- SAS Japanの紹介サイトが一番上にヒットするはずです（多分）。

ところで、

GLMSELECTできることは

- **名前の通り、**

PROC GLM + Selection

のイメージです。

GLMSELECTできることは、

以降で概略を順にご紹介します。

GLMSELECTの主な機能

1. カテゴリ変数を含む回帰で変数を”逐次選択”
2. 逐次選択の方法をより細かく設定可能
3. LAR(2002/2004) and Lasso(1996)
4. 入力データを学習・検証・テストデータに分割し、モデルを検証・評価
5. その他いろいろ

GLMSELECTの主な機能(1)

1. カテゴリ変数を含む回帰で変数を”逐次選択”

カテゴリ変数を含む回帰で”逐次選択”

- カテゴリ変数を説明変数として含む線形回帰をSASで実現しようとすると、

これまでは

1. PROC GLM
2. **ダミー変数を作成して** PROC REG
3. SAS Enterprise Miner

1. PROC GLM

```
PROC GLM DATA=Test;
  CLASS Group ;
  MODEL  Y= Group X1 X2 /SOLUTION;
  OUTPUT OUT=Out  P=Predict_y;
RUN;
```

←カテゴリ変数を指定

→でも、変数選択の機能がない

2.ダミー変数を作成して PROC REG

```
/* ダミー変数を作ってから */

PROC REG DATA=Test2;

MODEL  Y= {Group_A Group_B} X1 X2
                    /SELECTION=STEPWISE;

RUN;
```

→しかし、ダミー変数を意図どおりに作るって手間がかかったりする

そこで、PROC GLMSELECTの登場

```
PROC GLMSELECT DATA=Test;
  CLASS Group ;
  MODEL  Y= Group X1 X2
              /SELECTION=STEPWISE;
RUN;
```

でも、色々と注意が必要です

- **総当り的な方法はありません。**
- **REGのデフォルトの動きとは異なります。**

ということで、次にいきましょう。

GLMSELECTの主な機能(2)

2.逐次選択の方法を、より細かく設定可能

逐次選択の方法がより細かく設定可能

- PROC REGでは、F統計量（正確にはそのp値）に基づいて変数の選択（取り込みと削除）を行なっています。

逐次選択の方法がより細かく設定可能 (続き)

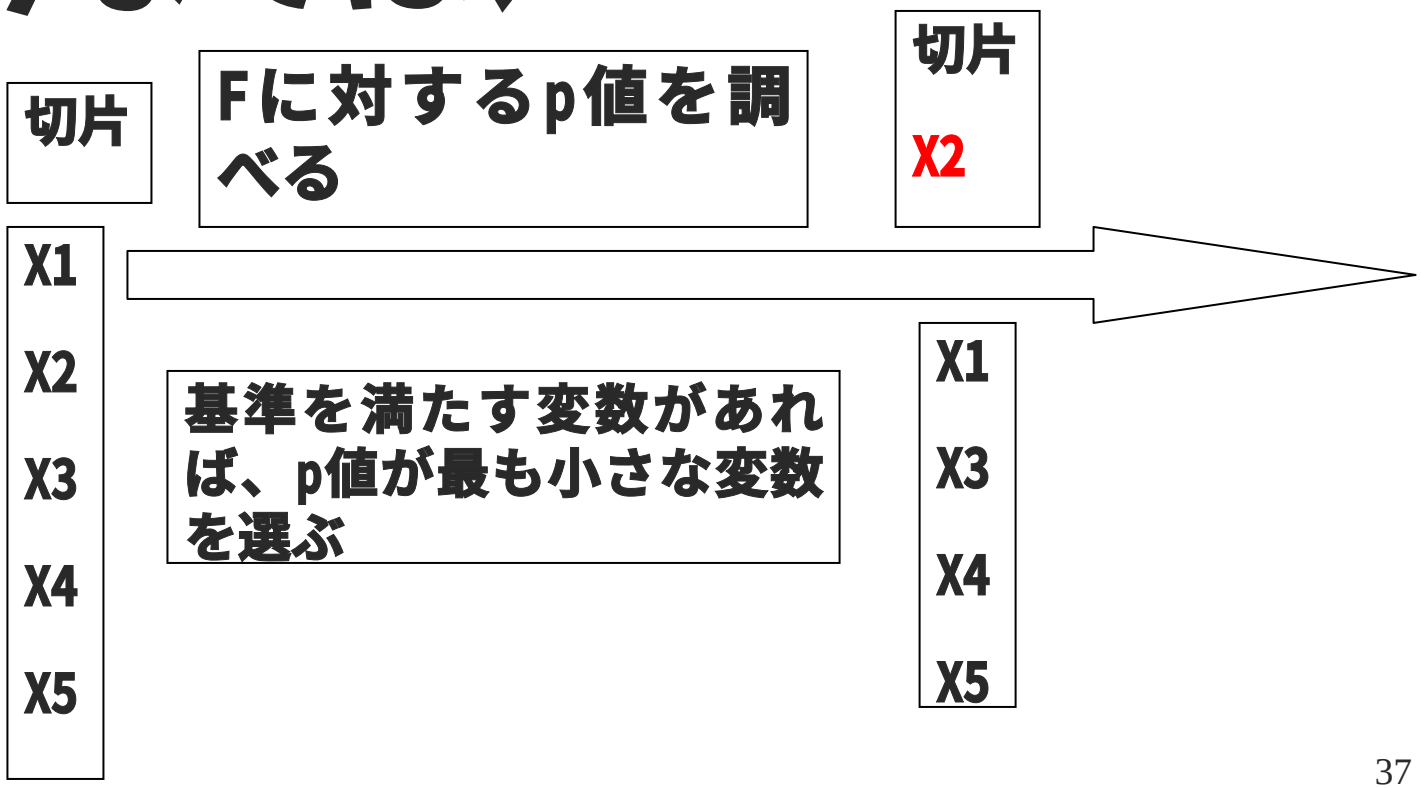
- PROC GLMSELECTでは、SBC、AIC、AICCなどの増減や、バリデーションデータの使用、クロスバリデーションに基づく逐次変数選択法などが用意されています。

どのようなことができるかというと、、、

- **たとえば、**Forward selection **(変数増加法)を例として**

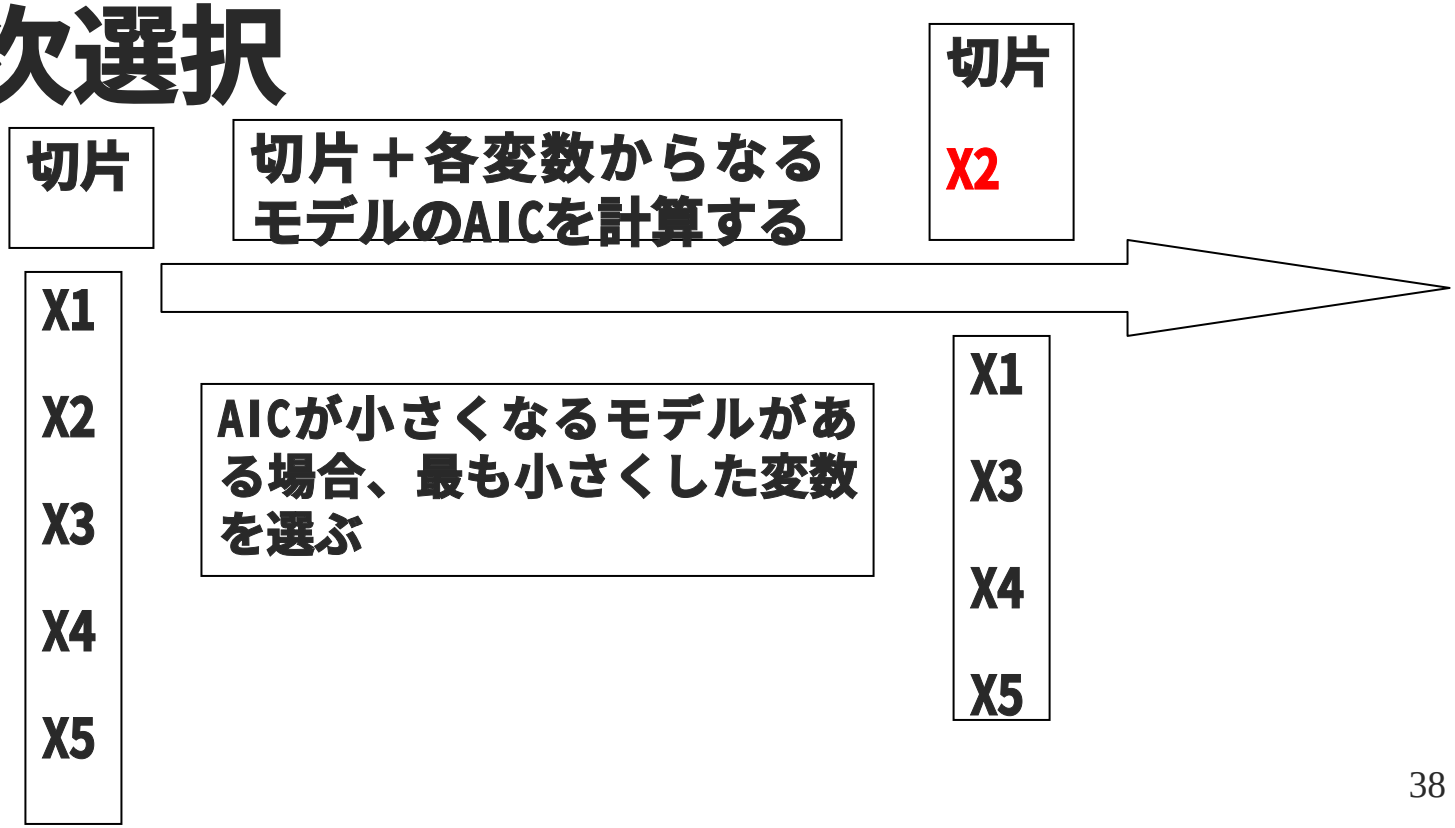
逐次選択の方法がより細かく設定可能 (続き 2)

- 今までは、



逐次選択の方法がより細かく設定可能 (続き 3)

- GLMSELECTにおいてAICで逐次選択

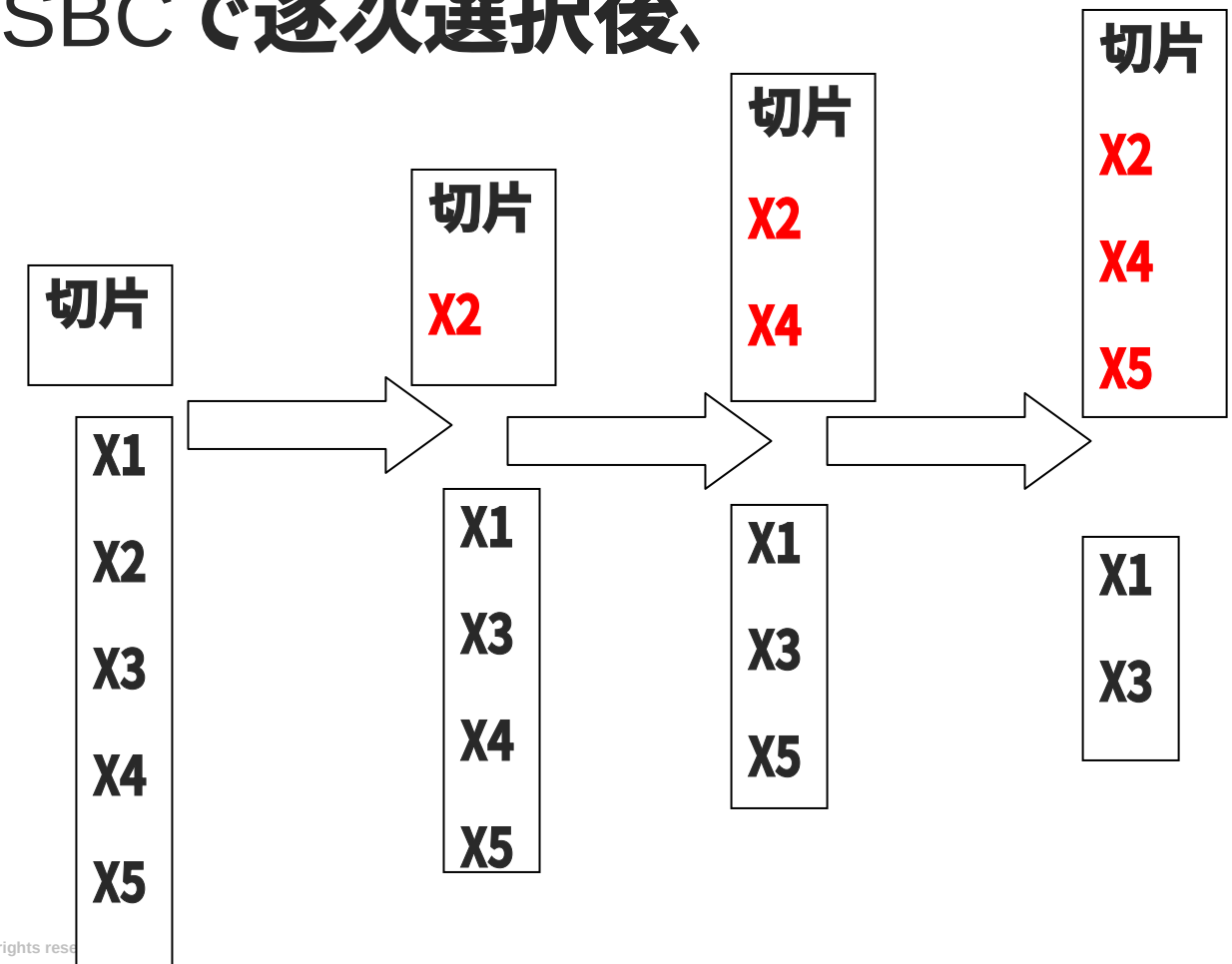


PROC GLMSELECTのプログラム

```
PROC GLMSELECT DATA=WORK.Test;
  MODEL Y=X1 - X5
      /SELECTION=FORWARD(SELECT=AIC);
RUN;
```

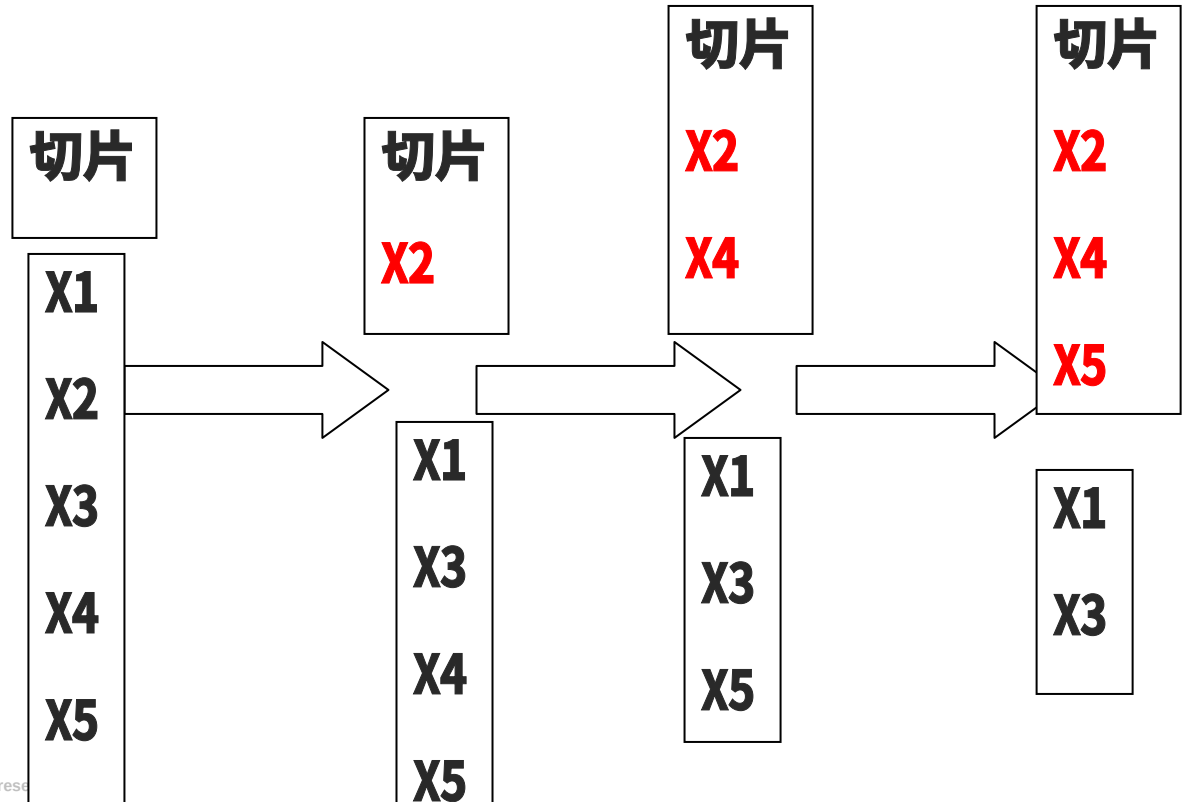
逐次選択の方法が より細かく設定可能 (続き 4)

- SBCで逐次選択後、



逐次選択の方法がより細かく設定可能 (続き 5)

- バリデーションデータに対して最もエラーが小さいモデルを選択する



プログラム

```
PROC GLMSELECT DATA=WORK.Test
                              VALDATA=WORK.Valid;
  MODEL Y=X1 - X5
      /SELECTION=FORWARD(SELECT=SBC
                                 CHOOSE=VALIDATE);
RUN;
```

MODELステートメントのオプション

- SELECTION=FORWARD
- SELECTION=BACKWARD
- SELECTION=STEPWISE
- SELECTION=LAR
- SELECTION=LASSO

デフォルト設定では、

- SBCという情報量規準に基づいて変数が取捨選択されます。
  (SELECT=SBC)

SELECTION=のサブオプション

サブオプション	FORWARD	BACKWARD	STEPWISE	LAR, LASSO
STOP=	X	X	X	X
CHOOSE=	X	X	X	X
STEPS	X	X	X	X
MAXSTEPS	X	X	X	X
SELECT=	X	X	X	
INCLUDE=	X	X	X	
SLENTRY	X		X	
SLSTAY		X	X	
DROP			X	
LSCOEFFS				X

GLMSELECTの主な機能(3)

3.LAR(2002/2004) and Lasso(1996)

LARとLasso(1)

- 実データを使った通常の線形回帰では、説明変数間に”相関”が見られる（多重共線性）のが常ですから、

LARとLasso(2)

- **“安定的”なモデルを作ることは悩ましい問題です。**

LARとLasso(3)

- 先に取り上げましたが、変数選択を行って変数を減らすことは、よく行われます。
- “小さなモデル”で表現できれば、それに越したことはありません。

LARとLasso(4)

- “Lasso”**は、回帰における変数選択と、いわゆるリッジ回帰のアイデアをミックスしたあたりに動機があります。**

LARとLasso(5)

- LARは、**変数増加法的なアプローチである、「Stagewise回帰」の改良形でもあり、Lassoとも関連があります。**

Lassoとは？

- Least Absolute Shrinkage and Selection Operator
- **オリジナルは、**

“Regression Shrinkage and Selection via the Lasso”, Tibshirani, *J. R. Statist. Soc.* B, 1996

Lassoとは？（続き）

- 回帰モデルにおける通常の最小2乗法では

$$\min \Sigma (y - \beta x)^2$$

となるようなβを見つけることです。（以降、式はやや省略形です。）

Lassoとは？（続き 2）

- Lasso では、制約（ペナルティ項）がついています。

$$\min \Sigma(y - \beta x)^2$$

$$subject\ to \Sigma \mid \beta \mid \leq t$$

Lassoとは？（続き 3）

- 制約式の t を大きくすれば、通常の線形回帰と同等です。

$$\min \Sigma(y - \beta x)^2$$

$$subject\ to \Sigma \mid \beta \mid \leq t$$

Lassoとは？（続き 4）

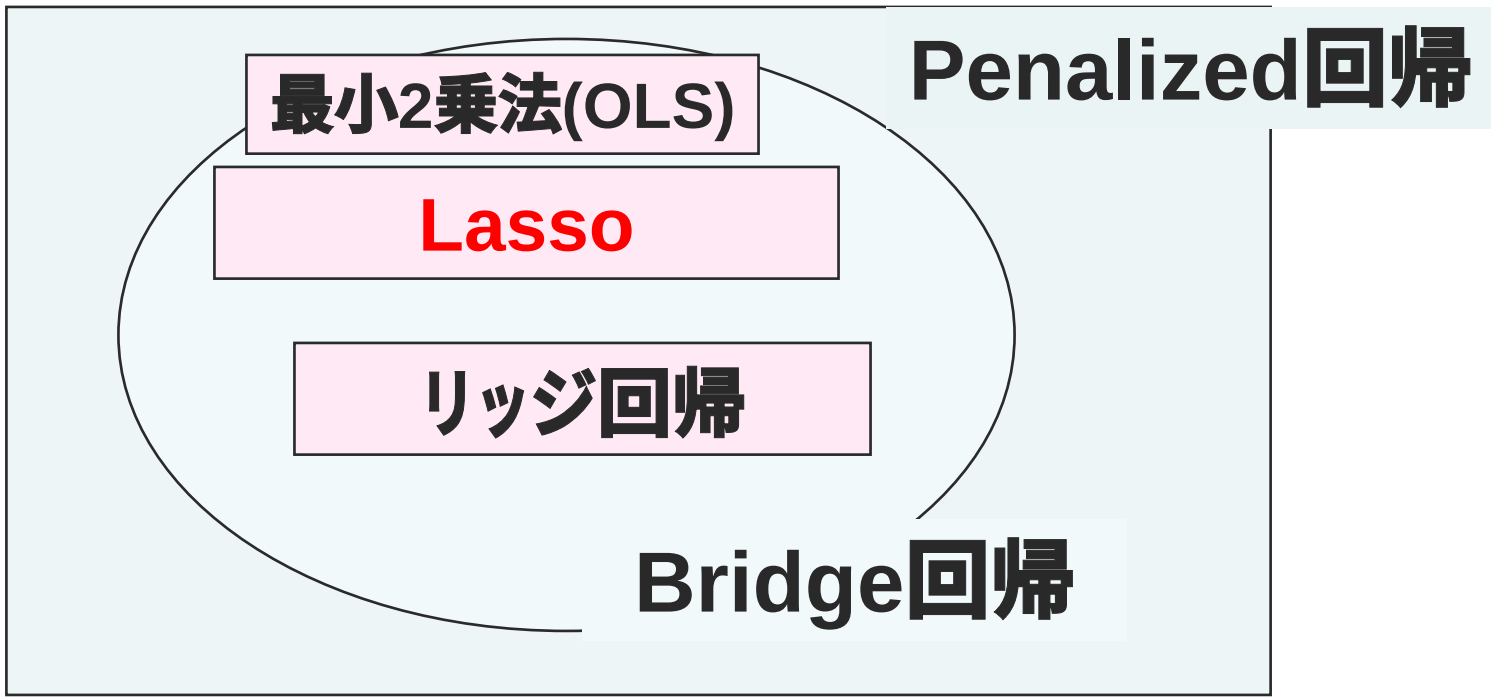
- 制約式の t を小さくすると、幾つかの回帰パラメータの0と異なり、残りは全て0となる性質があります。

Lassoとは？（続き5）

- GLMSELECTでは、Efronらの方法（後で出てきます）に基づき、LARと関連づけてLassoの解を求めています。

Lassoとは？（続き 6）

- Lassoの位置づけ



LARとは？

- Least Angle Regression
- By Efron, Johnstone, Hastie and Tibshirani

http://www-stat.stanford.edu/~tibs/ftp/LeastAngle_2002.pdf

元に戻って、LARとは？（続き）

- 基本的には、モデルに変数を1個ずつ入れていきます。
- 「ある角度」が等しくなるよう、選択していきます。

LARとは？（続き 2）

- 詳細に関しては、Efronらの論文を参照してください。
- 理屈は、簡単そうでそれなりに難しいです。

LassoとLARの利用方法

- SELECTION=LASSO
- SELECTION=LAR

GLMSELECTの主な機能(4)

4. 入力データを学習・検証・テストデータに分割し、モデルを検証・評価

入力データを学習・検証・テストデータに基づいたモデルの検証・評価

- 各データに対して平均2乗誤差(ASE)を計算します。
- 検証データは変数選択の過程で利用することもできます。

入力データを学習・検証・テストデータに分割し、モデルを検証・評価 (続き)

- 学習・検証・テストデータを直接指定

```
PROC GLMSELECT DATA=Test
               VALDATA=Test2
               TESTDATA=Test3;
```

(以下略)

入力データを学習・検証・テストデータに分割し、モデルを検証・評価 (続き 2)

- PARTITIONステートメントで元データを分割

```
PROC GLMSELECT DATA=WORK.Test;
  MODEL Y=X1-X5
        / SELECTION=STEPWISE(SELECT=VALIDATE);
  PARTITION FRACTION(VALIDATE=0.3);
RUN;
```

GLMSELECTの主な機能(5)

5. その他いろいろ

その他いろいろ

- ODS Graphicsでグラフを作成
- BY変数によるマルチスレッド対応
- 選択された説明変数名をマクロ変数として取り込む → REGやGLMプロシジャと連携可能
- スコアリング機能

GLMSELECTはここまでです。

- おさらいです。

GLMSELECTの確認

- 名前の通り、

PROC GLM + Selection

のイメージです。

GLMSELECTの機能の確認

1. カテゴリ変数を含む回帰で変数を”逐次選択”
2. 逐次選択の方法をより細かく設定可能
3. LAR(2002/2004) and Lasso(1996)
4. 入力データを学習・検証・テストデータに分割し、モデルを検証・評価
5. その他いろいろ

GLMSELECTの入手方法の確認

- 日本のGoogleかYahoo!でGLMSELECTで検索を。
- SAS Japanの紹介サイトが一番上にヒットするはずです（多分）。

GLMSELECTの参考文献

- “Introducing the GLMSELECT PROCEDURE for Model Selection” by R. Cohen, SAS

  http://www2.sas.com/proceedings/sugi31/207-31.pdf

- “Least Angle Regression” by Efron, Hastie, Johnstone and Tibishirani, Stanford Univ.

  http://www-stat.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf

SAS9.2の話。

- 来年 (2007年) にリリースという噂。
- 分析に関する追加機能がたくさんあるらしい。

SAS9.2の話（続き）。

- さらに興味のある方は、SUGI31のこちらをどうぞ。

“You Can’t Stop Statistics: SAS/STAT Software Keeps Rolling Along”

http://www2.sas.com/proceedings/sugi31/185-31.pdf

アドインのプロシジャ

- PROC GLMSELECTや PROC QUANTREGのように、後からインストールして利用可能となるプロシジャは、他にもあります。

たとえば PROC GLIMMIX

- いわゆる「一般化線形混合モデル」を扱うプロシジャです。
- SAS Learning Session 2006 でお話があるようです。

たとえば PROC PANEL

- いわゆる「パネルデータ」に対する解析。SAS/ETSのTSCSREGプロシジャの後継という扱いです。

PROC QUNATREGとは？

- “Quantile Regression”を行うものです。
- “分位点回帰”と訳されることが多いようですが、まだ定訳は存在しないようです。私は、とりあえずこの表現を使用しています。

分位点回帰とは？

- 分位点回帰は、たとえば以下のソフトウェアでも実現できるらしいです。
  - Stata, TSP, Xplore など
  - S-PLUS/R のサンプルコード

分位点回帰とは？（続き）

- 日本でも、研究者の間では取り上げられているようですが、まだそれほど一般には普及していない模様です。

分位点回帰とは？（続き 2）

- 理論は、結構昔からあったようです。

Koenker, R. and Bassett, G. W. (1978), “Regression Quantiles,” Econometrica, 46, 33–50.

なぜ分位点回帰？

- 回帰モデルにおける通常の最小2乗法には、計算は比較的容易なのですが、弱点が色々とあります。

なぜ分位点回帰？ (続き)

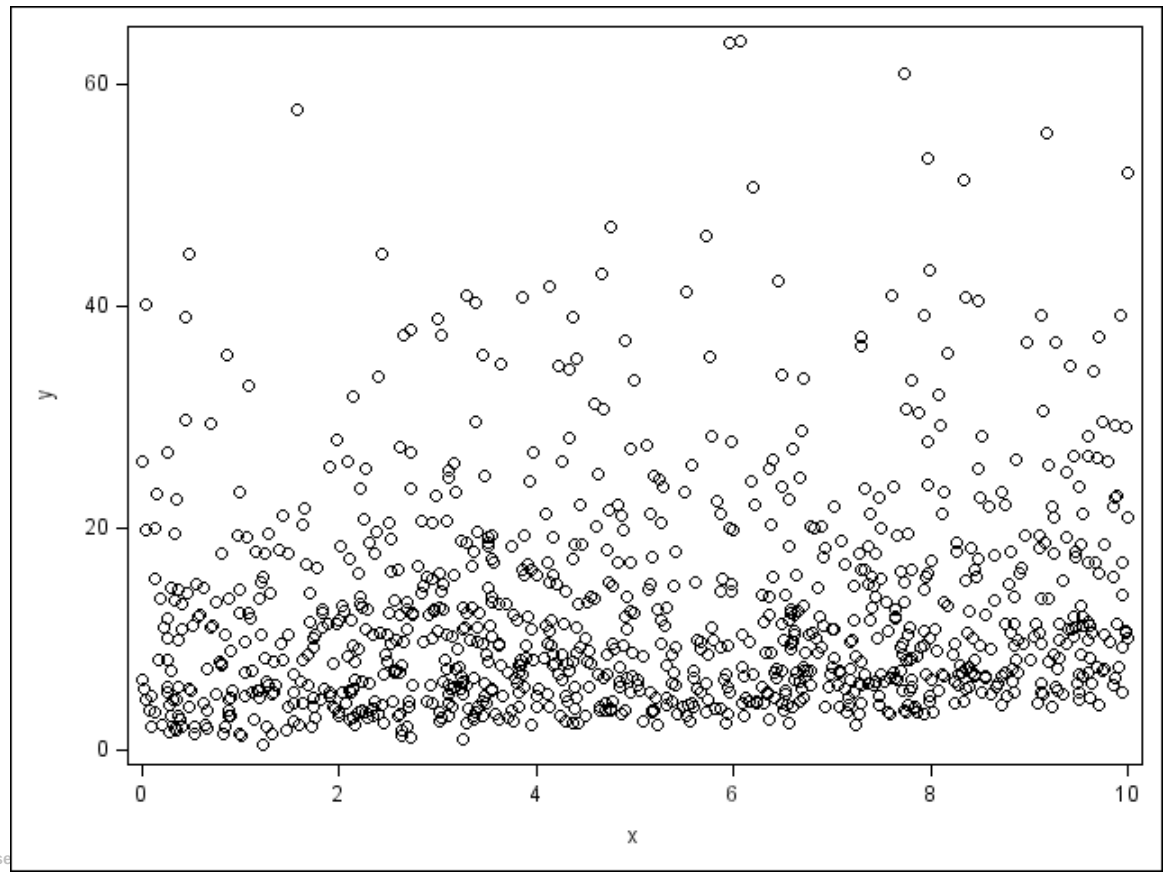
- たとえば、誤差項は、独立同分布 (正規) である、などの仮定が与えられています。

なぜ分位点回帰？（続き 2）

- 誤差項が正規分布に従う、ということはなかなか。。。
- 外れ値などにも弱いですし。。。

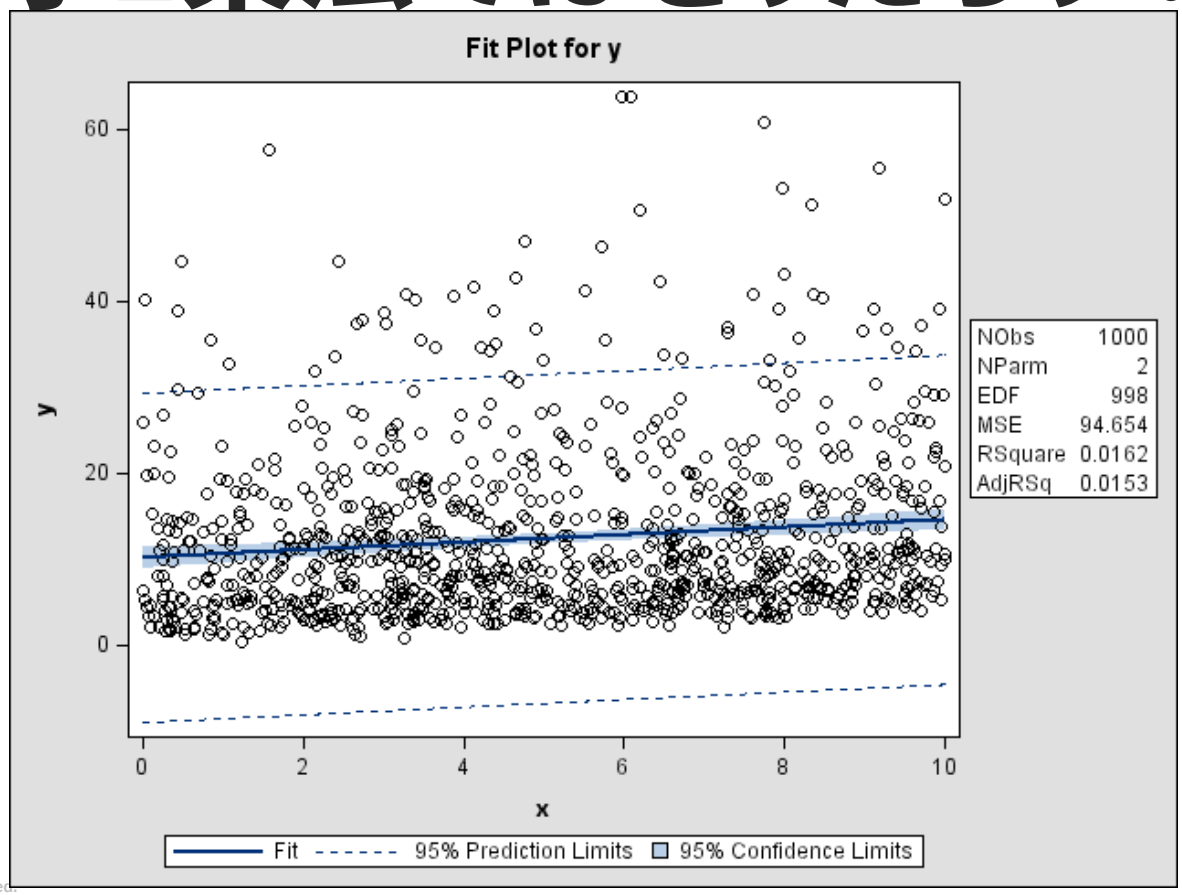
なぜ分位点回帰？ (続き 3)

- たとえば。



なぜ分位点回帰？(続き 4)

- 最小2乗法ではどうだろう？

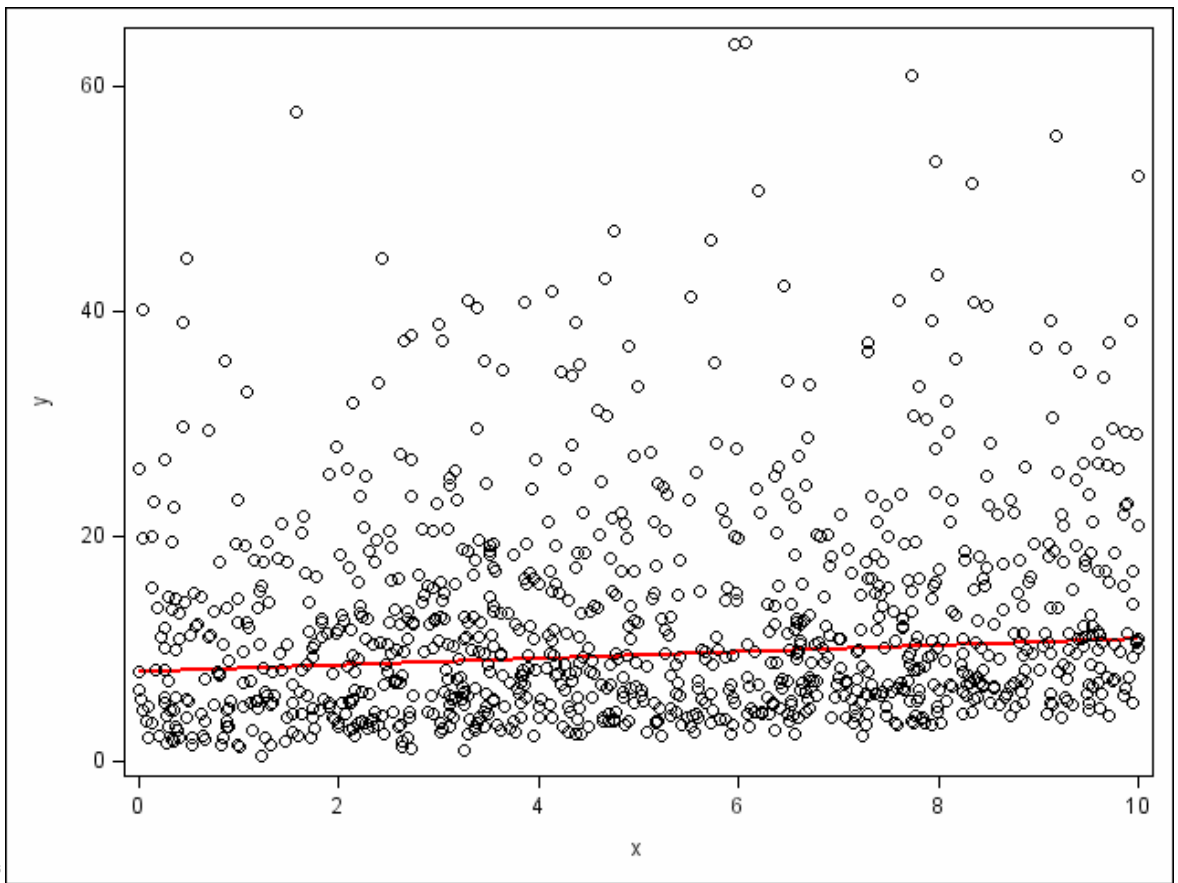


なぜ分位点回帰？(続き 5)

- 普通の線形回帰は、結局のところ”平均”について着目していることになっています。
- でも、中央値とか端の方の分位点に関心があることも。

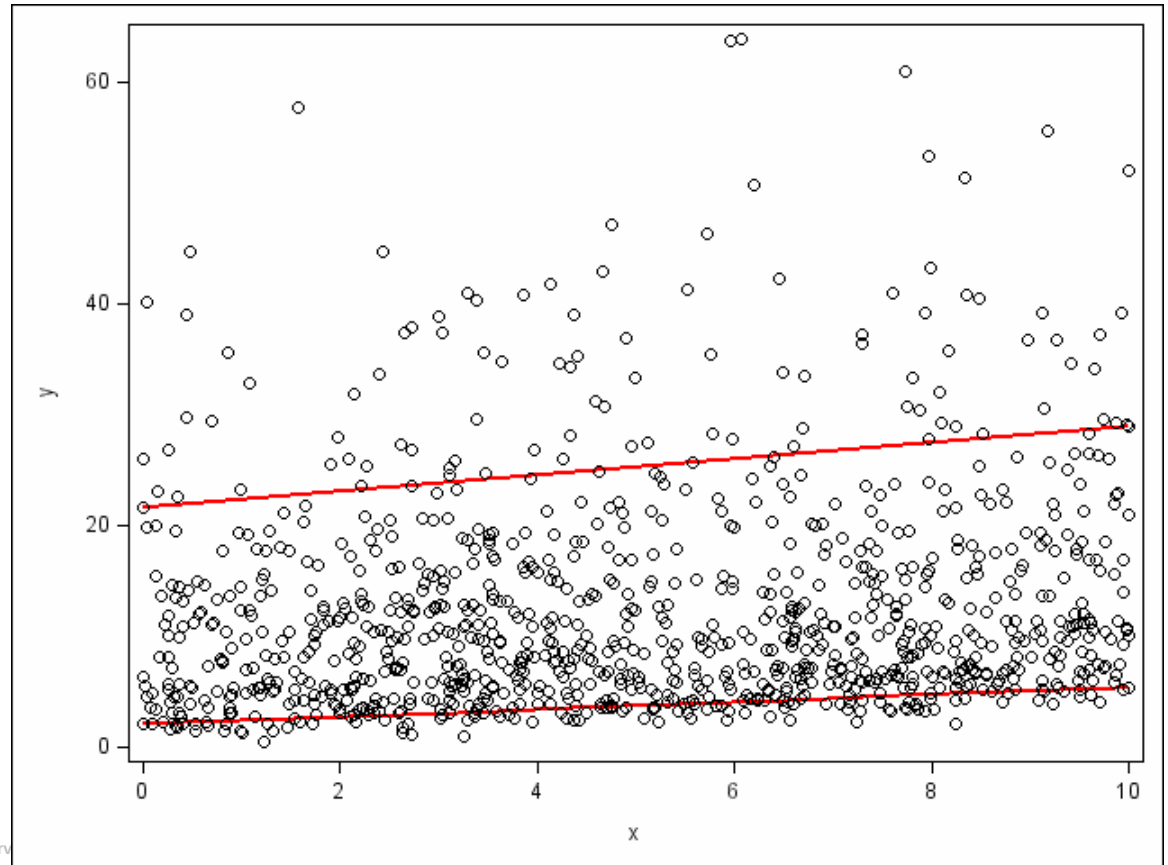
なぜ分位点回帰？(続き 6)

- たとえば、中央値に関しては



なぜ分位点回帰？（続き 7）

- 端の方を説明するには、



そこで、QUANTREG

- 米国SAS Institute Inc.では、新たにQUANTREGプロシジャを開発しています。
- SAS9.1.3において評価版として公開しています。

QUANTREGは、

- 今のところ、32bit Windows環境でのみ利用可能なモジュールが提供されています。

QUANTREGを使うには

- 以下のURLあたりから、EXEファイルをダウンロードしてインストールしてください。

http://support.sas.com/rnd/app/da/quantreg.html

それよりも、

- 日本のGoogleかYahoo!でQUANTREGで検索を。
- SAS Japanの紹介サイトが一番上にヒットするはずです（多分）。

で、分位点回帰とは、

- もう少し正確なところ、ということですが、、、
- まず、平均、中央値、分位点のおさらいを。

平均の考え方

- 平均は、全て足して、データの個数で割った答え、ですが、
- 平均は以下の式を最小とするξと考えることもできます。

$$\sum_i (y_i - \xi)^2$$

中央値の考え方

- 中央値は、データを並べ、その真ん中にあるデータのことですが、
- 中央値は、以下の式を最小とするξと考えることもできます。

$$\sum_i | y_i - \xi |$$

分位点の考え方

- じゃあ、25%点とか、95%点はどうやって計算する？

- 中央値と同じように、データを並べて、探せば良いわけですが、

分位点の考え方 (続き)

- τ -分位点 (パーセント点) は、以下の式を最小とする ξ と考えることができます。

$$\sum_i (y_i - \xi) \cdot (\tau - 1(y_i - \xi > 0))$$

通常の線形回帰は、

- 通常の線形回帰は以下の式を最小とするβを探す

$$\sum_i (y_i - X\beta)^2$$

$$\updownarrow$$

$$\sum_i (y_i - \xi)^2$$

中央値に基づく分位点回帰は？

- “中央値回帰”は、以下の式を最小とするβを探す

$$\sum_i |y_i - X\beta|$$

$$\updownarrow$$

$$\sum_i |y_i - \xi|$$

そこで、分位点回帰は、

- “τ-分位点回帰”は、以下の式を最小とするβを探す

$$\sum_i (y_i - X\beta) \cdot (\tau - 1(y_i - X\beta > 0))$$

$$\updownarrow$$

$$\sum_i (y_i - \xi) \cdot (\tau - 1(y_i - \xi > 0))$$

なお、

- “中央値回帰”は、分位点回帰の特殊例と考えることができます。

また、

- **中央値に基づく分位点回帰は、**Least Absolute Value(LAV) regression**として、**SAS/IML**のサブルーチンとして存在していました。**
- L1-**回帰などとも呼ばれます。**

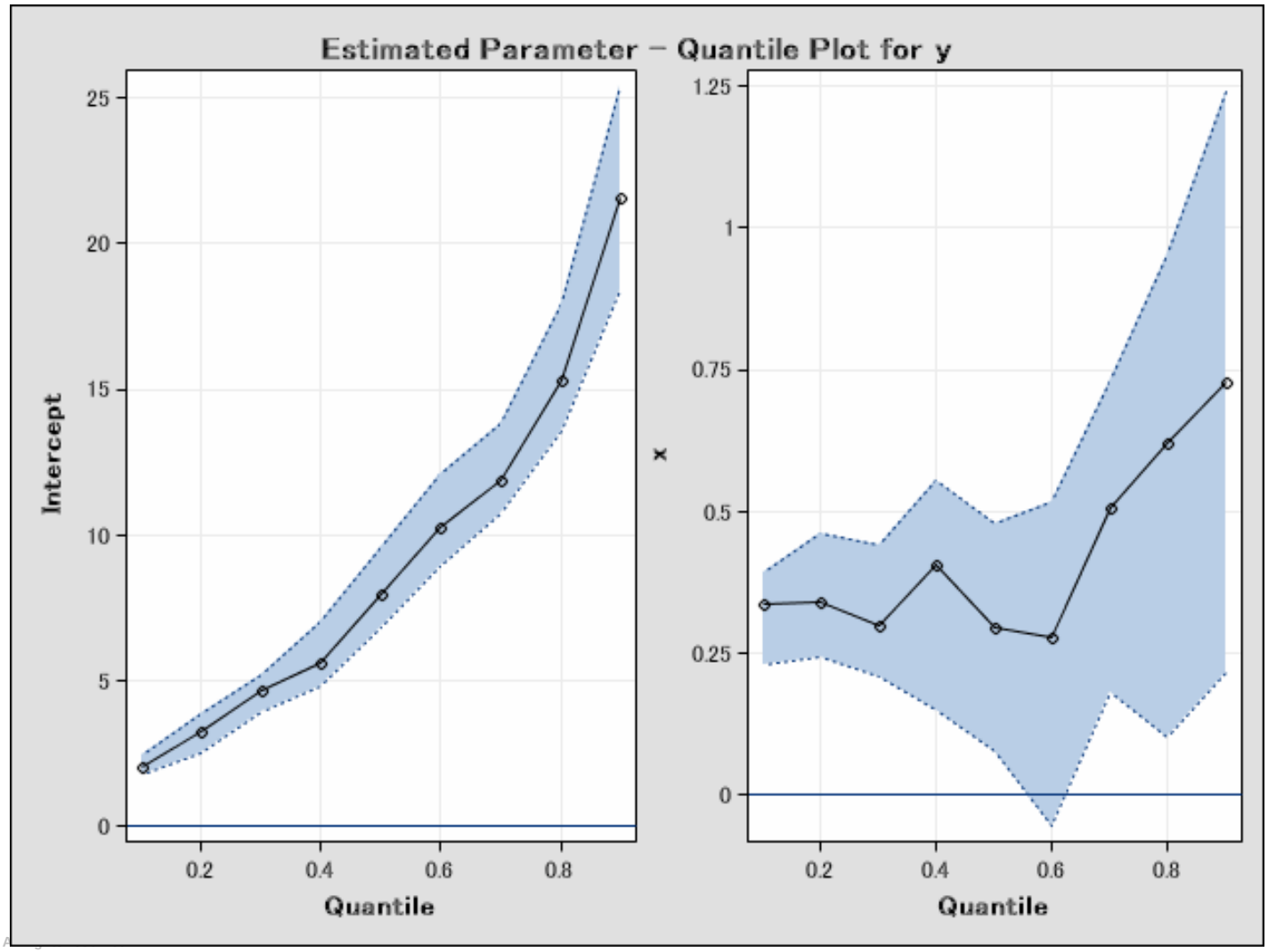
QUANTREGプロシジャは

- このような分位点回帰について、まとめて面倒を見ることができます。

QUANTREGの使用例

```
ODS GRAPHICS ON;
PROC QUANTREG DATA=Work.test2;
  MODEL Y=X /
                QUANTILE=0.1 to 0.9 BY 0.1
                 PLOT=QUANTPLOT;
RUN;
```

“Regression Quantile Process”



出力の見方

については、後述の参考文献を。

分位点回帰の解を求める

- τ-分位点回帰”は、以下の式を最小とするβを探す、でしたが、

$$\min_{\beta} \sum_{i} (y_i - X\beta) \cdot (\tau - 1(y_i - X\beta > 0))$$

解を求める (続き)

- 実は、このβは線形計画問題 (Linear Programming, LP)に落とすことができます。
- 線形計画問題に対しては、色々なアルゴリズムがあります。

PROC QUANTREGで利用できる最適化手法

- PROC QUANTREGステートメントで、ALGORITHM=オプションを使用します。

PROC QUANTREGで利用できる最適化手法 (続き)

- ALGORITHM=SIMPLEX

  →単体法 (シンプレックス法) で解く

- ALGORITHM=INTERIOR

  →内点法で解く

PROC QUANTREGで利用できる最適化手法 (続き 2)

- ALGORITHM=SMOOTH

  →”スムージング法”で解く

  →連続微分可能性をうまく保つように目的関数を調整する、PROC QUANTREGの開発者のオリジナル(?)

外れ値とLeverage Pointsの検出

- “ロバストな距離”に基づいて、Leverage Pointを報告します。

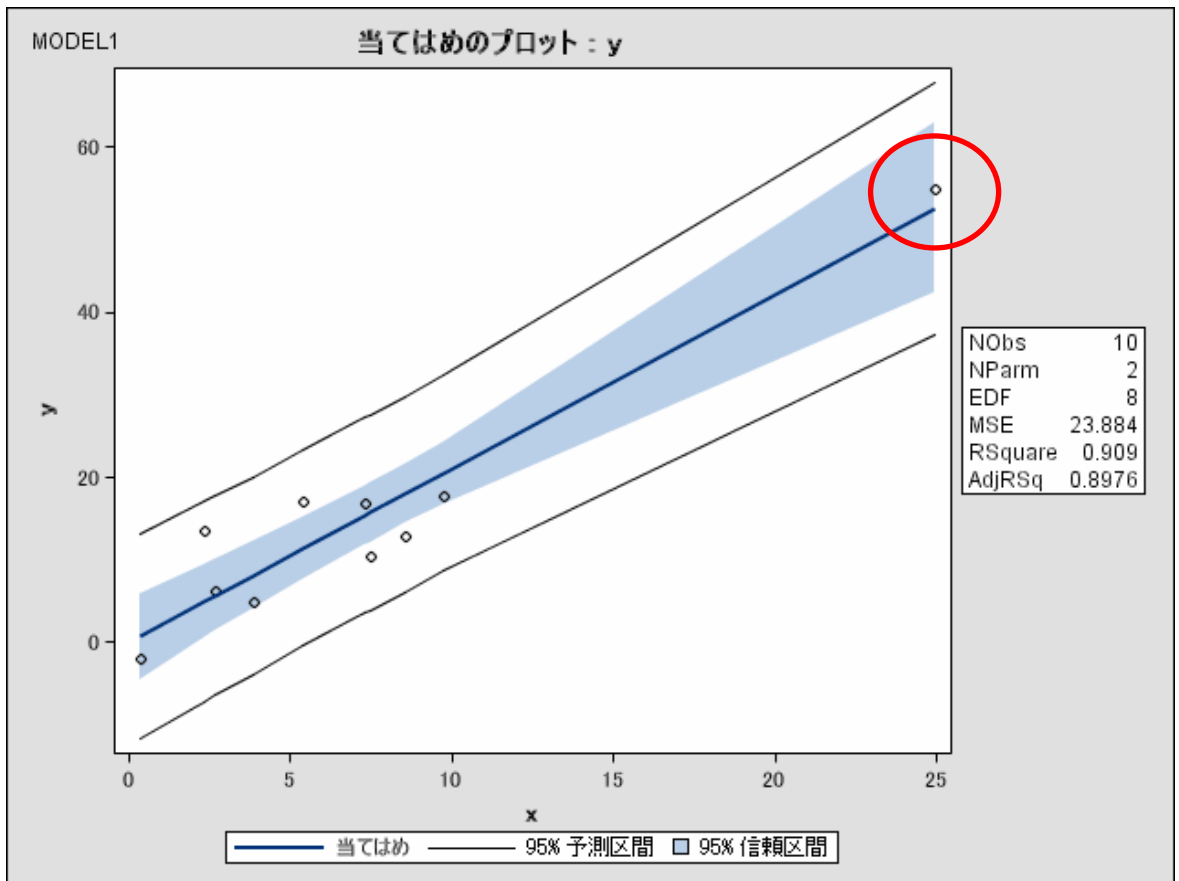
- 得られた分位点回帰の結果から、外れ値(Outliers)を報告します。

Leverage Points とは？

- Leverage Pointとは、説明変数の空間において「外れた」値のことです。

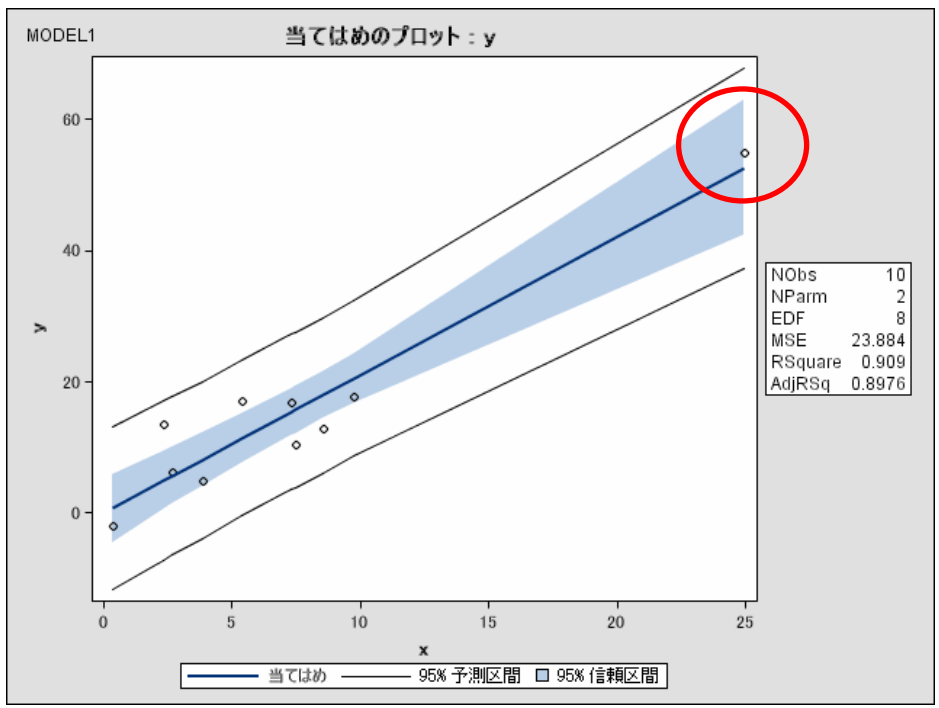
Leverage Points とは？（続き）

- 端的には、こういった点のことです。



Leverage Points とは？（続き 2）

- こういったX方向で外れた点があると、よい推定ができないかもしれません。

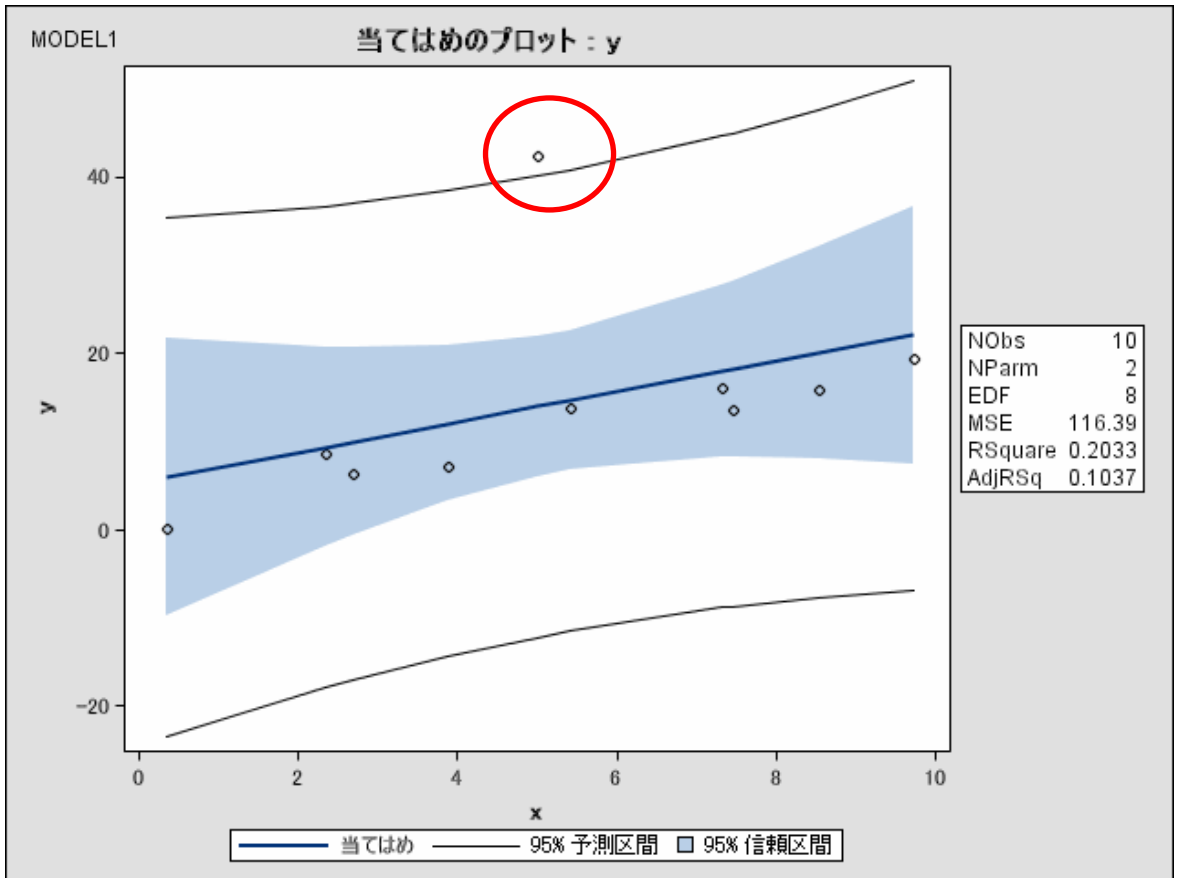


外れ値とは？

- 被説明変数に関して、大きく外れた点です。

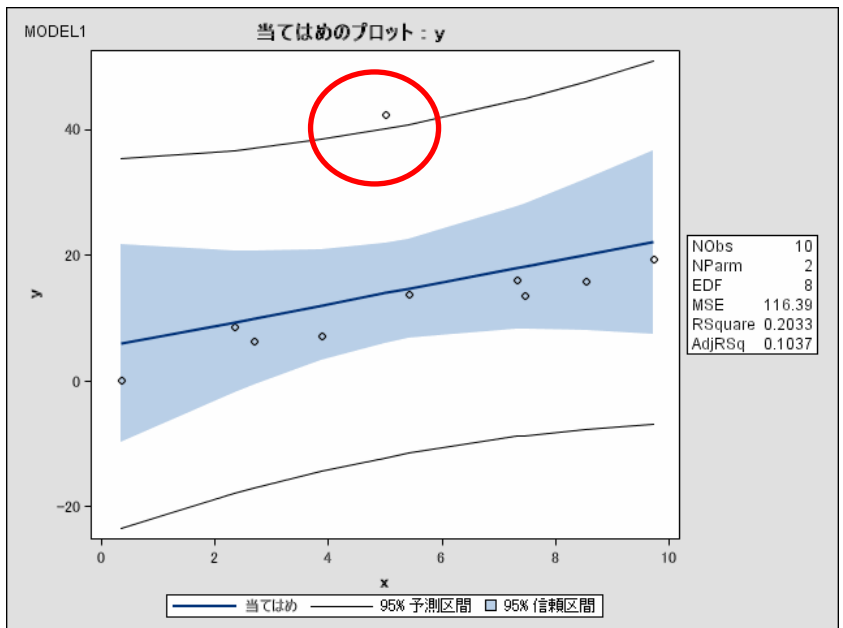
外れ値 とは？ (続き)

- 端的には、こういった点のことです。



外れ値とは？ (続き 2)

- こういった、y方向に外れた点があると、推定結果がこの点に引っ張られてしまいます。



外れ値とLeverage Pointsの検出

- 一般に、これらの点には色々な意味で注意する必要があります。
- 実は、分位点回帰は外れ値には強いですが、Leverage Pointsには強くありません。

そういうときは、

- ROBUSTREGプロシジャが役に立つかもしれません。

QUANTREGのその他の機能

- 推定されたパラメータに対して
  - 信頼区間を算出
  - 相関行列、分散共分散行列を算出
  - パラメータ=0を帰無仮説とした検定
- ODS Graphics によるグラフの描画

QUANTREGのその他の機能 (続き)

- 処理のパラレル化に対応しています。
- これは、PERFORMANCEステートメントで設定します。

QUANTREGに関する参考文献

- TUTORIAL ON QUANTILE REGRESSION

By Xuming He and Ying Wei, ENAR 2005

http://www.stat.uiuc.edu/~x-he/ENAR-Tutorial.pdf

- An Introduction to Quantile Regression and the

QUANTREG Procedure, By Colin Chen, SUGI 2005

http://www2.sas.com/proceedings/sugi30/213-30.pdf

QUANTREGに関する参考文献 (続)

- Quantile Regression, By Koenker, R. and K. Hallock, (2001) Journal of Economic Perspectives, 15, 143-156. http://www.econ.uiuc.edu/~roger/research/rq/QRJEP.pdf

再度のお願い

- **これまでご紹介したPROC GLMSELECT とPROC QUANTREG は、SAS 9.1.3 では評価版の扱いとなっています。正規版と同等にお考えにならないようお願いします。**
- **次期リリースSAS 9.2 では、正規版のプロシジャとなる予定ですが、その際に構文の変更などが行なわれる可能性があります。**

今日のおさらい

- 線形回帰モデルにおける新しい変数選択法を提供する

GLMSELECT

それと、

- 分位点回帰に対応した

QUANTREG

今日のお話しはここまでです

- ありがとうございました。
- ご質問？

THE POWER TO KNOW®