渡辺研講究　3.3.節概説

東京工業大学　情報理工学院　数理・計算科学コース

発表者　16M30250　林　直輝

2016/5/12　発表

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

自己紹介

名前：林　直輝(Hayashi Naoki)

略歴：2016年3月 東京工業大学 生命理工学部 生命情報専攻 卒業
- 入学は2013年度 i.e. 早期卒業
- 1類に落ち、転類もできなかった
- カリキュラムとして数学の講義なかったので、工学部と数学科で聴講
- 卒研テーマはヒト腸内細菌と疾病の関係性データベース・ネットワーク構築

研究的興味：応用数学

ポリコレな趣味的興味：平坦な曲面の構成と可視化、ラーメン
- 東工大ロボット技術研究会で理論よりのことをやって遊んでいる
- 一方で高レイヤ実装が好き　ProcessingとかMATLABとかツクールとか

座右の銘：

自己紹介

名前：林　直輝(Hayashi Naoki)

略歴：2016年3月 東京工業大学 生命理工学部 生命情報専攻 卒業
- 入学は2013年度 i.e. ***早期卒業***
- 1類に落ち、転類もできなかった
- カリキュラムとして数学の講義なかったので、*工学部と数学科で聴講*
- 卒研テーマはヒト腸内細菌と疾病の関係性データベース・ネットワーク構築

研究的興味：応用数学

ポリコレな趣味的興味：平坦な曲面の構成と可視化、ラーメン
- 東工大ロボット技術研究会で理論よりのことをやって遊んでいる
- 一方で**高レイヤ実装**が好き　ProcessingとかMATLABとかツクールとか

座右の銘：**Done is better than best.**
- 先んずれば人を制す/**厚い皮膚より速い足**/*嘆くヒマがあるなら戦うことだ*　etc.

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

はじめに

3.2.節では最小2乗法を扱った。

次の2点で不満が残る:

・予測精度

・結果の解釈

はじめに

3.2.節では最小2乗法を扱った。

次の2点で不満が残る:

- 予測精度
- 結果の解釈

はじめに

最小2乗法の特徴:

バイアスが小さく、バリアンスが大きい

以下の事実が知られている[統計数理の講義]:

- バイアス0　バリアンス$\sigma^2(X^tX)^{-1}$
- バイアス0つまり不偏推定量では上のバリアンスが最小
- 不偏性を犠牲にすればより小さなバリアンスが得られる

はじめに

Shrinking(3.4.節)やいくつかの係数を0にすることで予測精度が改善される

バイアスを僅かながら犠牲にしてバリアンスを小さくすることで予測精度を改善する

はじめに

3.2.節では最小2乗法を扱った。

次の2点で不満が残る:

・予測精度

・結果の解釈

はじめに

予測変数が多いとき

結果への寄与の大きい変数に着目したい

そのような変数だけの小さな部分集合が欲しい

しかし最小2乗法では全変数を用いる

はじめに

本節では線型モデルの部分集合選択を概説する

後半の節は次数落とし手法だけでなくバリアンスの制御手法も扱い、線型モデルに限らない一般のモデルについて扱う→7章

部分集合選択は変数の部分集合だけを考える

数多の部分集合選択手法がある

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

3.3.1. Best subset slct

Best subset selection（総当り法）

- 変数の部分集合を組合せ最適化で選択
- RSSが最小となる部分集合を選択
- 部分集合の濃度$k \in \{0, .., p\} =: \mathbb{N}_{\leq p}$毎に組合せ最適化を行う
- 変数の最大数pが30~40のとき有効
- 選ばれた部分集合はネストされると限らない

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \qquad (3.2)$$

3.3.1. Best subset slct

Figure3.5.（右図）は前立腺がんの例でBest subset selectionを実行したときの結果を示している

縦軸がRSSなため、紙面下部に近いほど目的を達成できている

横軸は部分集合の濃度

最下部の赤い曲線が最適解

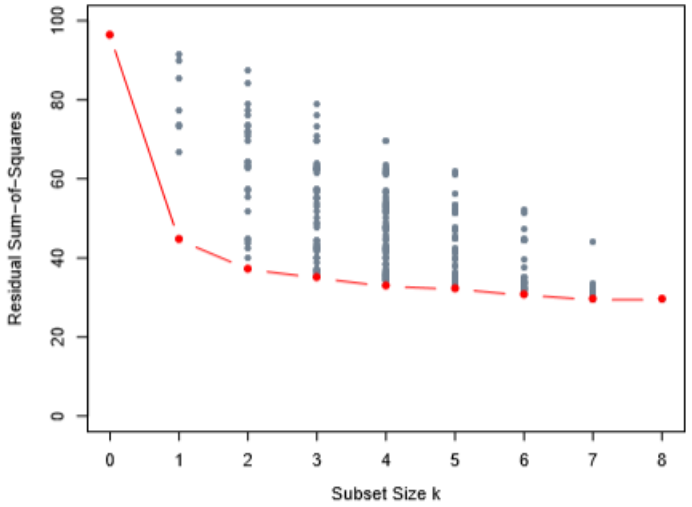


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

3.3.1. Best subset slct

最適解の曲線は部分集合の濃度からRSSへの函数として、単調減少なため最適なkを選べない

- kの選択方法はバイアス・バリアンスのトレードオフを孕む問題
- 使われる基準は様々であり、AICが一般的
- 詳細と他の手法は7章

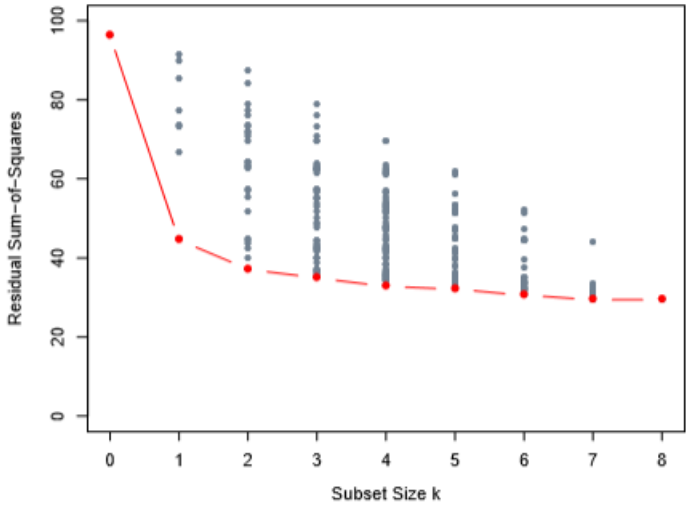


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

3.3.2. Best subset slctの弱点

Best subset selectionは組合せ最適化を総当りで解いている

→p>40では実用的でない

→(実用面で)より良い解の探索経路が必要

3.3.2. Forward stepwise slct

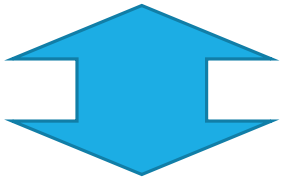
Forward stepwise selection

- 切片から初めて以降RSSが最小になるように予測変数をモデルに加えていく
- 貪欲法により組合せ最適化を近似的に解く
- フィットした変数にQR分解を用いることで、次に加える変数を素早く計算（演習3.9.）

3.3.2. Forward stepwise slct

Forward stepwise selection

- 貪欲法故、部分的な最適化(sub-optimal)でしかない



次の理由で使われる

- 計算可能性~pが大きく(特にp≫N)ても計算できる
- 統計的理由~束縛の強さ故、バリアンスがbest subsetより小さい(バイアスは大きくなることがある)

3.3.2. Backward stepwise slct

Backward stepwise selection

- 全変数を考慮したモデル(full model)から始める
- フィットへの影響が少ない予測変数から削る
- Z-scoreが小さい(係数0でないとは言い難い)変数から削っていく(演習3.10.)
- N>pのときでなければ使えない

$$\text{Z-score} \quad z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \qquad (3.12)$$

$$v_j = \left(X^T X\right)^{-1}(j,j)$$

3.3.2. 比較の図

Figure3.6.(右図)はbest-subset回帰、forward-、backward-stepwiseによる小規模シミュレーション結果を示している

N=300,標準正規分布に従う変数(p=)31

best-subset回帰、forward-、backward-stepwiseはいずれも非常に似通った結果を示した

緑の曲線Forward stagewiseについては3.3.3.で述べるが、ここでは最小化に時間がかかっている

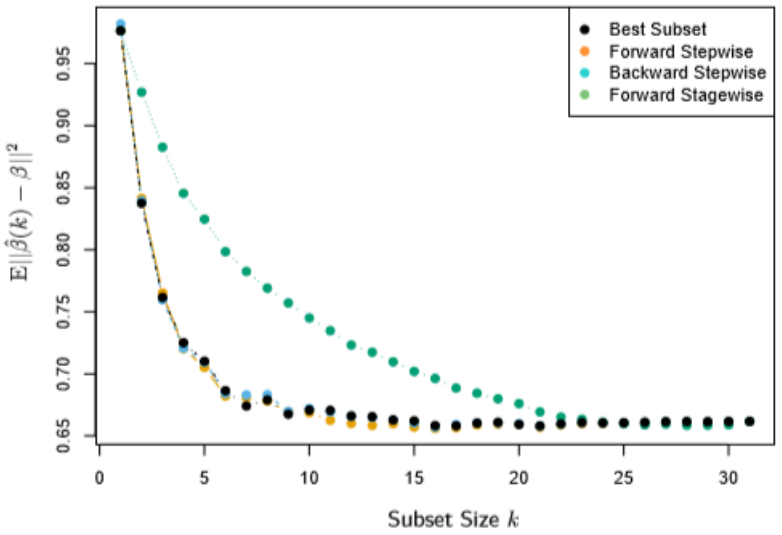


FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat{\beta}(k)$ *at each step from the true* β.

3.3.2. アラカルト

Hybrid stepwise selection

- ForwardとBackwardの双方を各ステップで検討
- 統計フリーソフトRのstep函数など

F-統計

- 多様さを与える変数を追加、そうでないものを除去
- 多重学習問題で正しく計算できない時代遅れ手法
- 探索過程で計算しないため標準偏差が有効でない
- →ブートストラップ(8.2.節)

3.3.2. Remark

多層的な予測変数を記したダミー変数のように、変数はグループ化されがちである

Rのstep函数のような、賢いstepwiseの手順は一度に全てのグループに対して自由度を適切に計算するように変数の追加または除去を行う

目次

3.3.3. Forward stagewise slct

Forward stagewise selection

- Forward stepwise slct.よりも更に束縛した手法
- $\bar{y}$に等しい切片から始める
- 中心の予測変数は係数を全て0とする
- 各ステップで残差と最強の相関を持つ変数を識別
- 線型回帰の選択変数の残差の係数を計算し、現在の係数を追加する
- 相関がなくなるまで続ける
- p<Nのときの最小2乗法

3.3.3.　Fwd stpw slctとの違い

変数追加時に調整される変数が存在しない

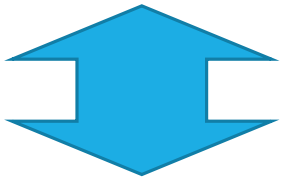
→pよりもステップ数が大きくなる

→最小2乗のフィットに達するまでに時間がかかる

→使えない手法とされた

3.3.3. 逆転劇

時間がかかるため歴史的に使えない手法とされた



高次元問題で有用と判明、不採用が覆った

とりわけ超高次元問題において分散を減少させる

詳細は3.8.1.節で扱う

3.3.3. Figure3.6.での評価

Figure3.6.(右図)では最小化に時間がかかっている

すべての相関係数を10^{-4}未満にするために1000》p=31超の段階を踏んだ

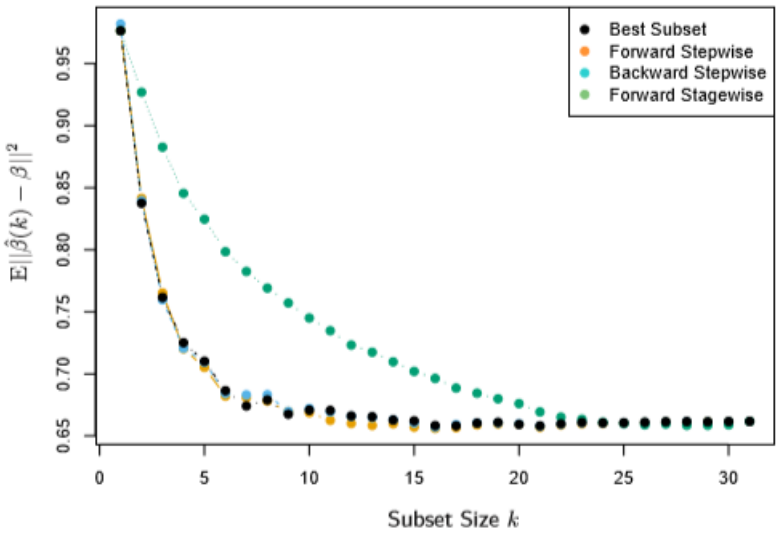


FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat{\beta}(k)$ *at each step from the true* β.

目次

3.3.4. 前立腺がんの例

Table3.3.(右表)は
- 最小2乗法(LS)
- Best subset slct
- リッジ回帰
- Lasso
- 主成分回帰分析(PCR)
- 部分最小二乗法(PLS)

の前立腺がんの例における各予測変数の係数を示している
- LSとbest subset以外は後節で扱う
- 10fold-クロス・ヴァリデーションに基づいて予測誤差の推定値の最小化をするように選ばれたものである

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

3.3.4. Cross-Validation

10foldの場合の簡単な説明

- パラメータのフィットにデータの9/10を使う
- 残り1/10で予測誤差を計算する
- 1/10の部分を入れ替えていき、結果の予測誤差を平均する

→予測誤差推定量曲線をパラメータから予測誤差への函数として得る

詳細は7.10節

3.3.4. Remark

既にデータをサイズ67の学習集合とサイズ30の試験集合に分けている

縮退させるパラメータの選択は学習過程の一部で行う

→クロス・ヴァリデーションは学習集合に適応される

試験集合は選択したモデルの挙動を判定することになる

3.3.4.推定量曲線

Figure3.7.(右図)は各手法で予測誤差の推定量曲線を示している

多くの曲線について、最小に近くなる誤差を超えると非常に平坦である

図にはクロス・ヴァリデーションによって計算された推定標準偏差のエラーバーも含まれている

”1-標準偏差”ルールという、最も倹約なモデルを最小値の1-標準偏差の中で拾い上げる(7.10.節244ページ)

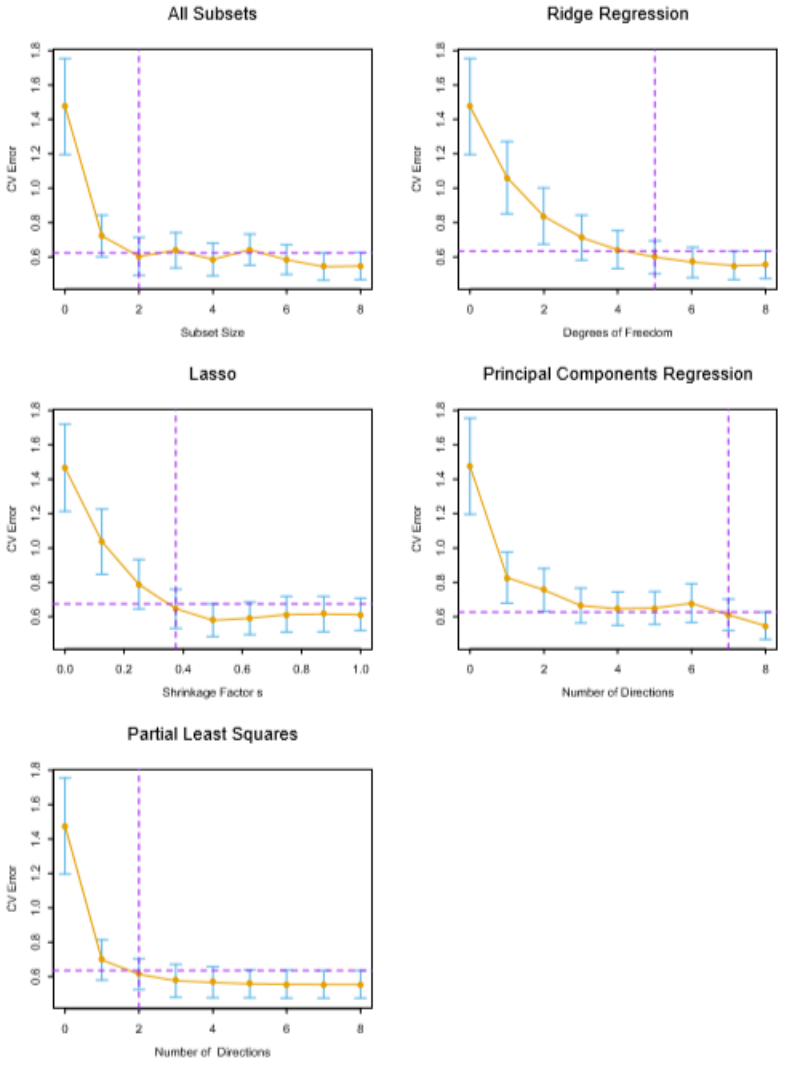


FIGURE 3.7. *Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that method. The horizontal axis has been chosen so that the model complexity increases as we move from left to right. The estimates of prediction error and their standard errors were obtained by tenfold cross-validation; full details are given in Section 7.10. The least complex model within one standard error of the best is chosen, indicated by the purple vertical broken lines.*

3.3.4. 例でのBest subset

Best-subset slctは予測変数lcavolとlweightを選択

表の最後の2行は予測誤差の平均(と推定標準偏差)を、試験集合を越えて与える

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

目次

(・自己紹介)

・はじめに　～　最小二乗法では不満

・3.3.1.　～　Best subset selectionについて

・3.3.2.　～　個の弱点とStepwise selection

・3.3.3.　～　Forward stagewise selection

・3.3.4.　～　前立腺がんの例について

・演習　～　Fwd stpwとBwd stpwについて

演習　雑感

演習　3.9.

Forward stepwise selection

- 切片から初めて以降RSSが最小になるように予測変数をモデルに加えていく
- 貪欲法により組合せ最適化を近似的に解く
- フィットした変数にQR分解を用いることで、次に加える変数を素早く計算（演習3.9.）

演習　3.9.

考えられるところまで考えた結果

→別資料（大意資料）

- 演習3.9.
  - QR分解でQが直交化されたベクトルたちを並べたもの
  - Qの直交系(直交q枠)と直交するように、ベクトルをX_2の列ベクトルからとってきて作る（シュミットの直交化）？
  - 最小2乗法の幾何的考察から直交するものがRSSを小さくする？

演習　3.10.

Backward stepwise selection

- 全変数を考慮したモデル(full model)から始める
- フィットへの影響が少ない予測変数から削る
- Z-scoreが小さい(係数0でないとは言い難い)変数から削っていく(演習3.10.)
- N>pのときでなければ使えない

$$\text{Z-score} \qquad z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \qquad (3.12)$$

$$v_j = \left(X^T X\right)^{-1}(j,j)$$

演習　3.10.

考えられるところまで考えた結果

→別資料（大意資料）

- 演習3.10.
  - 本文通りにZ-scoreが最小なものを除去するのであればTable3.2.より、lcpという予測変数を除去すれば良い
  - Z-scoreが最小なものを除去すればRSSの除去時の増分が最小になることの証明ができない
    - 増分が係数の増加関数であれば最小の係数（Z-score最小に対応？）が増分を最小化する？
    - →微分したが増加関数か判定できず
    - Z-score最小に対応←絶対値大な負数かもしれないので偽？

まとめ

最小2乗法では予測精度と結果解釈の点から不満が残る

→変数の部分集合を選択して精度向上・大局のみ見る

- 組合せ最適化を総当りでやってみる（best subset)
  - 時間がかかる(高精度かつp≦40では速いがp>40で計算不可)
- 組合せ最適化を貪欲法ライクにやってみる(stepwise)
  - pが大きくても良く、かつ総当りと遜色ない精度(Figure3.6.)
- 上のstepwise手法を更に束縛してみた(Fwd stagewise)
  - 予測誤差最小化が遅い
  - 一方、高次元問題で分散を小さくする