ブートストラップ回帰による平滑化

名田穂乃花　(指導教員：吉田 裕亮)

1　はじめに

回帰分析とは, いくつかの説明変数からなるモデルより目的変数の振る舞いを推定し予測を行う手法である. 一般に多くの説明変数を用いると与えられたデータへのフィティング精度は改善される. しかしこれにより与えられたデータへのオーバフィティングが発生し, 予測の精度が落ちる場合が多々ある. これは過学習と呼ばれる.

線形回帰モデルにおいてはブートストラップ回帰を用いることで, 回帰係数の分布を推定し, 真に有効な説明変数を選択して平滑化することが可能である.

本研究では, 周期的なデータに対する離散フーリエ変換を線形回帰モデルとみなし, ブートストラップ回帰を援用することにより, 最適な周期成分を抽出することを試みた.

2　線形回帰モデル

m 個の説明変数 $\{X_i\}_{i=1}^{m}$ の線形和で線形回帰モデルは, 以下の式で与えられる.

$$Y = b_1X_1 + b_2X_2 + \cdots + b_nX_m + \boldsymbol{\varepsilon}$$

ここで Y は目的変数, また ϵ は誤差項でありモデル式からの確率的な要因による擾乱を表す.

3　変数選択

変数選択とは, 回帰分析や判別分析を行う際に, 複数の説明変数の中から効率的に目的変数を説明するに足る説明変数を何らかの基準に従って選択することであり. 基準を満たす変数がなくなった時点で, 変数選択は終了となる. 変数選択には減少法, 増加法, 減増法, 増減法などの手法がある. 変数選択を用いることで, オーバーフィット問題や, 外れ値による影響を抑えた回帰が可能となる.

4　BootStrap

ブートストラップ法とは, 母集団の性質を推定するための手法であり, リサンプリングによって推定値の信頼性評価を目的としている. 再抽出データから得られたデータのばらつきを評価することで, 元のデータから得られる統計量の性質を導く.

その手続きは以下の通りである.

1. ある母集団からの大きさ n の標本 $\{x_1, x_2, x_3, \ldots, x_n\}$ について, 重複を許し, リサンプリングを行う.
2. このリサンプリング標本を $\{x_1^*, x_2^*, x_3^*, \ldots, x_n^*\}$ とする. それに基づき, 推定量を計算し, その値を $T_{(1)}$ とする.
3. 同様の手順を N 回繰り返し, N 個の推定値, $T_{(1)}$,$T_{(2)}$,$T_{(3)}$, ... ,$T_{(N)}$ を得る.
4. N 個の推定値から, 推定量の標準誤差は, 以下のように推定される.

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(T_{(i)} - \overline{T}\right)^2}, \quad \overline{T} = \frac{1}{N}\sum_{i=1}^{N}T_{(i)}$$

以上ことから, 誤差評価や区間推定が可能となる.

5　ブートストラップ回帰

回帰モデルにおける残差からリサンプリングすることで, 回帰係数の区間推定を行うブートストラップ手法をブートストラップ回帰という. 本研究では, 回帰係数の推定値の信頼区間内に 0 が含まれる場合, その係数は取り除くことが可能であると判断することにより変数選択を行う.

ブートストラップ回帰の手続きは以下の通りである.

1. 与えられた n 点のデータから, 通常の最小 2 乗法によって回帰係数 $b_1, b_2, \ldots, b_m$ を推定し, その値を $\widehat{b}_1, \widehat{b}_2, \widehat{b}_3, \ldots, \widehat{b}_m$, とする. その回帰モデルを $Y = \widehat{b}_1X_1 + \widehat{b}_2X_2 + \cdots + \widehat{b}_mX_m + \boldsymbol{\varepsilon}$ とする.
2. それぞれの観測点 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ の値に対して, Y の予測値とその観測値の誤差を残差という. その標本を $\{\epsilon_1, \epsilon_2, \epsilon_3, ..., \epsilon_n\}$ とする.
3. 説明変数側の値は固定したまま, 残差標本から n 個をリサンプリングし $\{\epsilon_1^*, \epsilon_2^*, \epsilon_3^*, ..., \epsilon_n^*\}$ とする.
4. 新しい z_i $(i = 1, \ldots n)$ の値を, 以下の式
$$z_i = (\widehat{b}_1, \widehat{b}_2, \ldots, \widehat{b}_m) \cdot \boldsymbol{x}_i + \epsilon_n^*$$
で与え, このデータから同様にして最小 2 乗法により新しい回帰係数 $(\widehat{b}_1, \ldots, \widehat{b}_m)$ を推定する.
5. この試行を N 回繰り返すとすると, N 組の回帰係数 $(\widehat{b}_1, ..., \widehat{b}_m)$ が得られる. これらから $b_1, b_2, ..., b_n$ の区間推定を行うことが可能となる.
回帰係数の推定値の信頼区間内に 0 が含まれる場合, その係数は取り除くことが可能であると判断できる.

6　離散フーリエ変換

n 等区間点で与えられた時系列データ $\{y_1, y_2, \ldots, y_n\}$ を三角関数の重ね合わせで表現する手法として離散フーリエ変換がある. 離散フーリエ変換では三角関数を周期的な入力データ, 説明変数とみなした線形回帰モデルとして考えられる. すなわち

$$y_j = a_0 + \sum_{k=1}^{n}\left(a_k \cos\left(\frac{2\pi kj}{n}\right) + b_k \sin\left(\frac{2\pi kj}{n}\right)\right) + \epsilon_j$$

の形式の展開を考える．すなわち，目的変数ベクトル $Y = (y_1, y_2, \ldots, y_n)$ に対して説明変数ベクトルが

$$C_k = \left(\cos\left(\frac{2\pi kj}{n}\right)\right)_{j=1}^{n} \quad k \text{ 周期 cos 成分}$$
$$S_k = \left(\sin\left(\frac{2\pi kj}{n}\right)\right)_{j=1}^{n} \quad k \text{ 周期 sin 成分}$$

で与えられたことになり，線形回帰モデルでは

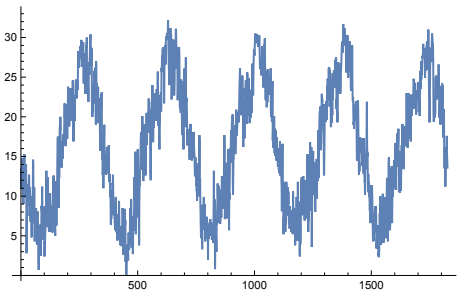
$$Y = a_0\mathbf{1} + \sum_{k=1}^{n}\left(a_k C_k + b_k S_k\right) + \boldsymbol{\varepsilon}$$

と表現される．このモデルにおいて $a_k^2 + b_k^2$ は k-周期成分のパワースペクトルと呼ばれ，この値が大きければ，与えられた時系列データには k-周期成分が多く内在していることを示している．

7　実データへの応用，三角関数の重ね合わせ

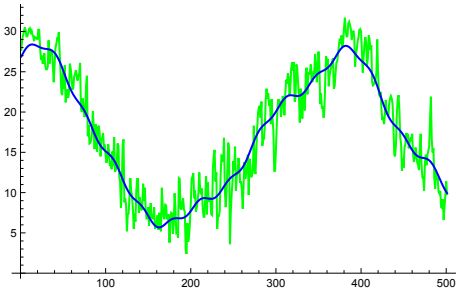
東京における 5 年分 ($n = 1826$ 日)(2021 年 11 月 30 日まで) の日平均気温データを用いる．

元のデータ

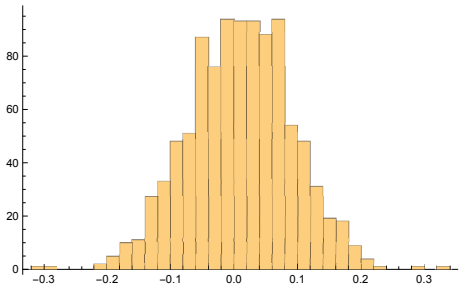


1. 離散データに対して離散フーリエ変換を行い，パワースペクトルを求める．
2. ある閾値でパワースペクトルをカットし，それより大きなパワースペクトルの周波数を残して大まかに平滑化する．
   ここでは，パワースペクトルの閾値を 65 に設定．なお，パワースペクトルは左右対称であるため、1 ～$\frac{n}{2}$ を見れば十分である．ここで残った周波数成分は，$C_0, C_4, S_4, C_5, S_5, C_{10}, S_{10}, C_{13}, S_{13}, C_{15}, S_{15}, C_{17}, S_{17}, C_{19}, S_{19}, C_{20}, S_{20}, C_{54}, S_{54}$ である．

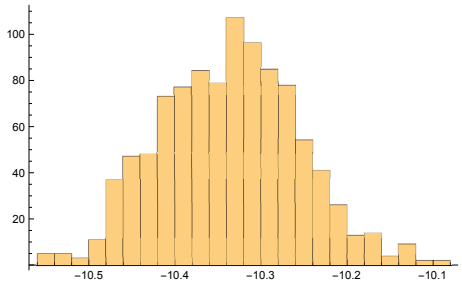
閾値を用いて平滑化したデータの一部
($n = 1000$～1500)



3. それを回帰モデルとしてブートストラップ回帰を行う．
4. 変数選択を行う．係数の区間推定として 0 を含むかどうかを基準に行う．
   例えば，係数の推定値のヒストグラムで見た場合に，以下のような場合，0 を含むと判断する．

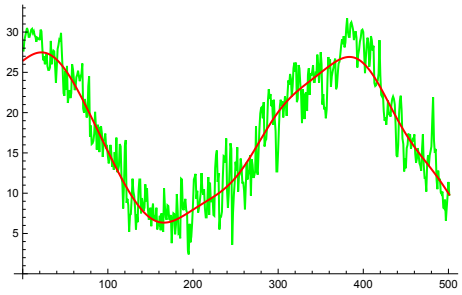


次の場合は当てはまらない．



この結果，$C_0, C_4, S_4, C_5, S_5, S_{10}, C_{13}, C_{15}, S_{15}, C_{17}, C_{19}, S_{19}, S_{20}, S_{54}$ が有効な周波数成分であると判断された．

変数選択後の平滑化気温データの一部
($n = 1000$～1500)



8　まとめ

離散フーリエ変換を線形回帰モデルとみなして，ブートストラップ回帰を行うことで，最適な周波数成分を選択することができた．

参考文献

[1] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York, NY (2017).