状態空間モデルの考え方・使い方

TokyoR #38

2014/4/19

@horihorio

自己紹介

- Twitter ID: @horihorio
- お仕事: 分析コンサルタント
- 興味: 統計色々/DB/R/Finance/金融業/会計
- 過去の発表: ここ
- 最近の出来事
  - 金融業以外の分析にも進出
  - 主に週末は子ども(そろそろ2歳)の相手中
  - 4月から何故か転職した

紹介しないこと

1. Rの色々な操作
2. 線型回帰の説明 ⇒ 前提知識とします
3. 状態空間モデルの色々な理論

！　別の本や資料を見て下さい　！

時系列モデルって

対象：時間経過に伴い変わるもの(アクセス数、株価…)
　個の変化を数式表現したもの

例えば：(最初に学ぶ)ARIMAモデルで
Web会員登録者数を予測

$$y_t = 0.86y_{t-1} + 0.34y_{t-2} + \varepsilon_{t-1} + \cdots$$

入力：過去の値
出力：将来の値

で、なに？

状態空間モデルの完成イメージ

同じく Web会員登録者数を予測

$$\begin{cases} y_t = H_t x_t + w_t & \text{（観測方程式）} \\ x_t = F_t x_{t-1} + G_t v_t & \text{（状態方程式）} \end{cases}$$

- y_t：Web会員登録者数
- x_t：SEO, Listing, TVCMコスト、キャンペーン、etc…
  ＋ 見えない状態（後述） などのベクトル

（参考 v_t, w_t：ノイズ, F_t, G_t, H_t：係数）

入力：色々な要因
出力：将来の値

構造が見える

次何する？を考える

今回のメッセージ

状態空間モデルは、
線型モデルの拡張！

これを、何回も繰り返すだけ

アジェンダ

1	線型回帰の拡張とは？
	• 个を考えることで、状態空間モデルの発想に慣れてください
2	状態空間モデルの考え方
	• 状態空間モデルの得意な面、不得意な面の両方を知ってください
3	状態空間モデルの使い方
	• 得意/不得意を踏まえ、ではどう使うべきか？を考えていきましょう

1. 線型回帰の拡張とは？

1	線型回帰の拡張とは？
	• 1個を考えることで、状態空間モデルの発想に慣れてください
2	状態空間モデルの考え方
	• 状態空間モデルの得意な面、不得意な面の両方を知ってください
3	状態空間モデルの使い方
	• 得意/不得意を踏まえ、ではどう使うべきか？を考えていきましょう

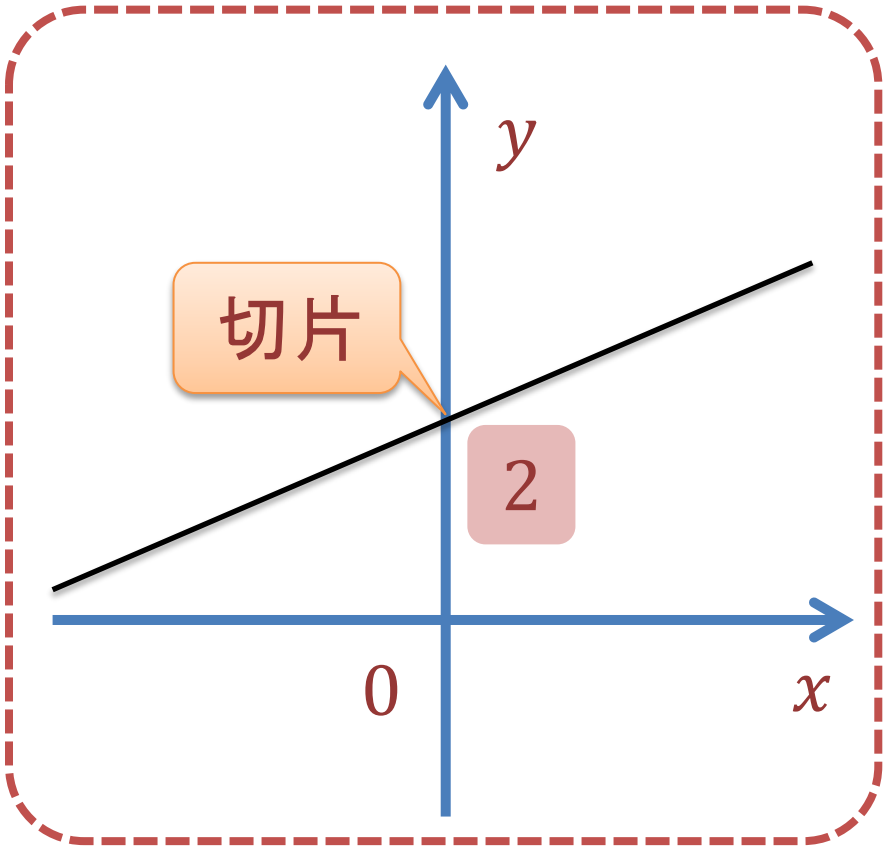
【中学の復習】切片とは？

例: $y = 2 + x/2$

$x = 0$ のとき $y = 2$ 。これを切片という。

言いかえ:

- 説明変数と
  無関係のゲタ
- 現状の持ち点



状態空間モデルの式

$$\begin{cases} y_t = H_t x_t + w_t & \text{（観測方程式）} \\ x_t = F_t x_{t-1} + G_t v_t & \text{（状態方程式）} \end{cases}$$

両者を
同時推定

- x_t：状態
- v_t：状態ノイズ（平均ゼロの正規分布）
- y_t：観測値
- w_t：観測ノイズ（平均ゼロの正規分布）
- F_t, G_t, H_t：それぞれ $k \times k, k \times m, l \times k$ の行列

これをいきなり理解する、って難しい。。。

状態空間モデルの式

話を簡単に。状態を一定とすると…

$$y_t = H_t x_t + w_t \quad \text{（観測方程式）}$$

$$x_t = F_t x_{t-1} + G_t v_t \quad \text{（状態方程式）}$$

単なる
重回帰

- x_t：状態
  x_t：説明変数
- v_t：状態ノイズ（平均ゼロの正規分布）
- y_t：観測値
- w_t：観測ノイズ（平均ゼロの正規分布）

本報告で言いたい
たった一つのこと

では状態方程式って

$$y_t = H_t x_t + w_t \quad \text{（観測方程式）}$$

$$x_t = F_t x_{t-1} + G_t v_t \quad \text{（状態方程式）}$$

- x_t：状態
- v_t：状態ノイズ（平均ゼロの正規分布）
- y_t：観測値
- w_t：観測ノイズ（平均ゼロの正規分布）

$$x_t = \begin{pmatrix} SEO_t \\ Listing_t \\ TVCM_t \\ CampaignFLG_t \\ \cdots \\ 状態_t \end{pmatrix}$$

説明変数

切片

状態方程式って

まとめると:

説明変数 ＋ 時間とともに動く切片

⇒「ゲタ」が都合よく動くので、柔軟性が高い

- なことは、以下Blogに書いている
  http://logics-of-blue.com/ローカルレベルモデル/

その他状態空間モデルのご利益

- 状態方程式、観測方程式の両方を同時推定している
  - モデル作成→残差を目的変数とし更にモデル、ではない

- データの増加や、昔から現在へと進むに連れて、勝手に係数が更新される
  ⇒(平たく言えば)勝手に賢くなる
  - 正確に言えば:予測分布、フィルタ分布、平準化分布を順次推定＆更新する
  - 弾道ミサイルの計算に採用された理由かも？

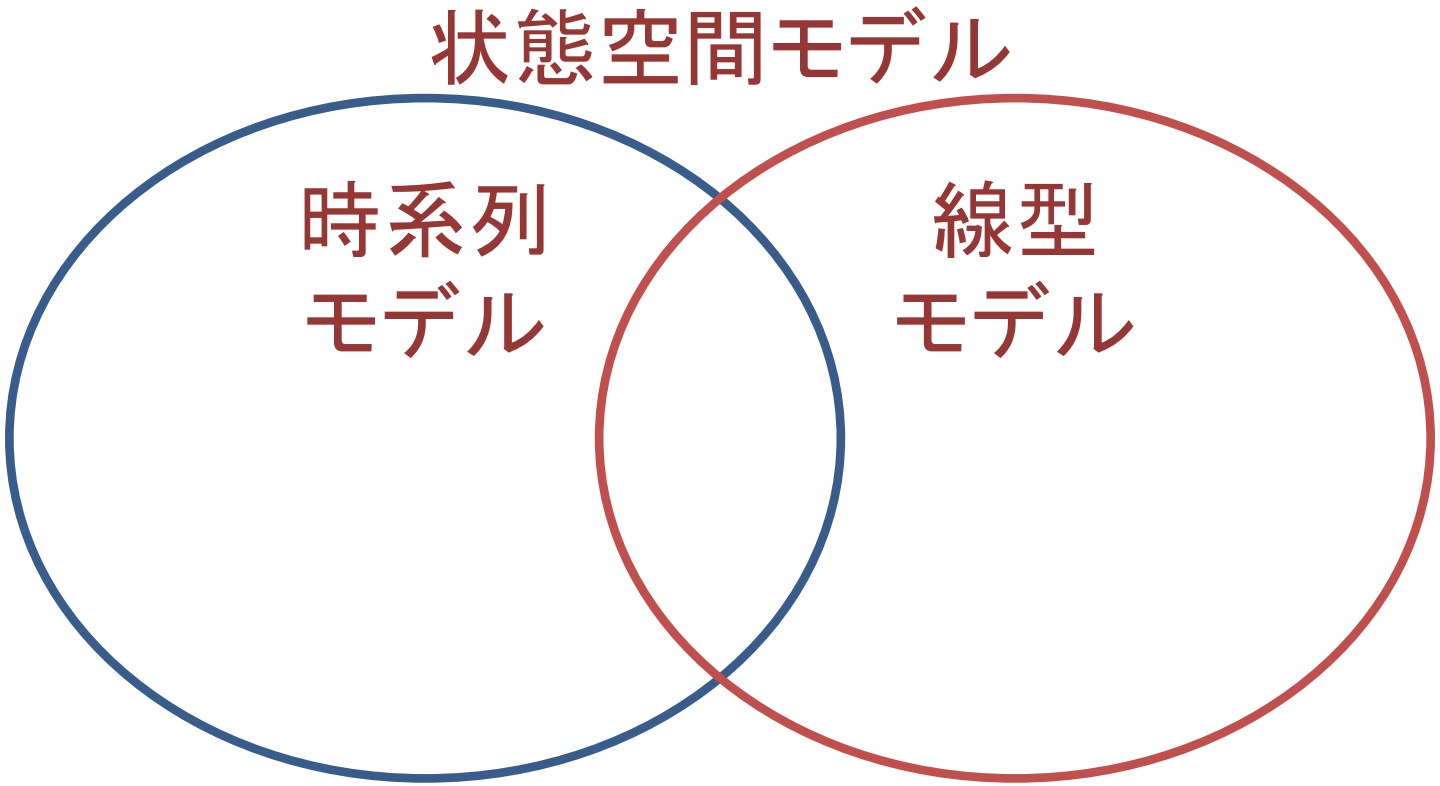
等々。詳しくは参考文献を。

2. 状態空間モデルの考え方

1	線型回帰の拡張とは？ • 1個を考えることで、状態空間モデルの発想に慣れてください
2	状態空間モデルの考え方 • 状態空間モデルの得意な面、不得意な面の両方を知ってください
3	状態空間モデルの使い方 • 得意/不得意を踏まえ、ではどう使うべきか？を考えていきましょう

時系列モデル＆線型回帰モデル

状態空間モデル＝時系列モデル＋線型モデル
使い方も足し算？なお話デス



時系列モデルとは？

発想： 過去の自分から、将来の自分を当てたい

ARIMAモデルの場合

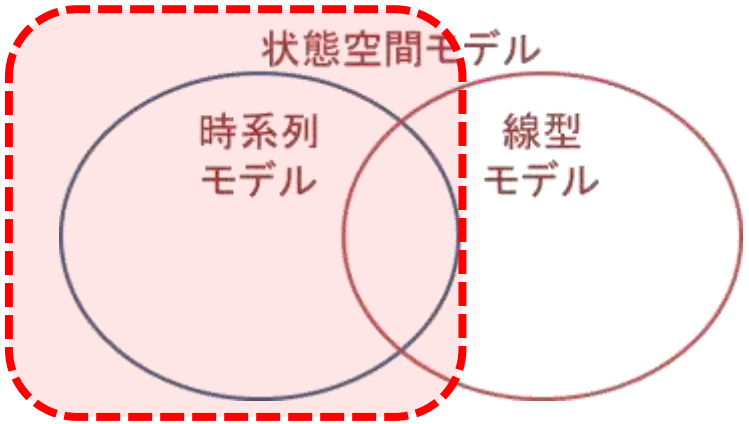
$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + e_t + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t$$

AR（自己回帰）

I（和分）

MA（移動平均）

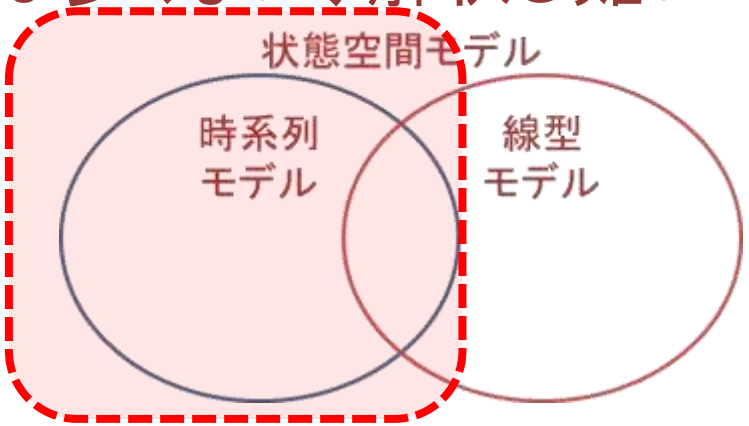
誤差項



時系列モデルとは？

（線型モデルと比較しての）メリット/デメリット例：

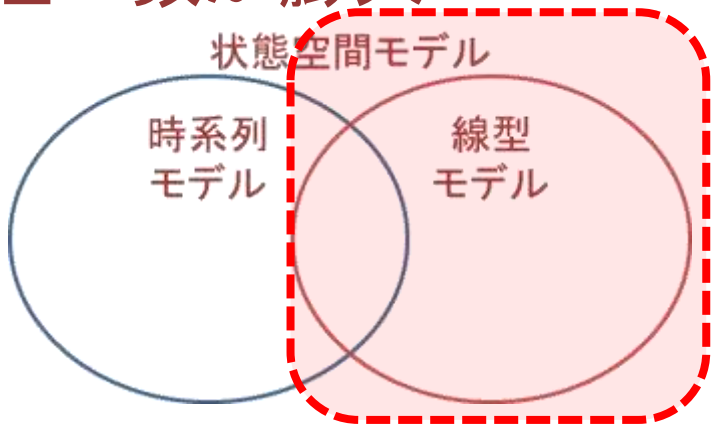
- メリット
  - 理論面：　観測値どおしの相関が導入されている
  - 実務面：　データ収集がそこそこ簡単
- デメリット
  - 理論面：　誤差が累積するので、長期予測はダメ
  - 実務面：　そんなに観測点は多くない、解釈し難い



線型モデルとは？

略w　（時系列モデルと比較しての）メリット/デメリット例：

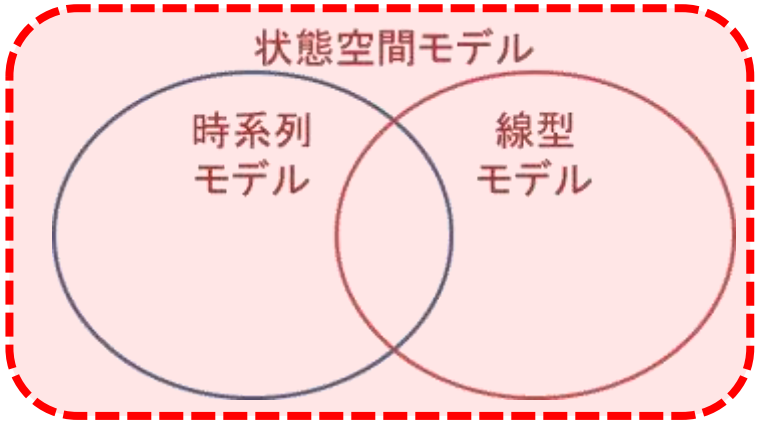
- メリット
  - 理論面：（基本モデルなので）道具や対応法が豊富
  - 実務面：データ数大、結果解釈が容易な場合も
- デメリット
  - 理論面：「変数間が独立」の仮定が厳しい…
  - 実務面：データ収集、前処理工数が膨大



では状態空間モデルとは？

理論面を話しても面白く無いので、実務面では:

- メリット
  - 時系列の構造が可視化できる
  - 状態成分で、明示的に周期性・トレンドが導入可能
- デメリット
  - よく部分最適解に陥り、変な係数が出てくる
  - 計算コスト(時間・工数)が高い
    ⇒モデル更新を
    ためらってしまいがち



3. 状態空間モデルの使い方

1	線型回帰の拡張とは？
	• 个を考えることで、状態空間モデルの発想に慣れてください
2	状態空間モデルの考え方
	• 状態空間モデルの得意な面、不得意な面の両方を知ってください
3	状態空間モデルの使い方
	• 得意/不得意を踏まえ、ではどう使うべきか？を考えていきましょう

最初におわび

本当はやりたかったこと:パッケージdlmを使った、TOPIXとCI先行指数の各成分との分析例

要は、一番基礎的なカルマンフィルタですが、

時間と体力＆気力的に無理でした…。

おわびに:

実務で使ってみての個人的FAQを列挙

でお茶を濁しマス

ケース1

効果として、これも、あれも、…考えられるけど、モデルに是非とも盛り込めない？

線型回帰：

結構対応可能

（ただし、IT/数理両面での前処理技術に依存 ）

状態空間：

結構無理

観測点が少ないため、変数数が少なくなりがち

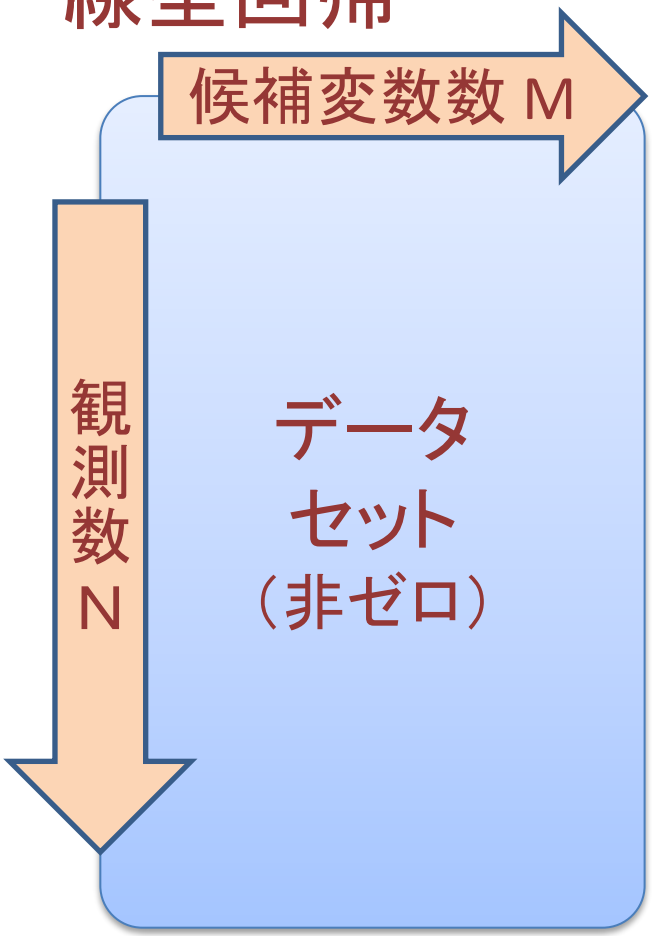
（次ページ参照）

⇒ ヒヤリング＆コミュニケーション時の留意事項

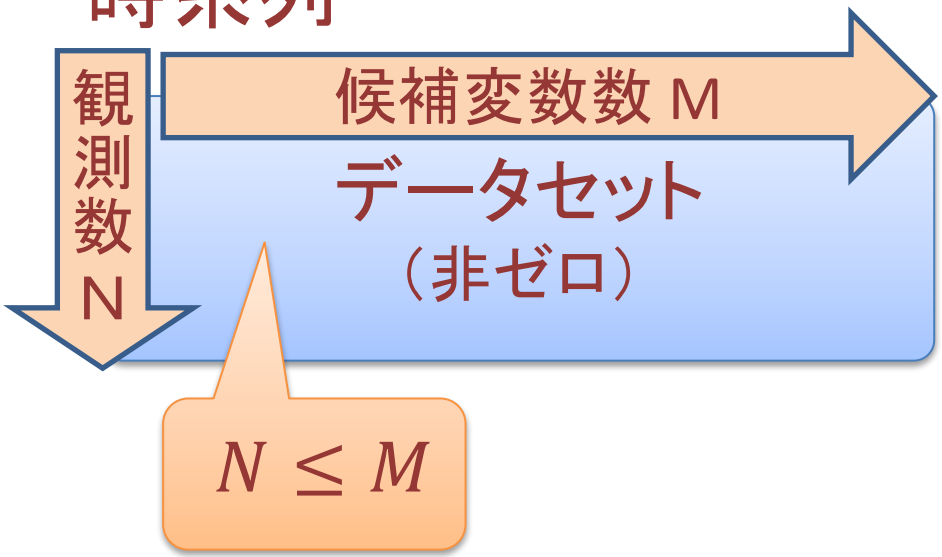
ケース1

「変数数が少なくなりがち」な理由

線型回帰



時系列



3年間の月次データならば、
方程式の最大変数数は36
ですよね…

ケース2

漠然と「状態」で説明されても困る。
年間、週次での周期成分は明示できないか？

明示するには：　説明変数に導入すれば良い。

（FLGを入れる、sin/cos成分を入れる、等々）

ただ問題点：

説明変数数はそう増やせない

（理由は、ケース1のとおり）

⇒対象の周期成分が強いならば、ARIMAモデルの方がむしろ適当といえる。

ケース3

VARモデルも、他の変数効果が導入できる。
状態空間モデルとの違いは？

まずVARモデルとは？

⇒ ARモデルの多変数版　$X_t = \sum_{i=1}^{p} A_i X_{t-i}$

違いは、係数の決定時点

- VAR:　推定時点＆期間でのスナップショット
- 状態空間:　逐次更新される

将来予測の際、状態空間では「将来の状態」についての仮定が導入可能。VARは何も出来ない。

ケース4

説明変数に自己回帰($y_{t-1}, y_{t-2}, \cdots$)を入れて良いか？

私がむしろ聞きたい(わりとマジで)

今のところ… 入れないことにしている

- 理論面： 単位根（次ページ参照）が怖い
- 実務面： 説明変数を解釈する際、循環参照になる
  ⇒ 実質、解釈できない

ケース4(補足)

そもそも単位根とは一体？

詳細は、TJOさんHPや、右の資料とか

端的にいえば:

$$y_t = \beta + \alpha y_{t-1}$$

を仮定して係数を推定する際、もし α が1に近いと(ランダム・ウォーク)、t 統計量が発散(p値→0)する。

⇒ ランダム・ウォークなので、将来予測が全く当たらない、「高度に有意」なモデルとなってしまうorz

ケース5

係数推定の際、尤度が部分最適解にならないか？

- はい。大概なりますw
- 実務的には
  - 部分最適解でも、大幅な業務効率改善になる
  - 同じ労力ならば、必死にGlobal Optimal求めるより、お客様と、運用や活用法を相談する方が報われることの方が多い、との実感

ケース6

その他何かあります？

- ネタ切れなので、会場に丸投げw
- 質問は、資料UPDATEの際に反映させる、かも知れません

まとめ

ビジネスでのデータ分析って

ビジネスの場合
- 分析して儲かる
- 相手が納得する

が大前提。

1. ビジネスの問題を
   A) 数学の問題に**翻訳**し
   B) 問題を数学的に解いて
2. 再度ビジネス世界に**翻訳**

で
- **「翻訳」**の妥当性
- 「1.から2.」&「A)からB)」の両方が妥当か

の両方を詰めるのは大変、だけど面白いです。

参考文献

- J. D. Hamilton, “Time Series Analysis,” Princeton Univ. Press, 1994

  和訳：沖本・井上(上下巻)、シーエーピー出版、2006年　は絶版

- 北川源四朗 『時系列解析入門』 岩波書店、2005年
  →中級者向け理論書。状態空間モデルベースの解説

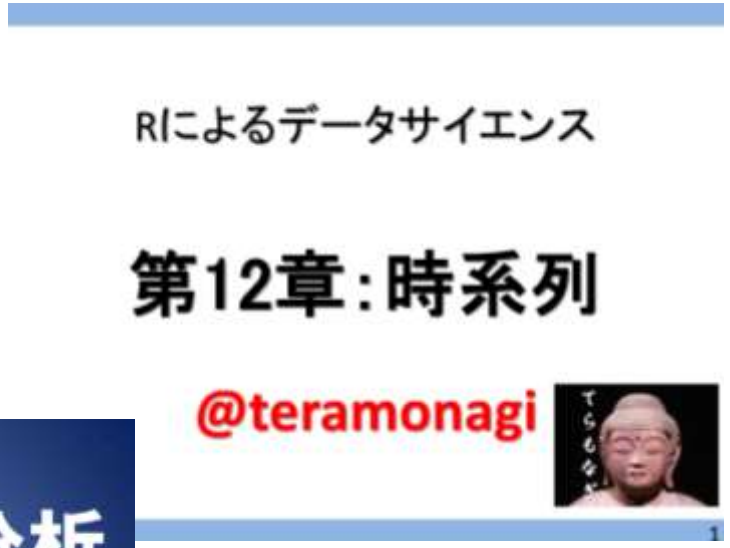
- J.J.F.コマンダー、S.J.クープマン、和合肇（訳）『状態空間時系列分析入門』 シーエーピー出版、2008年
  →具体的にデータを分析している過程もあり

Appendix

時系列モデルの解説

過去の Tokyo.R. で触れられた資料

http://lab.sakaue.info/wiki.cgi/JapanR2010?page=%CA%D9%B6%AF%B2%F1%C8%AF%C9%BD%C6%E2%CD%C6%B0%EC%CD%F7#p15



ARIMAでないが
関連して

時系列モデルの解説

使い方に側面を当てると…

http://www.slideshare.net/horihorio/howtousetimeseries

とりあえず
自己営業w

(一般化)線型モデルでの仮定

数点キツイ仮定を置いていることに留意

1. 各観測値は独立
   ⇒ 時系列データではありえない
   「状態空間は重回帰の拡張」は、この点で若干ウソ
2. 残差の分布が分かっている
3. 分散の構造が分かっている
4. 説明変数に何か変換(リンク関数)をし、線型結合することで、応答変数が表現できる

引用源:R-bloggersより(→これオススメ!)

http://www.r-bloggers.com/checking-glm-model-assumptions-in-r/