これまでの
プログラミング経験について

福田　太一

自己紹介

生年月日	**1996年8月27日愛知県生まれ**
所属	名古屋工業大学大学院・博士課程前期1年 舟橋研究室
趣味	料理, サイクリング, TVゲーム, ストレッチ
メールアドレス	f-taitai@i.softbank.jp, guhaabeshi@gmail.com
一言	データサイエンスの知識を深めるべく日々邁進中です. (現在はPythonに力をいれています.)

習得したプログラミング言語

言語	経験年数	習熟度 (5段階)	経験の具体例
R	2年間		• ロジスティック回帰 • 主成分分析 • 数理モデルの実装 • 時系列分析 (初歩)
Python	6ヶ月		• 主成分分析 • xgboostを用いたGBDT 回帰モデルの実装

これまでの
プログラミングの成果物

1. ロジスティック回帰を用いた要注意学生の推定
   (学部4年次の研究テーマ)

2. COVID-19の流行を予測するモデルの作成
   (博士課程前期の研究テーマ)

1.ロジスティック回帰を用いた要注意学生の推定

問題提起と実践した内容

何が起きているのか

教育現場での要注意学生
(将来留年・退学する学生)
[1] の存在

教員の負担が増える

指導量

時間

解決のために何をするのか

学生のデータを
データマイニング

要注意学生の
早期の抽出

教員の負担の軽減

[1]伊藤圭佑:“データマイニングによる『要注意学生』の発見に関する研究”、
平成 25 年度名古屋工業大学修士 論文、2013.

プログラムに用いたデータ (対象: 4年次学生110人)

目的変数 (分類したいデータ)

留年判定 :学生の留年判定を記録したデータ

(0: 留年している, 1: 留年していない)

説明変数 (分類のために必要なデータ)

GPA: 学生の成績の指標となるデータ

学生生活実態調査: 学生生活に関するアンケート調査の結果

学生生活実態調査のデータ

学生生活実態調査の結果からは
次のものが成績に影響すると仮説を立て, 採用した.

- 睡眠データ
  平日就寝時間, 平日起床時間, 平日睡眠時間
  休日就寝時間, 休日起床時間, 休日睡眠時間

- 住居・通学データ
  出身校所在地, 住所, 通学手段, 入構手段, 同居人

データマイニングの手法

これらの手法をプログラミング言語Rで実装.

ソースコード: bond2580/lo (github.com)

・ロジスティック回帰モデルL

$$L = \frac{exp(\alpha + \sum_{i=1}^{n} \beta_i x_i)}{1 + exp(\alpha + \sum_{i=1}^{n} \beta_i x_i)}$$

x: 説明変数　α, β: 定数

・主成分分析

$$z_n = a_{n1}x_1 + a_{n2}x_2 + \ \ldots \ + a_{nn}x_n$$

Z_n: 主成分得点, a: 係数, x: 説明変数

要注意学生の推定実験

1. ロジスティック回帰モデルの出力が閾値を超えた場合に要注意学生と推定.

2. 1.を採用データ, モデルへの入力, 変数選択法を変えて繰り返す.

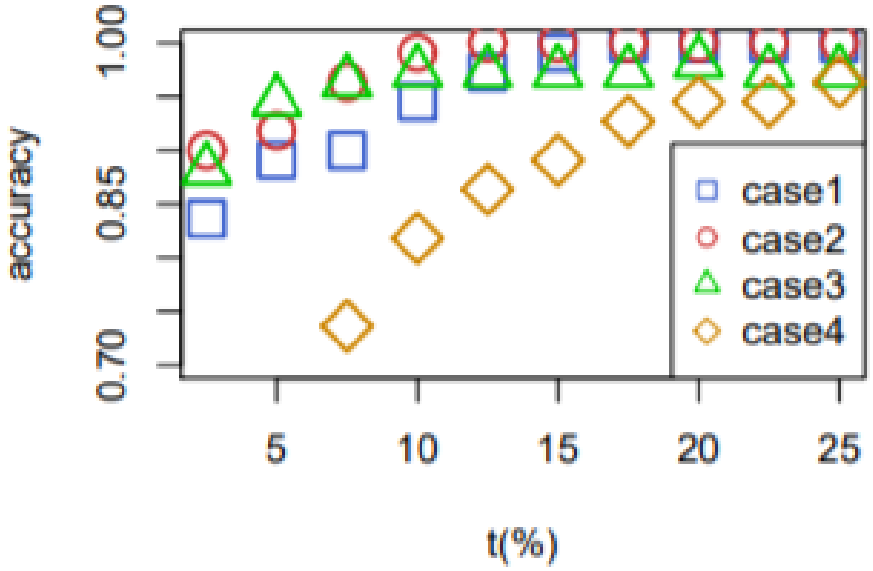
モデルへの入力	• **実データ** • **実データの主成分得点**
変数選択	• **全て投入 (強制投入法)** • **徐々に減らして最良化(ステップワイズ法)**

採用データ別最良モデル (ここではcaseと呼称する)

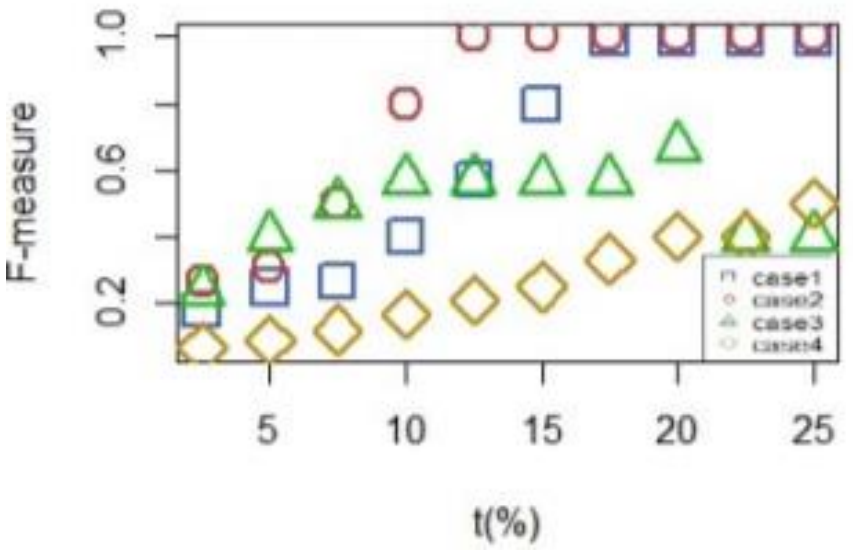
ケース	採用データ	入力	変数選択
case1	GPA, 睡眠 住居・通学	主成分得点	ステップ ワイズ法
case2	GPA, 睡眠	主成分得点	ステップ ワイズ法
case3	GPA 住居・通学	実データ	ステップ ワイズ法
case4	睡眠 住居・通学	実データ	強制 投入法

各ケースの精度の比較 (横軸: 閾値)

モデルの正答率



F値 (モデルの精度の高さ)



case2の高い推定率が確認できた.

各ケースの推定結果のまとめ

- case1がcase2より劣る
  住居・通学データは推定においてノイズ

- case2が全体的に推定率が高い
  - GPAと睡眠データは推定に有効
  - 入力は主成分得点，変数選択はステップワイズ法が有効

- case2がcase3より優れる
  推定においてGPA以外のデータも必要

むすび1

- プログラム概要

  GPAと学生生活実態調査の結果を用いた
  ロジスティック回帰分析による要注意学生の推定

- 結果

  GPAと睡眠データによるモデルの高い推定率を確認

- 課題
  - 過学習の回避
  - 新たな種類のデータの追加
  - GPAと睡眠データの関係性の調査

2. COVID-19の流行を予測するモデルの作成

2. COVID-19の流行を予測するモデルの作成

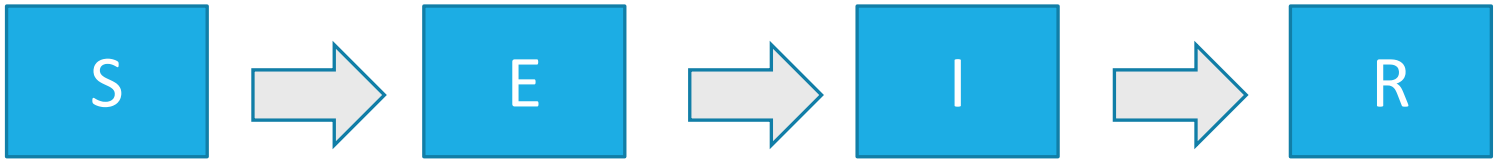
現在進行中の研究「COVID-19の流行に伴う人々の行動パターン分析」の一環として感染症流行予測モデル (SEIRモデル) [2] をRで実装.

ソースコード: bond2580/virus (github.com)

[2]「感染症流行の予測：感染症数理モデルにおける 定量的課題」,西浦 博, 稲葉 寿, 統計数理第54巻第2号, p.p461-480, 2006.

SEIRモデル とは

4つの感染段階の人口をボトムアップで予測



S: 感受性人口 (ウィルスに免疫を持たない人口)

E: 潜伏人口 (ウィルスに感染し, 潜伏している人口)

I: 発症人口 (発症し, ウィルスを人に移しかねない人口)

R: 回復人口 (症状が回復し免疫を獲得した人口)

S + E + I + R = 国内の総人口 (N)

SEIRモデルのパラメータ

R_o:基本再生産数　一人が感受性者にウィルスを移す人数の限界

e:平均潜伏期間　感染してから発症するまでの期間

l:平均発症期間　発症してから回復・死亡するまでの期間

モデルを用いた予測の手順

1. 過去のデータを用いてパラメータチューニング
2. I.の結果から未来のパラメータを導出
3. II. の結果をSEIRモデルに投入して未来の流行を予測

過去のデータを用いて パラメータチューニング

過去のデータ[3]の中でも明確なSとRの実測値と予測値の誤差が最も小さくなるパラメータをグリッドサーチ.

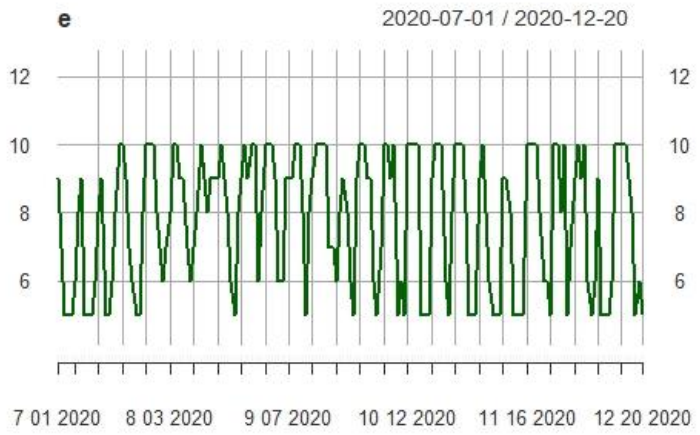
予測開始日: 2020年7月1日～12月20日

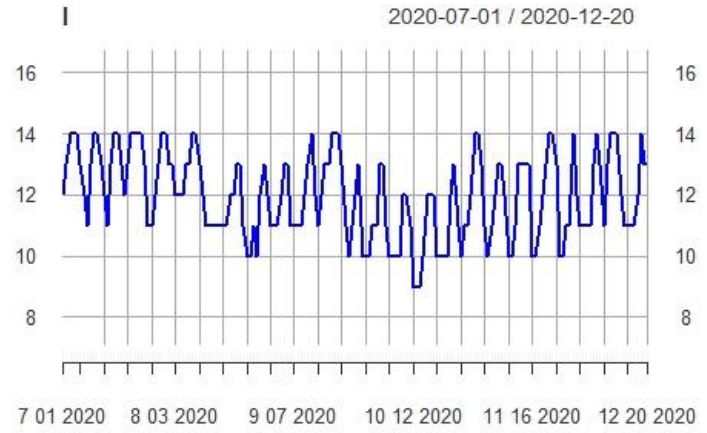
予測範囲:　直近7日間

評価指標:　平均二乗誤差

[3]新型コロナウイルス 国内感染の状況 (toyokeizai.net)

チューニングされた過去の最適パラメータ





パラメータ	区間
R0	0.1 ～ 3.0 (人)
e	5 ～ 10 (日)
I	8 ～ 14 (日)

未来のパラメータを推測

チューニング後のパラメータを時系列分析

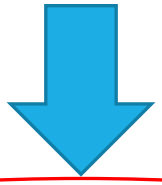
1. パラメータの周期性の確認
   → **周波数スペクトルの確認**

2. 周期性の有無に応じたパラメータ推測モデルの作成
   → **ホルトウィンタース法によって実装**

周波数スペクトルとは

時系列データの持つ周波数に応じた

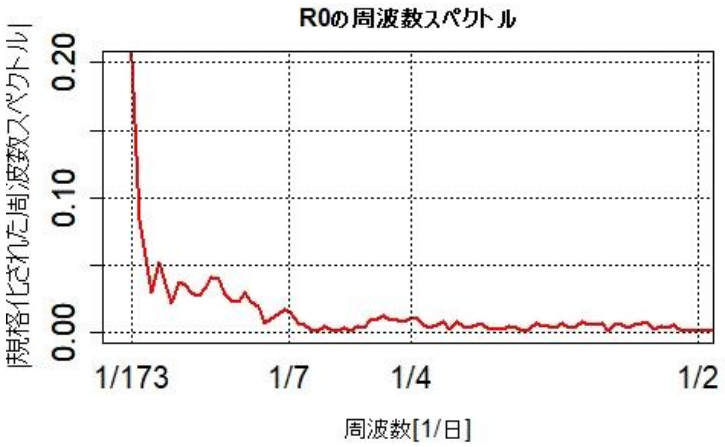
エネルギーの強さ

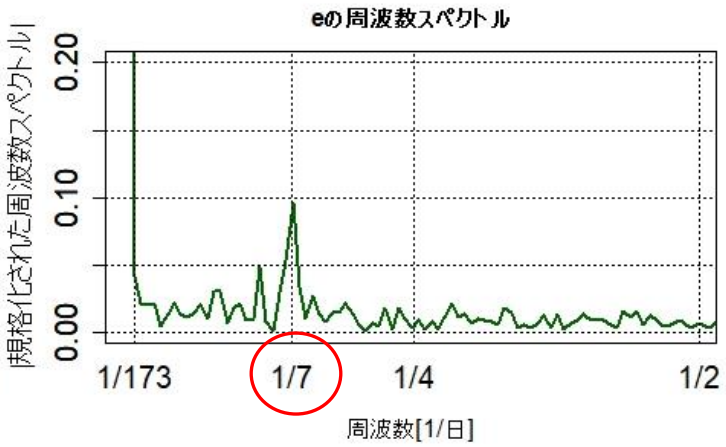


スペクトルの大きくなる周波数から

データの周期が分かる

チューニング後パラメータの周波数スペクトル





R0 ：周期性無し
e, I : **7日間**の周期性を確認

ホルトウィンタース法による モデル作成

時系列データの3つの成分を利用して推測モデルを作成

R0 : トレンド + レベル

e : トレンド + レベル + 周期性

I : トレンド + レベル + 周期性

トレンド : 時系列データの傾き

レベル : 時系列データのランダムに変化する値

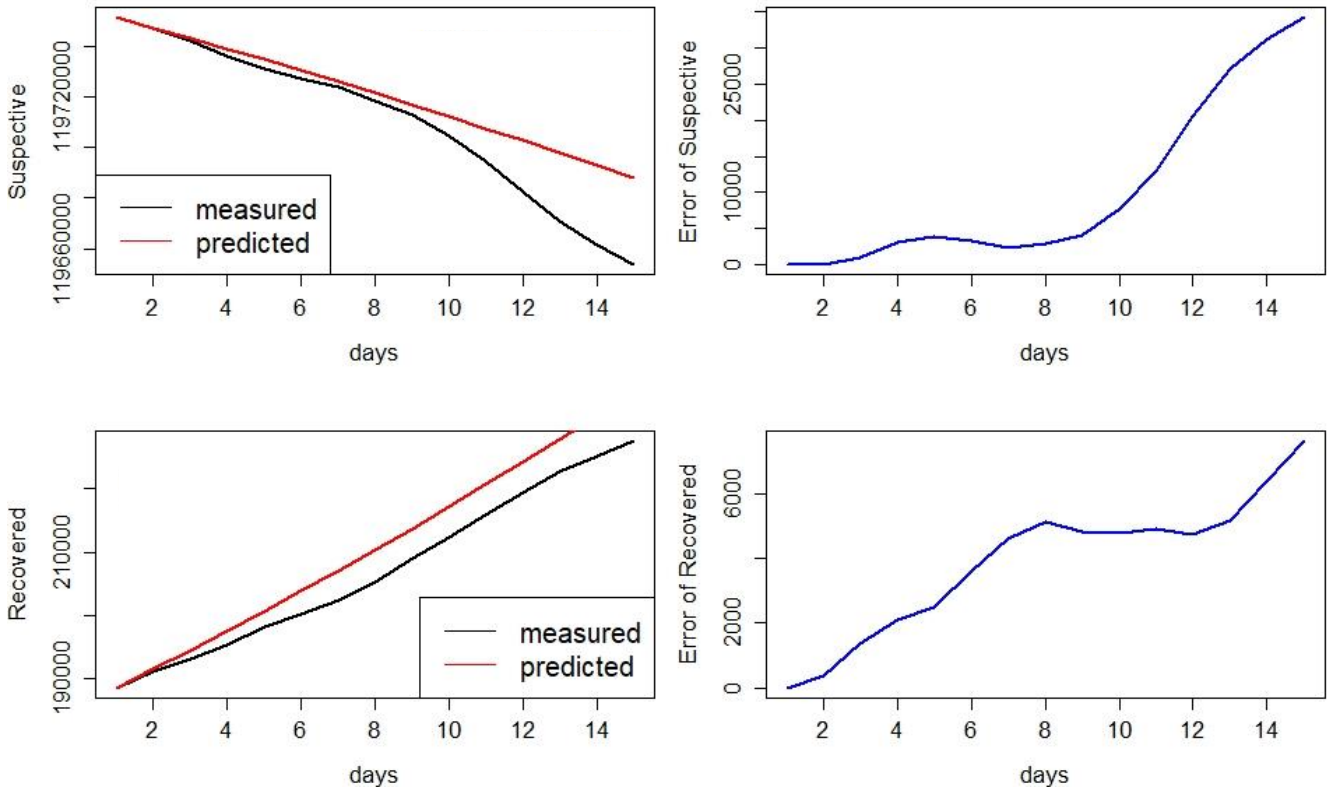
周期性 : 時系列データの周期性

推測された未来のパラメータ

日付	R0	e	I
2020/12/28	1.365	0.873	10.91

推測されたこのパラメータを使ってGotoキャンペーン
一時停止期間中(2020/12/28 ～ 2021/1/11)の流行を予測.

予測結果と実測値との誤差
([R0, e, l] = [1.365, 0.873, 10.91])



実測値との誤差が時が経つにつれ大きくなっていく

むすび2

- プログラミングの概要

  研究の段取りの一つとして

  COVID-19流行を予測するモデルを構築

- 結果

  予測値と実測値との誤差は最初は小さいが徐々に増大

- 課題

  実測値との誤差が小さくなるようにモデルを改良

  モデルの予測値と人々の行動パターンとの関連付け

最後に

最後までご覧いただきありがとうございました。