

# 国内における 新型コロナウイルスの 感染者数予測

—

G2

185719D 上田瑞宜, 185725J 識名俊希,  
185757G 鈴木海土, 185768B 佐藤優

# 目次

1. 目的・目標
2. アプローチ全体像
3. 予定していた実験計画
4. データセットの構築方法
5. 機械学習の進め方
6. 実験
  - 6.1 実験設計
  - 6.2 実験結果
  - 6.3 考察・自己評価
7. 残された課題

# 1.目的・目標

## <目標>

1週間後、1ヶ月後の国内のコロナ感染者数予測する。

## <目的>

現在新型コロナウイルスが流行しているが、新規感染者数の増減を予想することは極めて困難である。そこで、コロナに関係していると思われるデータを用いて、近い将来の新規感染者数予想を試みることによって、傾向と対策に繋げる。

## 2.アプローチ全体像(1)

どのようにしてデータを予測するか？

- ・感染症予測にはどのようなモデルを使用するのが良いか。



SARIMAモデルとSEIRモデルにしよう

## 2. アプローチ全体像(2)

どのようなデータが必要となってくるか？

- ・感染者数
- ・総人口
- ・退院者数
- ・死亡者数
- ・潜伏期間？
- .....

## 2. アプローチ全体像(3)

推定区間、検証区間を決定しよう

→ 6月23~30, 7月1~7

モデルの評価をどうしようか

→ 決定係数

### 3. 予定していた実験計画

- 1 国内でのCOVID-19の症例データ収集
- 2 使用するモデルの選択
- 3 モデルの実装
- 4 パラメータ調整
- 5 データ範囲の調整
- 6 モデルの評価

## 4. データセットの構築方法

日本国内におけるコロナ感染者数、死者数、入院者数などが含まれているデータを探す。

実際に個人でコロナ感染者予測を行っているサイトで使用しているデータや国などが出しているデータを使用。

1日ごとの新規感染者数や総感染者数、入院数、死者数など含まれているデータによって使い分けて使用。

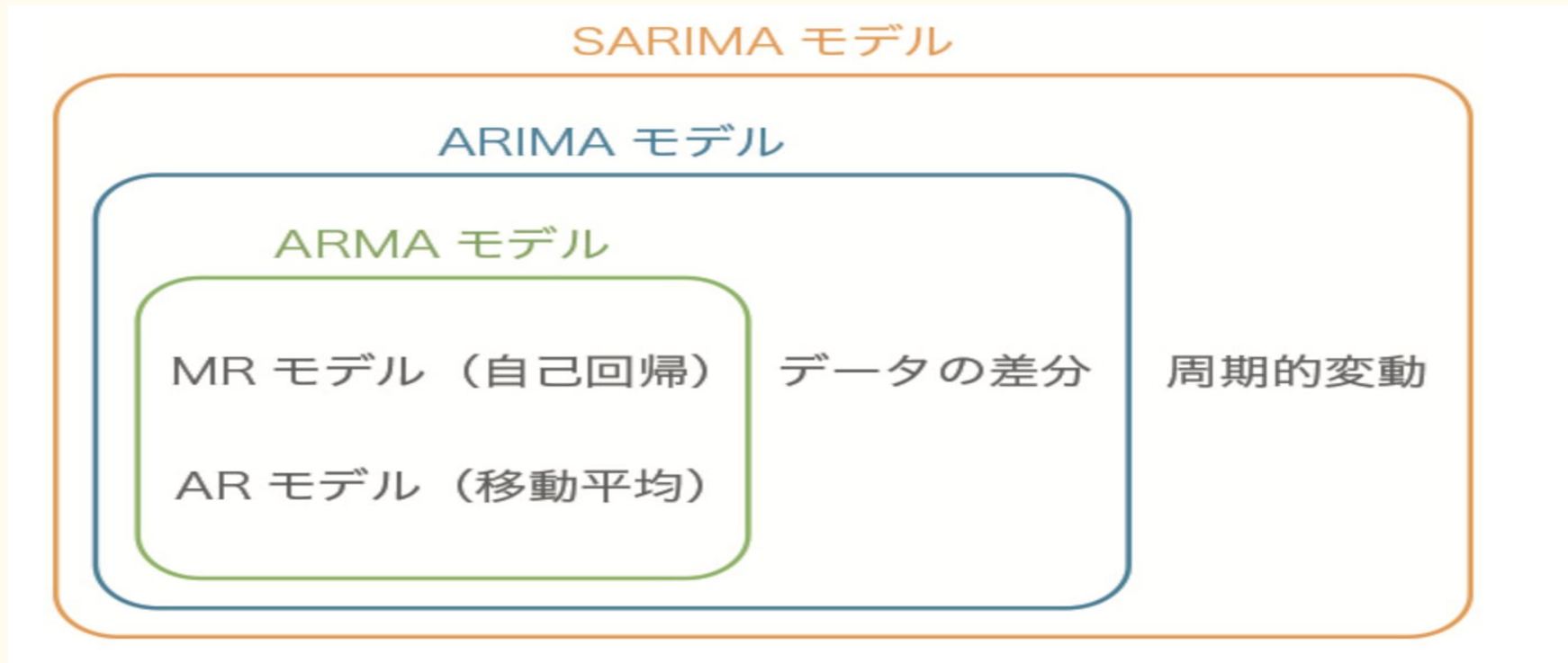
SARIMAモデルの場合は新規感染者数データを使用した。

(人口密度や人の移動率をNTTdocomoが提供しているデータを使用)



# 5.機械学習の進め方 ～SARIMAモデル～

SARIMAとは？



## 5.機械学習の進め方 ～SARIMAモデル～

- 使用した学習器(学習方法)は、...

`statsmodels.api.tsa.SARIMAX()`

- パラメーターはどのようにして調整したのか？

パラメーターの内の`order(p,d,q)`と`seasonal_order(sp,sd,sq)`の値の組み合わせを総当たりで求めその中で最もAIC値の低かったものを採用した。

## 5.機械学習の進め方 ～SEIRモデル～

### ・使用した数理モデル SEIRモデル

$$\begin{aligned}\frac{dS}{dt} &= -\beta \frac{SI}{N} \\ \frac{dE}{dt} &= \beta \frac{SI}{N} - \epsilon E \\ \frac{dI}{dt} &= \epsilon E - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

感染症流行の数理モデル

S:感染しうるもの、免疫を持たず感染可能(Susceptible)

E:感染症が潜伏期間中のもの(Infected)

I:感染症が発症しているもの、接触した感染しうるもの(S) に病  
気を感染(Infected)

R:感染後死亡者、もしくは免疫を獲得した者(Recovered)

N:全人口,  $S+E+I+R$

$\beta$ :感染率(The infectious rate)

$\epsilon$ :暴露後に感染症を得る率(The rate at which an  
exposed person becomes infective)[1/day]

$\gamma$ :除去率(The Recovery rate)[1/day]

## 5.機械学習の進め方 ～SEIRモデル～

### ・パラメーターの調整方法

以下の対数尤度関数の最大値を求めることで、

感染率  $\beta$  ・暴露後に感染症を得る率  $\epsilon$  ・除去率  $\gamma$

を最適化する。

$$\sum (\text{評価関数の計算結果}) - (\text{実データ}) \cdot \log |(\text{評価関数の計算結果})|$$

(実際はマイナスを取って最小化問題として、Scipyのminimize関数を用いた)

## 6.1 実験 ～実験設計～

- SARIMAモデル
  - データ取得、整形
  - とりあえず実行（ここまでが設計上）

### 3月24日~3月30の予測



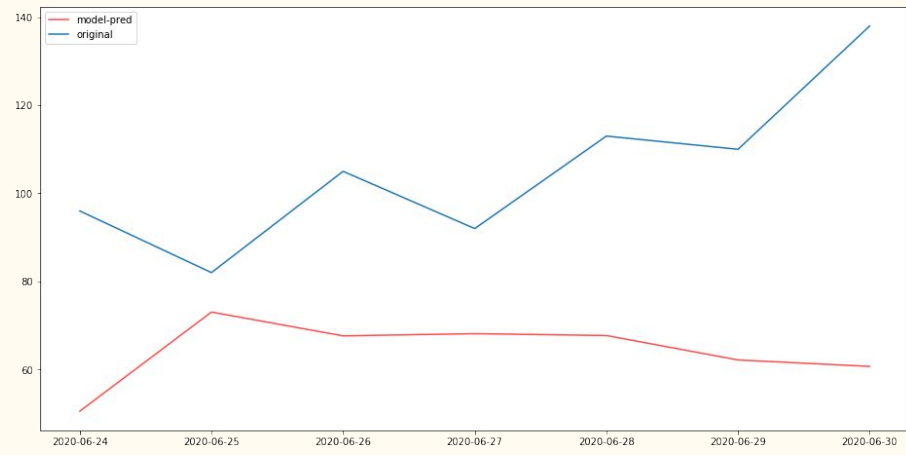
### 4月24日~4月30の予測



### 5月24日~5月30の予測



### 6月24日~6月30の予測



## 7月1日~7月7の予測

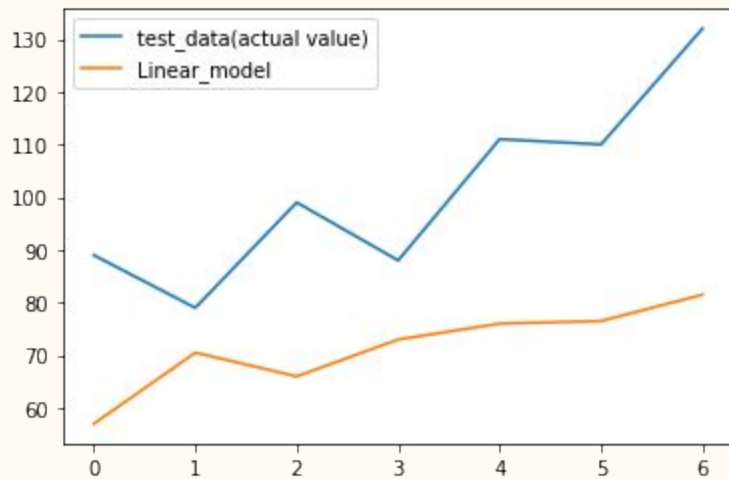
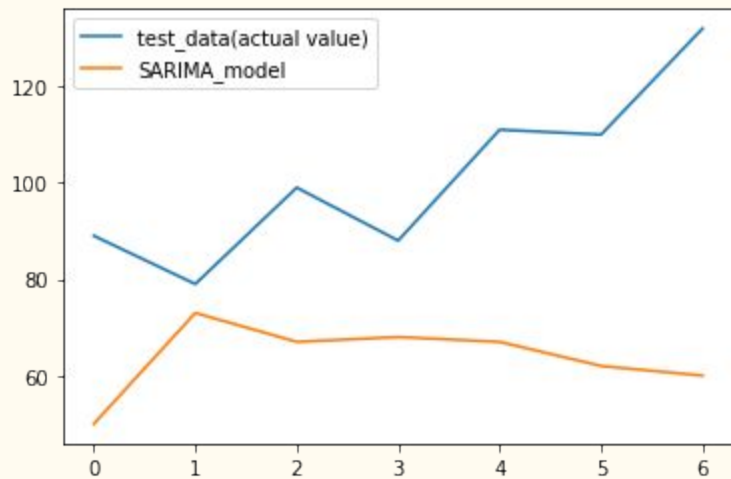


- 予測結果が安定しない (予測区間によってはめちゃくちゃ近い時もある、全然ダメな時もある (後者が多い) )
- パラメータを変えて実行 (悪化)

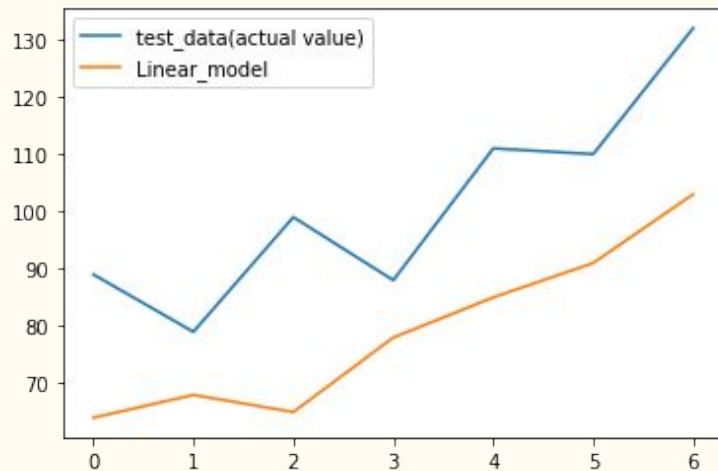
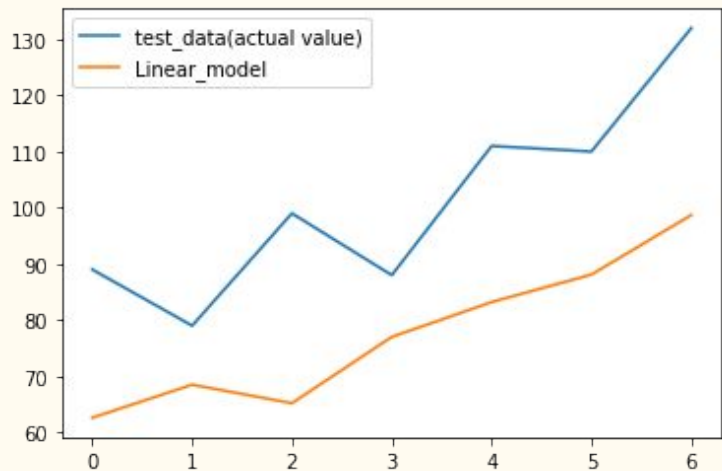
## 6.1 実験 ～実験設計～

- SARIMAモデル
  - 改善を試みる(感染者数の爆発的に増加した日の予測結果を近づけたかった)
  - 感染者数の増減は人の密集率に関係していると仮定し、新たに線形回帰モデルを実装してSARIMAモデルと結果を合成してみた。

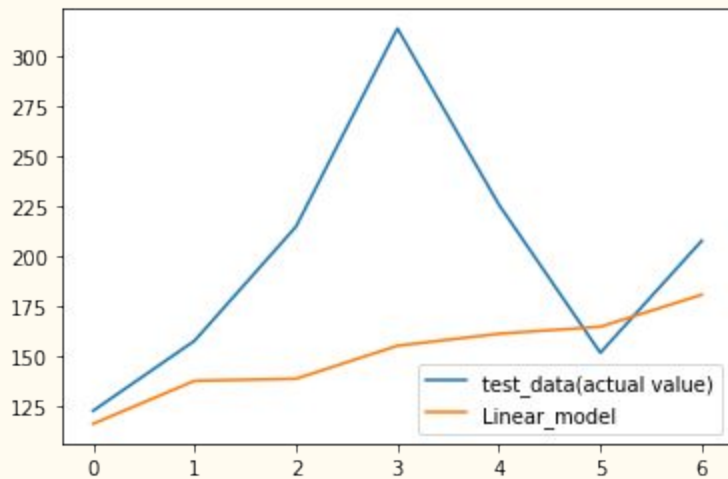
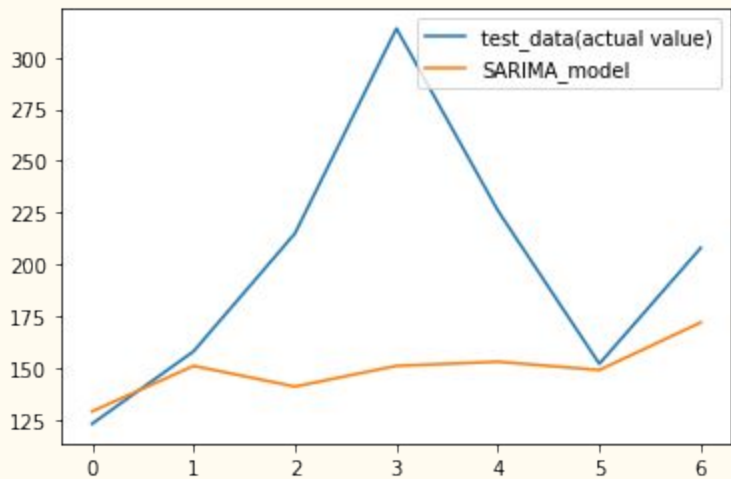




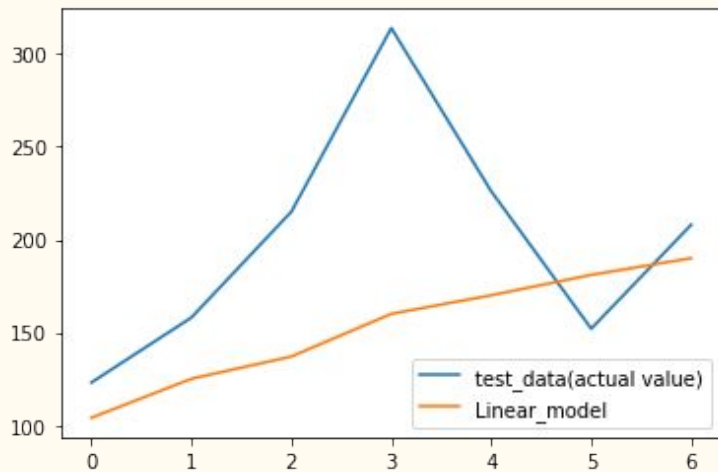
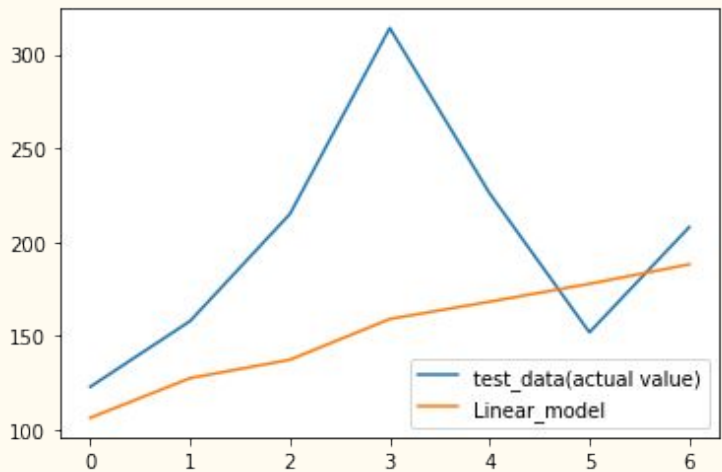
合成割合  
 SARIMA:LR  
 左上 1:0  
 右上 0.5:0.5  
 左下 0.1:0.9  
 右下 0:1



test\_data  
 6月23~6月30



合成割合  
 SARIMA:LR  
 左上 1:0  
 右上 0.5:0.5  
 左下 0.1:0.9  
 右下 0:1



test\_data  
 7月1~7月7

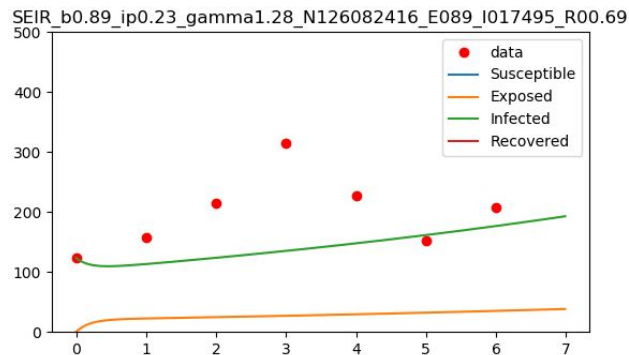
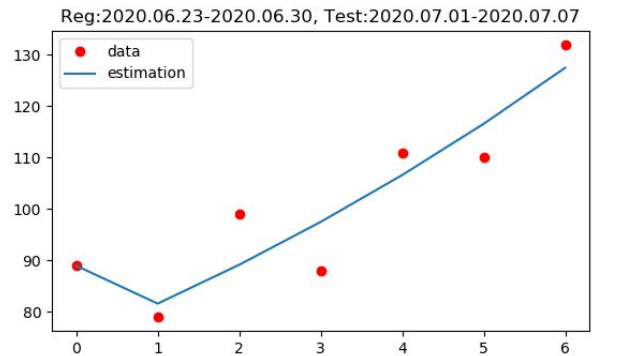
## 6.1 実験 ～実験設計～

- SARIMAモデル
  - 線形回帰モデルの方が、より良い結果が得られた。
  - 人の密集率は関係していると考えられる。
  - 爆発的に増加した日の予測結果は、ほとんど変わらなかった。

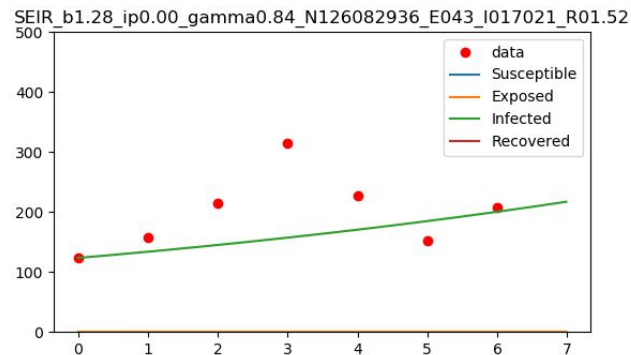
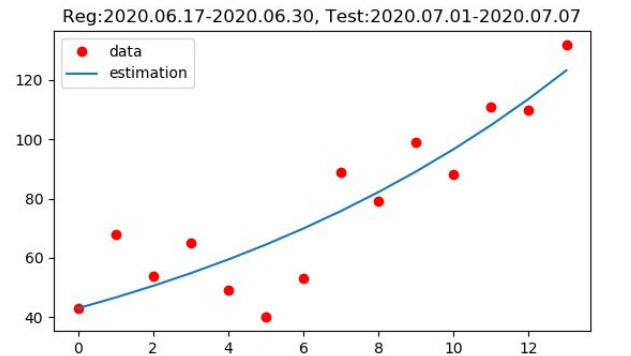
## 6.1 実験 ～実験設計～

- SEIRモデル

- パラメータ推定に使用する実データ区間を7日、14日などで変化させてみる
- モデルの初期値として用いる新規感染者数などの数値はどの段階のものを用いるかを検証する

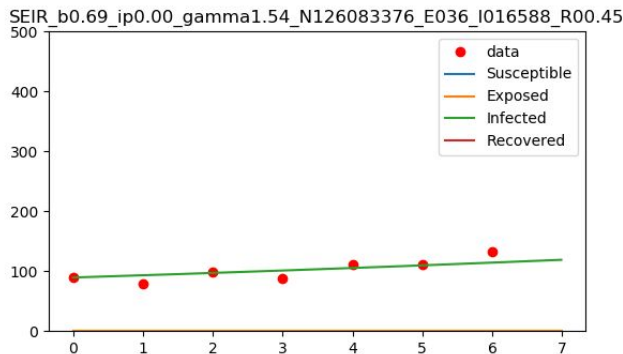
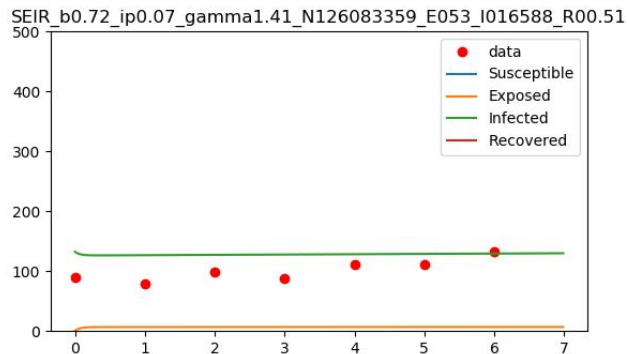
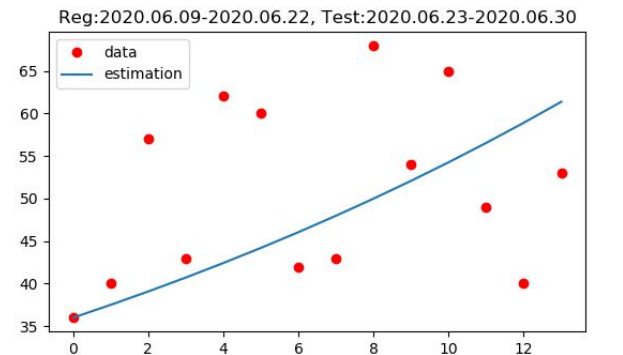
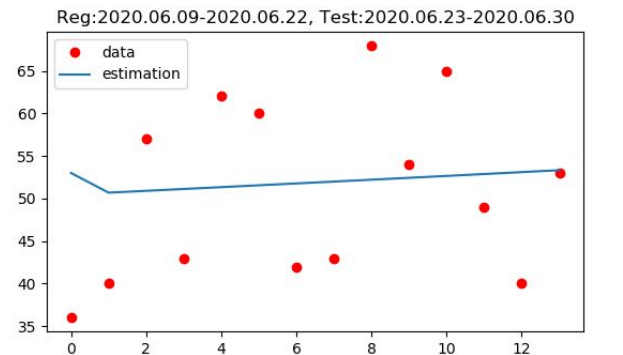


パラメータ推定区間：直近7日間  
決定係数：-1.081797804912084



パラメータ推定区間：直近14日間  
決定係数：-0.44047214184083394

パラメータ  
推定区間には  
予測区間  
(テスト区間)より  
14日間前ま  
でを採用



新規感染者数  
の初期値には  
区間の初日  
の数値を採用

新規感染者数：区間最終日  
決定係数：-2.619383670666088

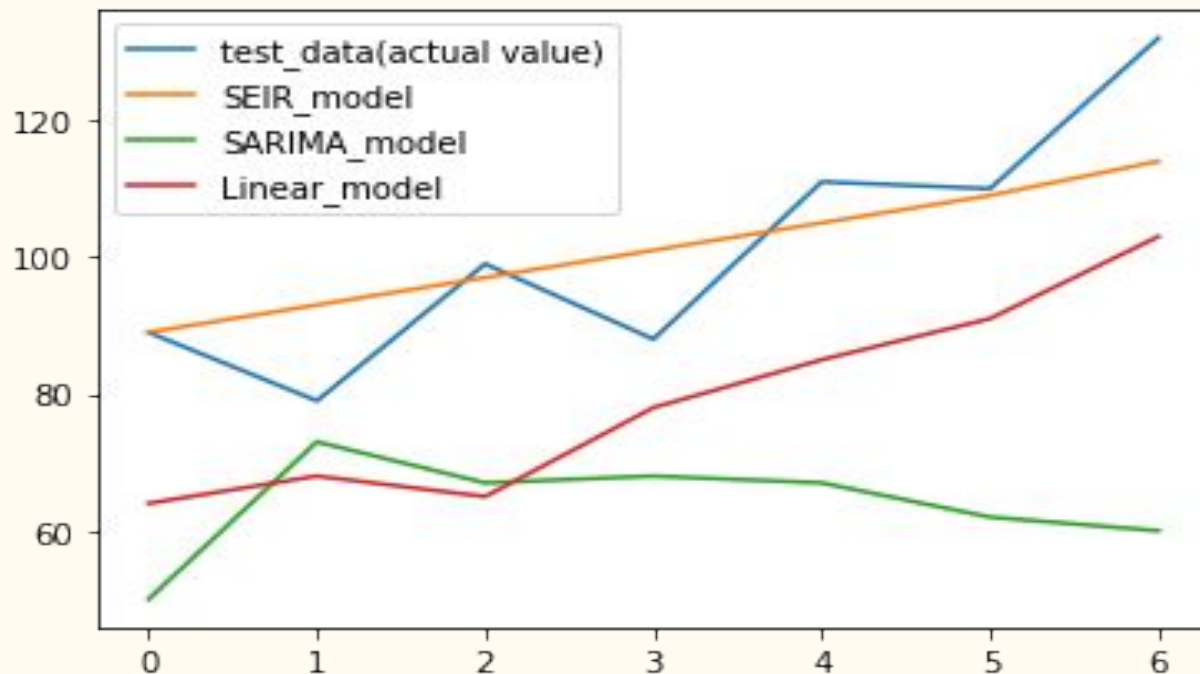
新規感染者数：区間初日  
決定係数：0.6282428719670415

## 6.1 実験 ～実験設計～

- 全体
  - コロナのような感染症の予測をするのに有効なモデルを知る
  - どちらのモデルの方が優れている(より正確に予測できた)か調べる
  - 予測する区間は7日間とする
  - 実際に得られた予測データから評価モデル(決定係数)を用いて、どちらのモデルが優れていたか判断する

## 6.2実験 ～実行結果～

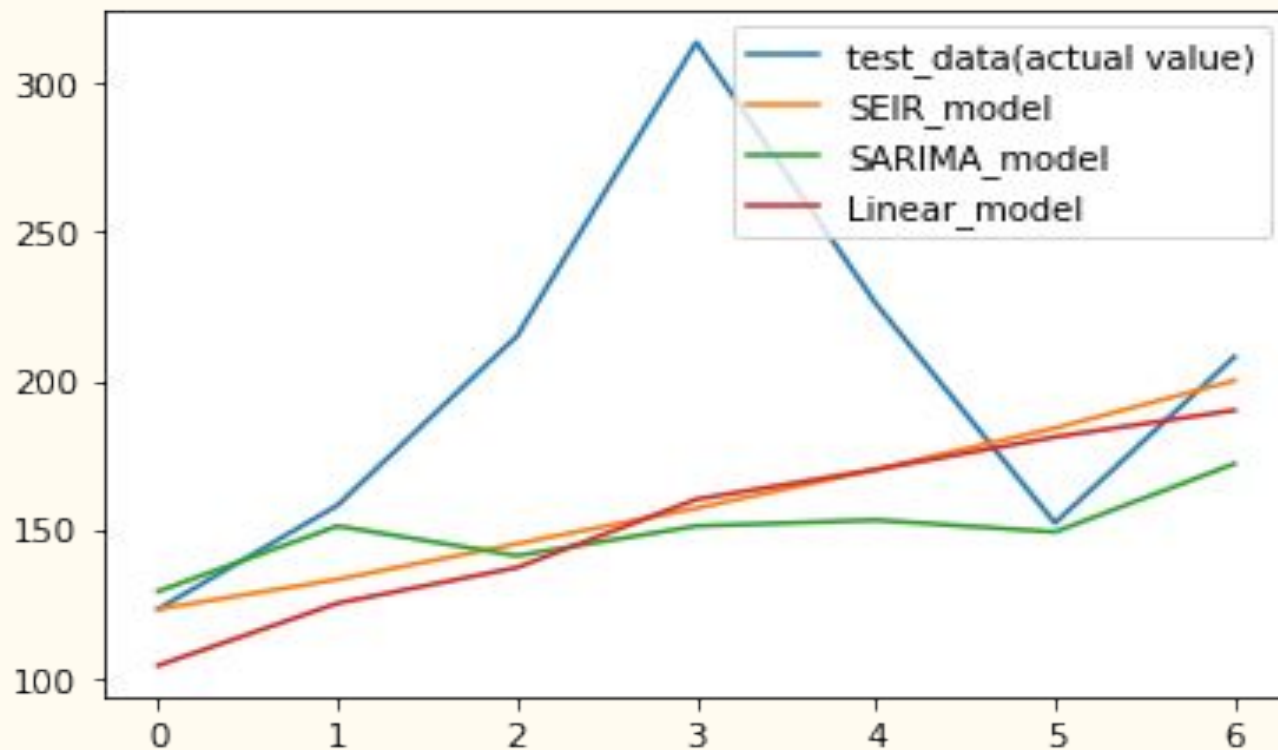
6月23~30日の区間







## 6.2実験 ～実行結果～



7月1～7日の区間



## 6.2実験 ～実行結果～

SARIMAモデルとSEIRモデルで予測した値を決定係数で評価した結果

6月23～30日の区間 SARIMA  5.384 SEIR  0.6287915291992177

7月1～7日の区間 SARIMA  0.618 SEIR  -0.4400735669398044

どちらの区間もSEIRモデルの方が1に近い値であるためコロナ感染者を予測するにはSEIRモデルの方が良い

6月23～30日

test : [89, 79, 99, 88, 111, 110, 132]

seir : [ 89, 93, 97, 101, 105, 109, 114] => 0.6287915291992177

sarima : [50, 73, 67, 68, 67, 62, 60] => -5.384

linear : [64, 68, 65, 78, 85, 91, 103] => -0.997

7月1～7日

test : [123, 158, 215, 314, 226, 152, 208]

seir : [123, 133, 145, 157, 170, 184, 200] => -0.4400735669398044

sarima : [129, 151, 141, 151, 153, 149, 172] => -0.618

linear : [104, 125, 137, 160, 170, 181, 190] => -0.484

## 6.3実験 ～考察・自己評価～

- ・推定期間によって決定係数の値が異なる
- ・パラメータの調整によってより精度をあげられる
- ・SARIMA、SEIRそれぞれ一長一短
- ・第二波やPCR検査数の違いなどでも値が大幅に変わるため継続的な予測が必要
- ・爆発的な感染者数の増減は予測できなかった。

## 7. 残された課題

- 実行結果より、推定の期間などで決定係数が大きく異なる結果となった
  - どの期間のデータを使えば最適解が得られるのか、推定する期間の長さはどれくらいが適切な  
のか調べる
- いずれのモデルでも、7月1日~7日の区間のように感染者数が急激に増減する場合に対応出来ない
  - より詳細なパラメータを含めたり、人の行動の予測をする