日本の交通事故死傷者数の予測
— ARIMAXモデルによる時系列データ分析 —

18-430 花澤楓

1. まえがき

時系列データとは，気温や株価の推移などの日，あるいは年や月，時間など一定の間隔で取られた，一連のデータのことをいう．本研究では，2006年1月から2019年12月までの日本における交通事故死傷者数のデータから未来の交通事故死傷者数を時系列モデリングによって分析・予測・評価する．また，総務省による家計調査の酒類・自動車購入費の2つの変数を外生変数としてモデルに組み込んだ場合と組み込まなかった場合，それぞれの結果も比較検討する．

2. 手法

ある時点tにおけるデータをy_tと表記する．

2.1 定常性

「得られたデータが同一の分布に従う」という条件のもとで時間に依存した前提条件を考えることで時系列データ解析が可能となる．その前提条件となるのが弱定常性というものである．弱定常性は以下のように定義される．定常性は同時分布や基本統計量の時間不変性に関するものであり，何を不変とするかによって弱定常性と強定常性の2つに分かれる．
任意のtとkに対して、

$$E(y_t) = \mu$$
$$Cov(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k$$

が成立する場合、過程は弱定常といわれる．
一方，強定常性は同時分布が不変である，というものである．
任意のtとkに対して、$(y_t, y_{t+1}, \cdots, y_{t+k})'$ の同時分布が同一である場合、過程は強定常といわれる．

2.2 White Noise

時系列データの統計学的ばらつきをWhite Noiseで表現する．
各時点でのデータが互いに独立でかつ同一の分布に従う（強定常）系列はiid系列と呼ばれる．
すべての時点tにおいて

$$E(\epsilon_t) = 0$$
$$\gamma_k = E(\epsilon_t, \epsilon_{t-k}) = \begin{cases} \sigma^2 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

が成立するとき，ϵ_tはWhite Noiseと呼ばれる．

2.3 自己共分散・自己相関

異なる時点のデータとの関係性を表す統計量が自己共分散・自己相関である．自己共分散の値が大きければ，「t時点のデータが大きければt-k時点のデータも大きくなる」とわかる．
k時点前の値との自己共分散は,

$$\gamma_{kt} = \mathrm{Cov}(y_t, y_{t-k}) = E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})] \tag{1}$$

自己相関は自己共分散の値を最大で+1,最小で-1に標準化したものであり，以下のように表せる．

$$\rho_{kt} = \mathrm{Corr}(y_t, y_{t-k}) = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t)}\mathrm{Var}(y_{t-k})} \tag{2}$$

2.4 ARIMAX Model

非定常過程に対して差分を取ることで定常過程にデータを変換したものを自己回帰和分移動平均モデル(Auto Regressive Integrated Moving Average Model)といい，外生変数が入ったARIMA ModelをARIMAX(ARIMA with eXogenous variables) Modelという．
r個の説明変数があり，時点tにおけるk番目の説明変数を$x_{k,t}$とおくとARIMAX(p,0,q)モデルは以下のように表せる．

$$y_t = c + \sum_{i=1}^{p} \alpha_i y_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{r} \beta_k x_{k,t} \tag{3}$$

2.5 RMSE

実際のデータ(y_t)からモデルによる推測値$(\hat{y_t})$を引いたものを残差と呼び，以下のように表す．

$$e_t = y_t - \hat{y_t} \tag{4}$$

RMSE(Root Mean Square Error)はT期間における予測残差の大きさを評価する．RMSEが小さければ小さいほど実データと推測値が一致している．

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \tag{5}$$

3. 対象データ

今回分析の対象とするデータは2006年の1月から2021年の12月までの月次の交通事故死傷者数である．まず，データをグラフでみる．

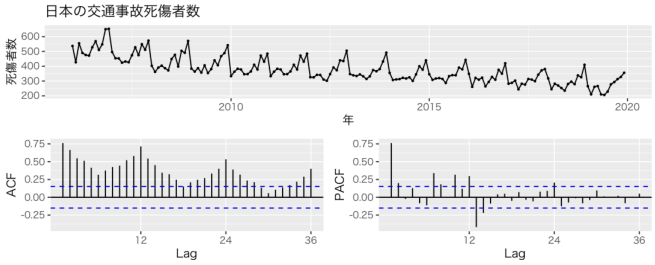

図 1　ACF:自己相関，PACF:偏自己相関

グラフから，死傷者数は一定の周期で推移しており，特に年末にかけて急激に上昇していることがわかる．月ごとのグラフは以下の通りである．

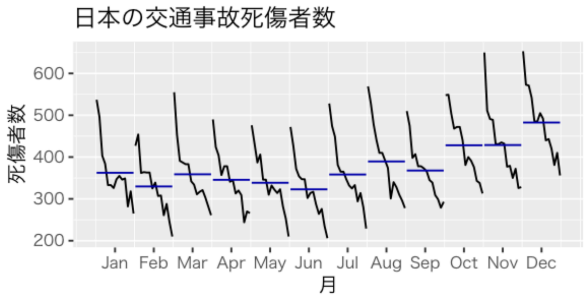

図 2　対象データ:月次

4. 実験方法

データを2006年1月から2018年12月までの訓練データと2019年1月から2019年12月までのテストデータに分けた．訓練データからモデルを推定して，テストデータとの残差で評価する．実験は以下の手順で行なった．

1. 対象データが定常性を満たすようデータを整形する．
2. 整形済みデータに対して，ARMAモデルを適用してモデルを推定する．
3. 推定されたモデルのRMSEによる評価

5. 実験結果

AR Model, ARIMAX Modelと，過去の平均値，直近の値を予測としてそのまま出したモデルのそれぞれについてRMSEを計算した．

推定されたモデル	RMSE	外生変数
AR(15)	0.2127643	なし
ARIMA(1,0,1)(0,1,1)[12]	0.1013616	自動車購入費
ARIMA(0,1,1)(0,1,1)[12]	0.1025004	酒類購入費
ARIMA(0,1,1)(0,1,1)[12]	0.1037038	なし
過去の平均値	0.3947943	なし
直近の値	0.4740179	なし

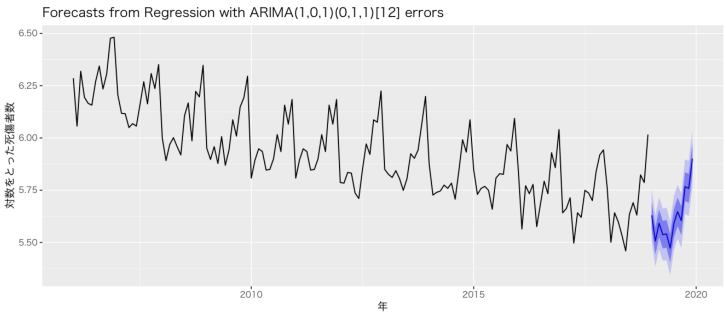

図 3　外生変数を自動車購入費としたARIMAXモデルの予測

6. 考察

AR ModelはRMSEがARIMAX Modelよりも2倍以上高い数字となったが，外生変数の有無は交通事故死亡者数にあまり影響を与えないことが実験結果から判明した．RMSEはかなり低かったので高精度で予測が行えた．時系列データは自らのデータから十分正確な予測が立てられるのではないかと考えた．

7. まとめ

各変数の相関などはうまく表せたが，因果関係は明らかにならなかった．今後は因果関係の有無を研究していきたい．

謝辞

参考文献

[1] 参考文献1