71745242 段 暁然

ベイズ統計学最終レポート

インフルエンザ定点報告数のベイズ重回帰分析

1. データの概要

今回分析するデータセットは 2013 年 40 週から 2019 年 20 週(2013 年 9 月 30 日~2019 年 5 月 19 日)までの東京都インフルエンザ定点報告数、気象データ、Google の検索トレンドをまとめたものである。インフルエンザ定点報告数は主に厚生労働省のホームページから取得したが、欠損値が存在していたため東京感染症情報センターのデータで補足を行なった。インフルエンザ定点報告数のデータは全て報告書として pdf 形式で保存されていたため、データの入力は全て手作業で行なった。気象データは気象庁のホームページより取得したものである。また Google の検索トレンドは Google Trends よりい csv フォーマットでダウンロードした。

2. 厚生労働省インフルエンザ定点報告: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/kenkou/kekkaku-kansenshou01/houdou.html
3. 東京感染症情報センター: http://idsc.tokyo-eiken.go.jp/weekly/
4. 気象庁ホームページ: https://www.data.jma.go.jp/obd/stats/etrn/index.php
5. Google Trends: https://trends.google.com/trends/?geo=US

NA 値を取り除いた後、このデータセットは合計 234 行と１０つの項目がある。それぞれの項目名と内訳は以下の通りである。また、表 1 はデータの最初の 5 行を表示している。

表 1

	week	inf_pts	temp	temp_range	rain	humidity	cloud	sunlight	wind	google
61	2014年49週	5.86	7.7	-0.8	5.0	52.0	3.4	8.76	2.4	153.75
62	2014年50週	10.15	7.2	1.0	1.5	53.0	5.1	8.47	2.6	150.88
63	2014年51週	20.60	5.5	-2.6	39.0	56.0	4.2	8.51	2.9	182.45
64	2014年52週	32.90	5.2	-3.3	5.5	52.0	4.1	8.63	2.5	259.53
65	2015年1週	11.05	5.1	-2.3	0.0	53.0	2.8	9.22	2.5	252.46

- week: 2013 年 40 週(2013 年 9 月 30 日)〜2019 年(2019 年 5 月 19 日)までの週情報
- inf_pts: 東京都のインフルエンザ定点報告数(医療機関あたりの患者数)
- temp: 東京都の一週間の平均気温©
- temp_range: 最高気温と最低気温の差
- rain: 一週間の合計降水量(mm)
- humidity: 一週間の平均相対湿度(%)
- okyo_cloud: 一週間の平均雲量(10 分比)
- sunlight: 平均全天日射量(MJ/㎡)
- wind: 平均風速(m/s)
- google: Google Trends より取得した「インフルエンザ」の検索トレンド(相対データ、無単位)

表 2

	inf_pts	temp	temp_range	rain	humidity	cloud	sunlight	wind	google
count	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000	234.000000
mean	6.101368	15.759829	8.798718	29.337607	67.551282	6.699145	13.335812	2.871795	76.546496
std	11.544633	7.671961	7.472889	41.595453	12.100772	2.027557	4.697020	0.549576	97.969083
min	0.000000	2.100000	-5.400000	0.000000	40.000000	1.500000	4.860000	2.000000	2.390000
25%	0.100000	8.275000	2.225000	2.500000	59.000000	5.300000	9.707500	2.500000	8.810000
50%	0.590000	15.950000	9.100000	15.500000	68.000000	6.900000	12.320000	2.800000	33.965000
75%	6.300000	21.900000	14.875000	39.750000	77.000000	8.400000	16.680000	3.200000	121.940000
max	64.180000	30.600000	23.800000	307.500000	94.000000	10.000000	25.770000	5.400000	607.000000

表 2.はデータセットの基本統計量をまとめたものである。各項目のカウント、平均、標準偏差、最小値、四分位偏差、そして最大値が表示されている。

2. 探索的データ分析

インフルエンザに関する事前調査から、ウイルスの流行は気温と湿度に大きく関わることがわかった。インフルエンザは飛沫感染をするタイプのウイルスで低温低湿度の環境で長く生き残ることができる。また、ネットが普及している今日、Google における「インフルエンザ」の検索数もインフルエンザの流行を予測する上で有効な指数となってくる。したがってこのセクションでは、目的変数であるインフルエンザ定点報告数と説明変数である気温、湿度、Google 検索数を詳しく見ていく。

2.1 目的変数: インフルエンザ定点報告数の分布

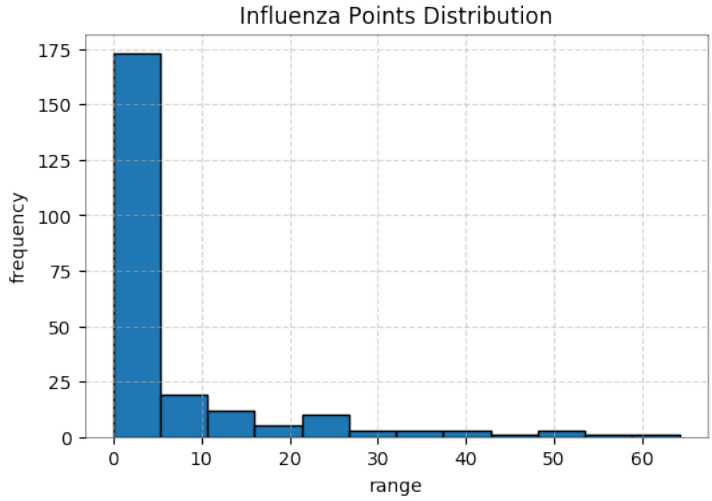

図 1

図 1.はインフルエンザ定点報告数の度数分布表である。インフルエンザ定点報告数の最大値は 64.18、最小値は 0、平均値は 6.1、中央値が 0.59 である。度数分布表から確認できるように左に歪んだ鋭い分布である。一年を通してインフルエンザが流行る時期は長くないため、このような分布になると考えられる。

2.2 説明変数: 東京都の週間平均気温の分布

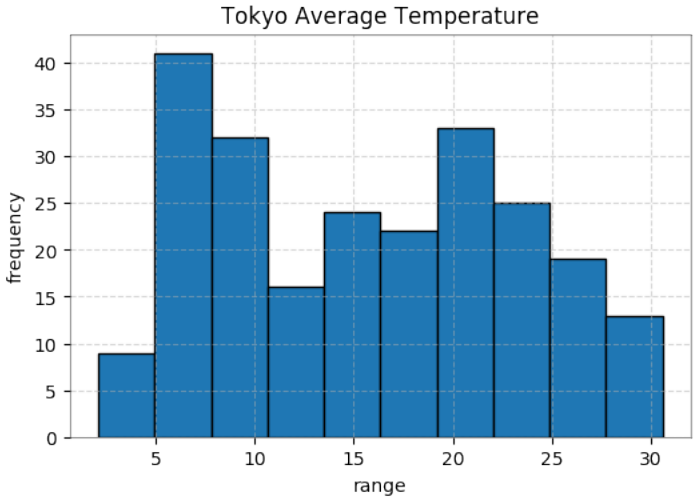

図 2

図 2 は東京都の週間平均気温の度数分布表である。東京の平均温度の最小値は 2.1、最大値は 30、平均値は 15.9、そして標準偏差は 7.6 である。標準分布に近いものになると予想していたが、チャートからピークが二つある分布になっていることがわかる。東京の冬は 0 度に近い気温が長く続くが、氷点下になることは少ないことが読み取れる。

インフルエンザ定点報告数はその週に医療機関から報告された患者数に基づいて計算されている。しかし、インフルエンザには 1~3 日間の潜伏期間がある上、症状が現れて患者が病院に訪れるまで時間がかかるため、一週間前から二週間前の気温の方が報告数に影響があると考えられる。よって、分析では気温を一週間と二週間前にずらしたものも用意するべきだと考えられる。

2.3 説明変数: 東京の週間平均湿度の分布

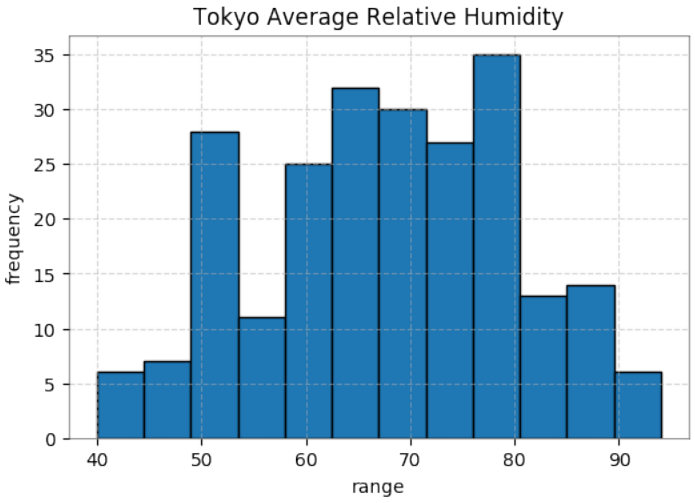

図 4

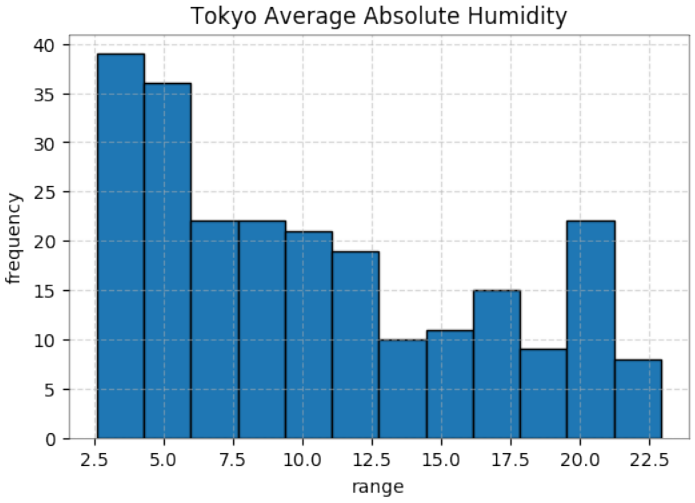

図 3

図 3 は東京都の週間平均相対湿度の分布を表している。相対湿度の最小値は 40%、最大値は 94%、平均値は 68%、標準偏差は 12.25 である。

湿度には相対湿度と絶対湿度の二つの種類がある。

- 相対湿度: ある温度の空気中に含みうる最大限の水分量（飽和水蒸気量）に比べて、どの程度の水分を含んでいるかを示す値で表す。

- 絶対湿度: 湿り空気（一般に存在する空気）中の乾き空気（全て水分を含まない空気）1kg に対する水蒸気の重量割合を示す。

インフルエンザ予測において相対湿度より絶対湿度が良いと考えられる。インフルエンザは飛沫感染型のウイルスで乾燥した空気の方が生存しやすいため、空気中の水蒸気量が予測において大切だと考えられるが相対湿度は空気中の実際の水蒸気量ではなく、気温に依存する相対的な尺度であるため、適していないと考えた。よって、相対湿度から絶対湿度への変換を行う必要があり、図４は変換を行った後の絶対湿度の分布を表している。

2.4 説明変数: Google 検索数

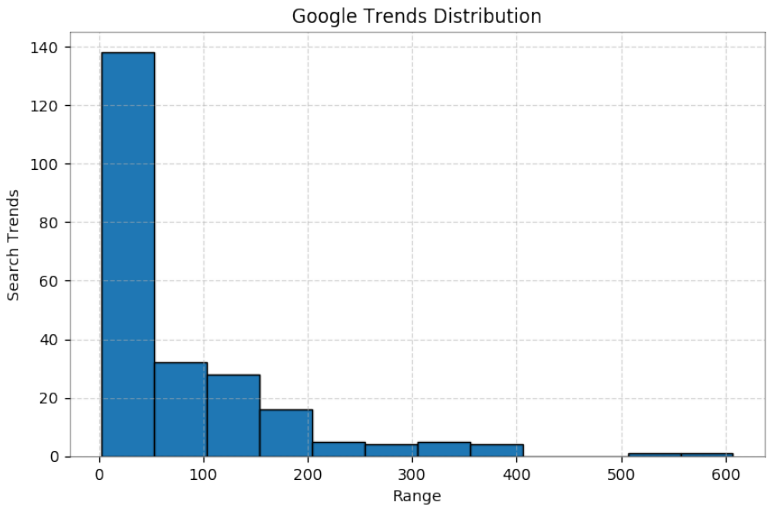

図5

図５は Google における「インフルエンザ」の検索トレンドの度数分布表である。日本ではインターネット及びネット接続端末の普及が進んでいるため、近年の情報収集はネットで行われることが多い。冬に熱を出した時、病院に行く前に先ずネットでインフルエンザ症状を調べて自己診断することが多いと考えたため、Google 検索のデータは予測において有効だと考えた。

Google Trends のデータは指定する期間によってデータ点の間隔が異なってくる。例えば一年以上の期間を取得しようとすると一ヶ月毎のデータがくるのに対し、一ヶ月や二ヶ月など短い期間を取得すると一日毎のデータがくる。また、Google Trends 上のデータは標準化されており、指定期間においてデータが 0 から 100 値をとるように設定されている。その期間における最大値は必ず 100 となり、他のデータは最大値を元に計算された相対的な指数である。最大値がわからないため、非標準化が難しい。

今回の分析の目的は相関を調べることであるため、相対的なデータでも問題はないが、インフルエンザの定点報告数は週ごとの値である。Google Trends ではデータの取得期間を選択できるものの、データ間隔の指定はできず、その上取得期間ごとに一般化されているため、データを利用する上で工夫が必要であった。先ず今回調べたい期間のデータを一ヶ月ごとに取得することで１日毎のデータを取得し、繋げ合わせた。次に調べたい期間の全体のデータを取得することで月毎のデータを取得する。月毎のデータを対応する日データに重りとしてかけることによって標準化の問を回避することができた。

2.5 まとめ

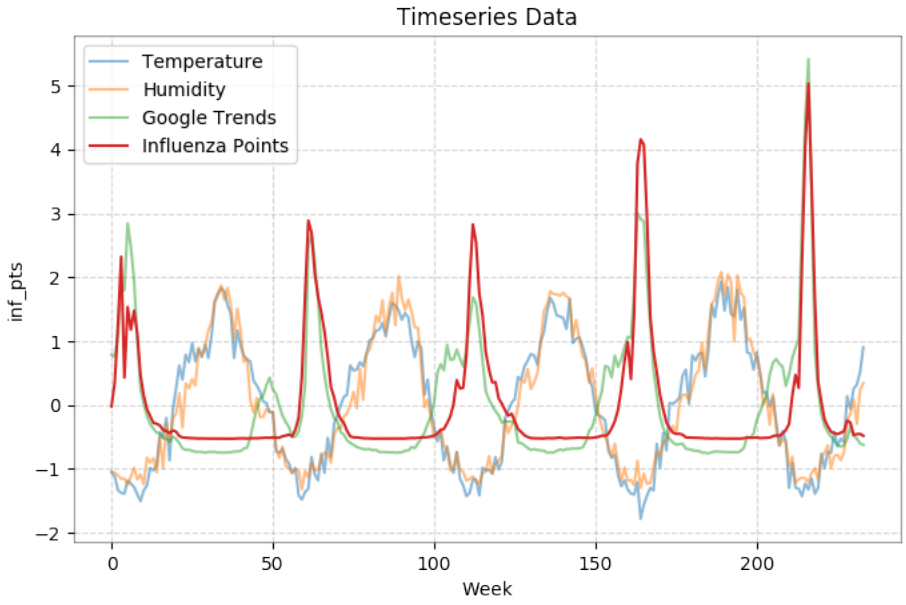

図 6

図 6 は標準化した気温、湿度、Google Trends、そしてインフルエンザ定点報告数の時系列推移である。横軸が週数で、縦軸が標準化した指数である。先ず、Google Trends とインフルエンザ定点報告数のピークが一致していることから、この二つの変数の間には強い正の相関があると考えられる。次に気温と湿度の谷がインフルエンザ定点報告数のピークに来ていることがわかる。これはインフルエンザが寒く乾燥して時期において流行しやすいという事前調査の情報を裏付けする形になっている。このチャートより、この三つの説明変数はインフルエンザ報告数と強い相関を持っていると予想することができる。

3. インフルエンザ定点報告数との相関

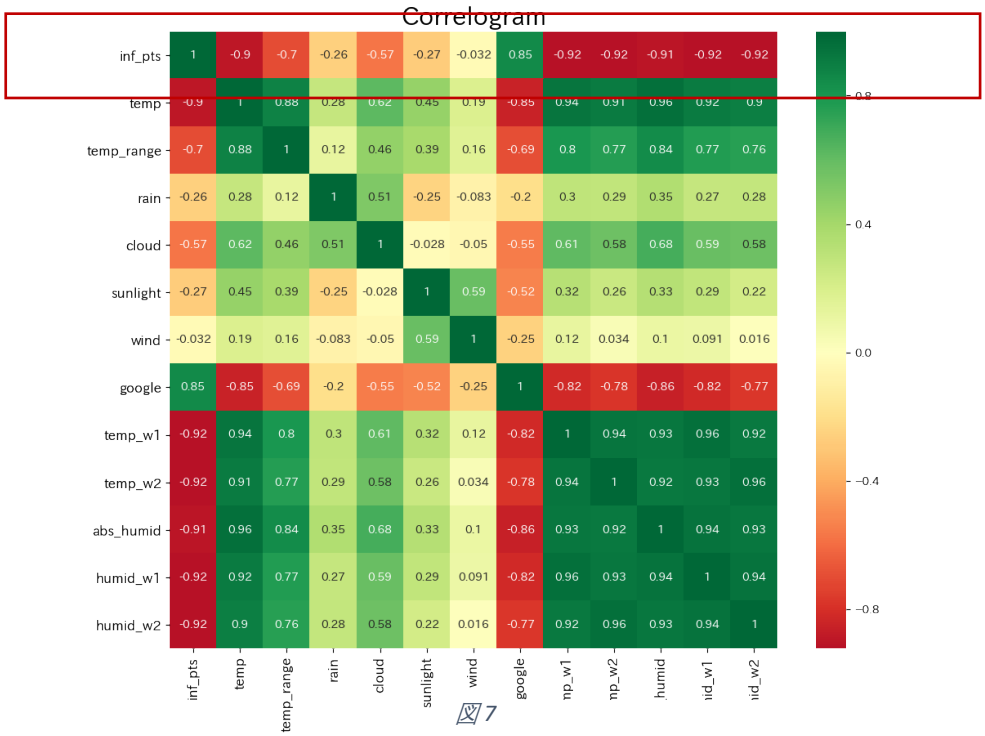

図 7

図 7 はデータセットの各変数の相関表である。合計で 13 個の変数があり、データの概要で説明した以外にも気温データを一週間と二週間前倒しした temp_w1、temp_w2 と湿度データを一週間二週間前倒しした humid_1、humid_2 が含まれている。相関を求める前にそれぞれの変数を対数化し、標準化した。対数化した理由として、インフルエンザ定点報告数のデータは指数的な性質を示していた分布だったからである。また元のデータではそれぞれの項目が異なるスケール上にあり、取りうる値の範囲も大きく異なるため、回帰係数を比較するのが難しい。標準化することにより、比較をより容易にできると考えたからだ。

このチャートはセルの色が濃いほど相関が強いことを表していて、赤方向が負の相関、緑方向が正の相関を示している。着目するべきは一行目のインフルエンザ定点報告数が他の変数間の相関である。先ず探索的データ分析でも予想したように、気温、湿度、Google 検索と非常に強い相関を見せている。また、温度差と雲の量とも中程度の負の相関がある。一方降水量、日射量、平均風速とは全く相関がないことが読み取れる。

この相関表を元に、インフルエンザ定点報告数と最も強い相関を示した変数を相関の強い順に 7 つ選択した: temp、temp_w1、temp_w2、abs_humid、humid_w1、humid_w2、google。しかし、ここでは気温と湿度に関する関数がそれぞれ三つと重複しているため、それぞれの項目の中で相関が最も弱いものを除外した(temp と abs_humid)。

4. ベンチマークの設定

4.1 評価関数

回帰モデルの精度を評価するために二つの誤差関数、Mean Absolute Error と Root Mean Squared Error を用いる。

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE)

$$RMSE = \frac{1}{n}\sum_{i=1}^{n} \sqrt[2]{(y_i - \hat{y}_i)^2}$$

二つの関数を利用するのはそれぞれに利点があるからであり、MAE は直感的に理解しやすいのに対して RMSE の方が大きな誤差を厳しく評価してくれる。

4.2 ナイーブベースラインの設定

モデルの精度を客観的に評価する上で、最低基準となるナイーブベースラインを設定する。今回はインフルエンザ定点報告数の中央値と実測値との差の絶対をベースラインとする。

$$median\ naive\ baseline\ MAE = MAE(median(y), y)$$

$$median\ naive\ baseline\ RMSE = RMSE(median(y), y)$$

MAE のベースラインは 0.7866、RMSE のベースラインは 0.9267 である。これから構築するモデルはこの数値を下回ることが目標となる。

4.3 スタンダード機械学習モデル

モデルの精度及び妥当性をより客観的に評価するため、ナイーブベースラインに加え、スタンダードの機械学習モデルも用意する。今回比較対象となるモデルは以下の通りである:

- Linear Regressor
- Elastic Net Regressor
- Random Forest Regressor
- Extra Trees Regressor
- Support Vector Regressor

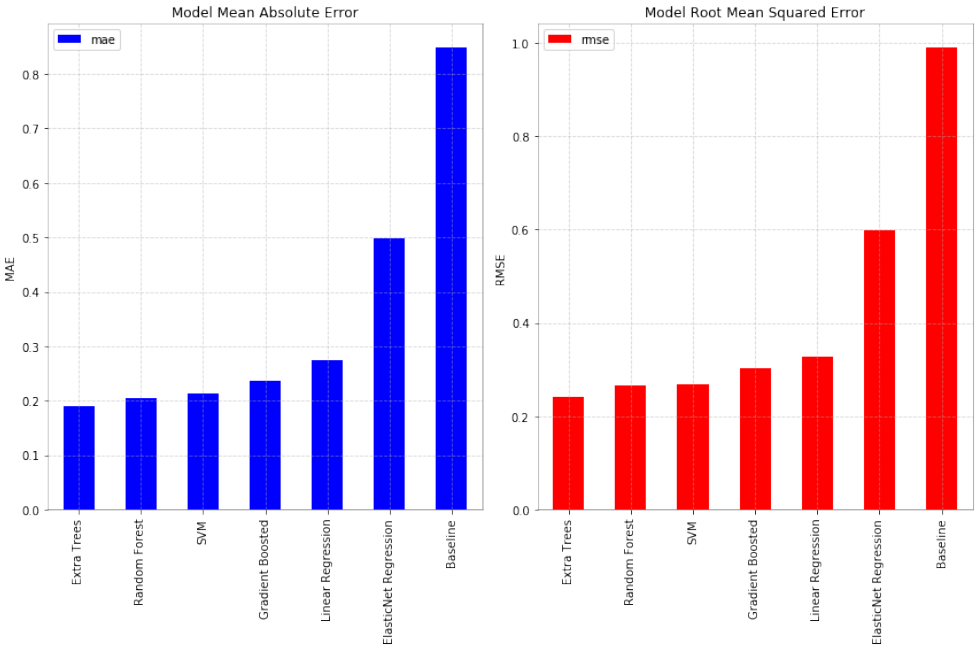

図8

	mae	rmse
Extra Trees	0.190917	0.241776
Random Forest	0.203898	0.265387
SVM	0.213961	0.269012
Gradient Boosted	0.237296	0.30428
Linear Regression	0.273938	0.32653
ElasticNet Regression	0.499299	0.598888
Baseline	0.849795	0.991545

表3

図８は各モデルのそれぞれの評価関数に対する結果を表したチャートであり、表３は具体的な数値である。図８と表３から Extra Trees Regressor が最も優れていて、ナイーブベースラインより約 77%精度が良い。

4.4 線形重回帰モデルの構築

ベイズ推定に入る前の最後のステップとして従来の線形重回帰モデルを構築する。このデータを重回帰分析した結果、回帰式は:

$$\hat{y} = -0.2 * temp_{w2} - 0.25 * humid_{w2} - 0.23 * temp_{w1} - 0.09 * humid_{w1} + 0.24 * google - 0.01$$

5. 線形回帰モデルのベイズ推定

5.1 仮説設定

探索的データ分析及び相関係数の分析の結果を元に、インフルエンザ定点報告数は temp_w2、temp_w1、humid_w1、humid_w2 とは負の相関、そして google とは正の相関があると仮定する。以下の R スタイルの式はこれらの変数がインフルエンザ定点報告数と線形的な関係にあるころとを意味している。

$$influenza \sim temp_{w1} + temp_{w2} + humid_{w1} + humid_{w2} + google$$

5.2 仮定よりベイズモデルを構築し、MCMC 法で事後分布からサンプリングする

この仮定を元に、PyMC3 ライブラリーを用いて事前分布を構築し、Markov Chain Monte Carlo 法で事後分布からサンプリングを行い、ベイズ推定を行うことができる。

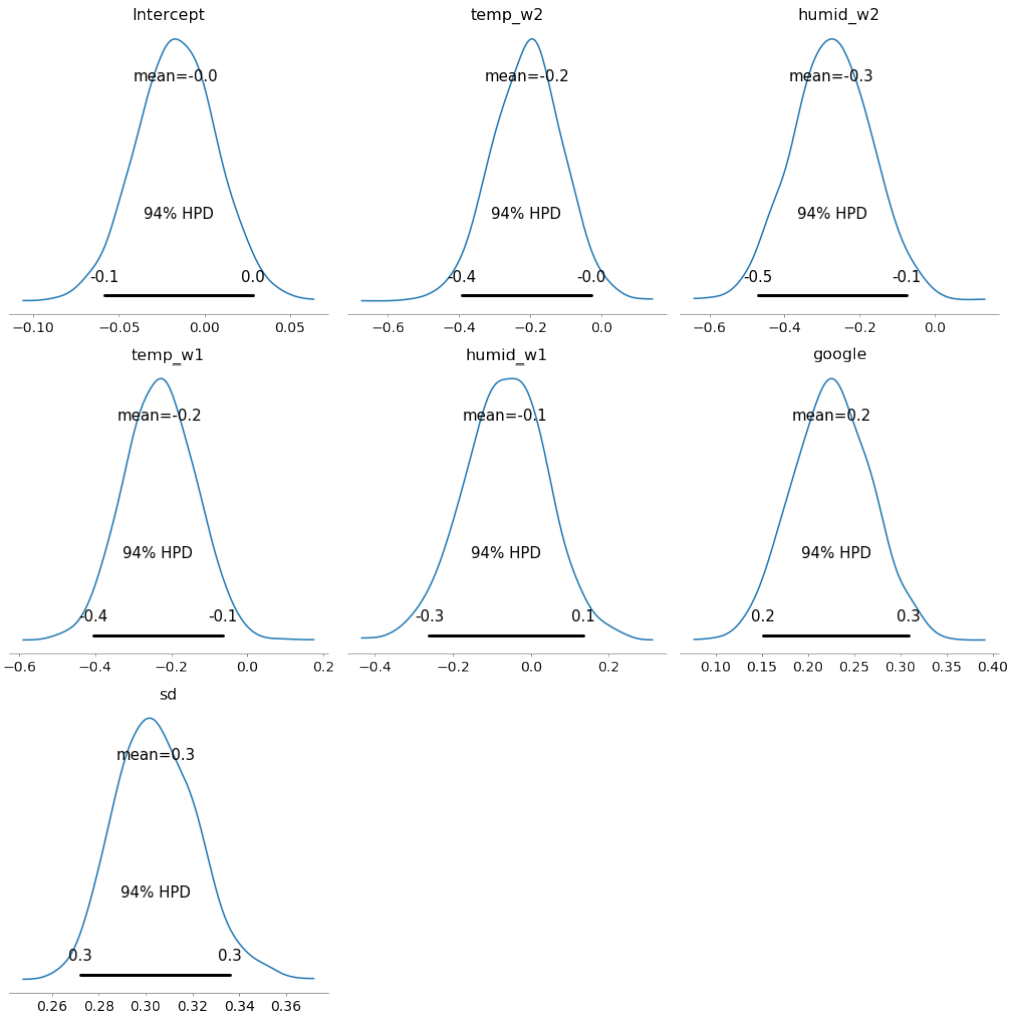

図9

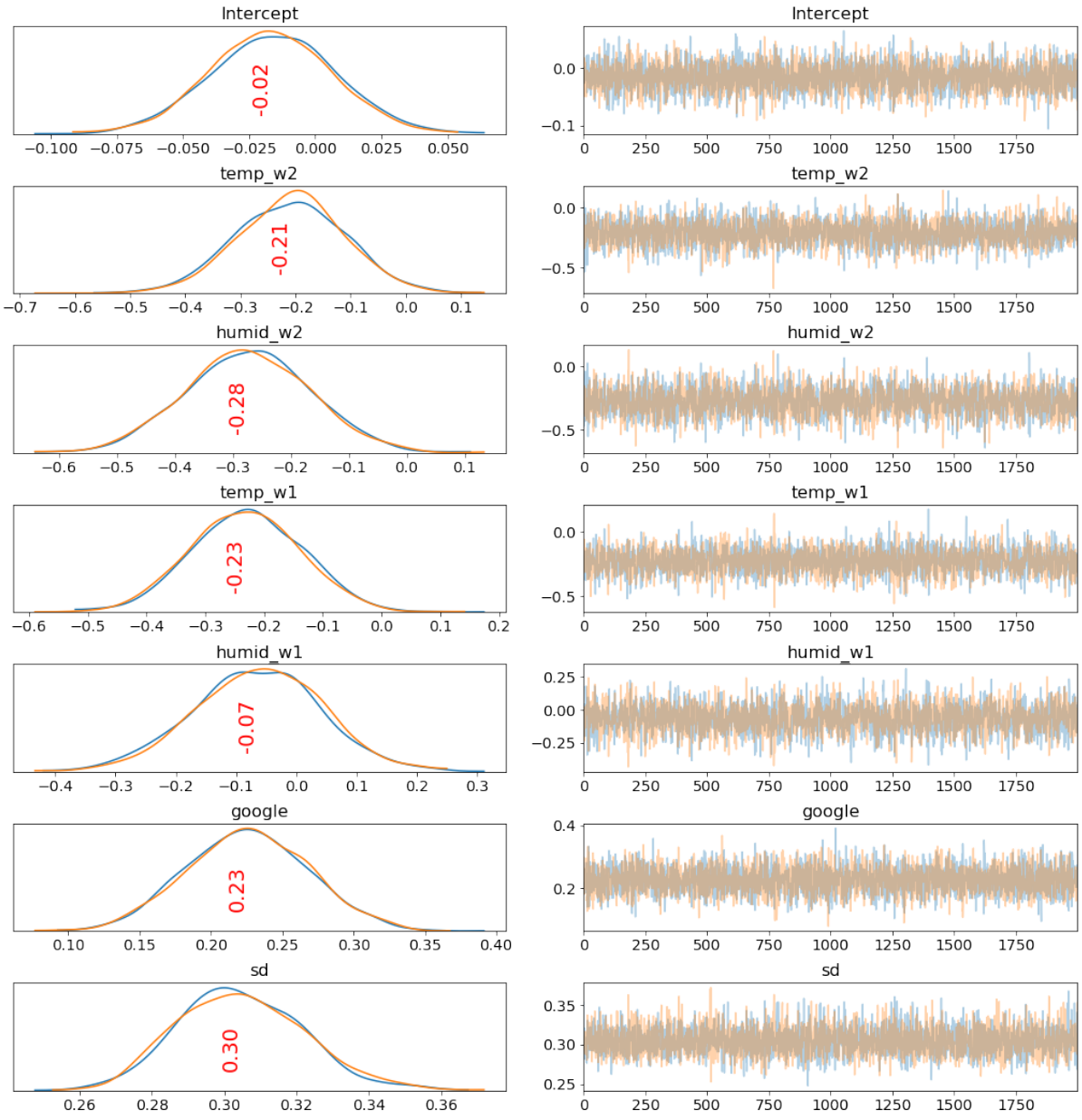

図 10

	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
Intercept	-0.016672	0.023180	0.000427	-0.064015	0.027000	3835.701321	1.000779
temp_w2	-0.210110	0.097103	0.001872	-0.403061	-0.024079	3012.343262	0.999887
humid_w2	-0.275258	0.106038	0.002222	-0.471862	-0.058245	2446.038387	1.000001
temp_w1	-0.231443	0.092945	0.001697	-0.404954	-0.050331	2814.506783	1.000646
humid_w1	-0.065935	0.105269	0.002185	-0.282716	0.131944	2049.853351	1.000754
google	0.225355	0.041959	0.000673	0.145895	0.309933	3697.872153	0.999824
sd	0.304492	0.017380	0.000290	0.271444	0.338240	3507.925219	0.999903

表 4

5.3 モデルの妥当性

これらの表と図は MCMC 方で事後分布より、2000 ランダムにサンプルしベイズ推定モデルを 2 回構築した結果である。図 9 は各変数に対する回帰係数の分布と 95%信頼区間である。図 10 の左側は回帰係数の事後分布であり、異なる二色の線は MCMC チェインを二回実行したことを意味する。そして、表 4 は結果の具体的な統計量をまとめたものである。

表 4 の全ての変数に対する R ハット値が 1.1 以下であることからモデルが収束していて妥当であることがわかる。収束している様子は図 10 の二つの事後分布が非常に近いことからも確認することができる。

5.4 変数の解釈と仮説の検証

図 9、図 10、及び表 4 の統計量から、以下のことが言える:

- temp_w2 (二週間前にずらした気温): 回帰係数の 95%信頼区間が-0.403 から-0.024 で、平均が-0.210 であることからインフルエンザ定点報告数と負の相関がある。
- temp_w1 (一週間前にずらして気温): 回帰係数の 95%信頼区間が-0.405 から-0.05 で、平均が-0.231 であることからインフルエンザ定点報告数と負の相関がある。
- humid_w2 (二週間前にずらして湿度): 回帰係数の 95%信頼区間が-0.472 から-0.058 で、平均が-0.275 であることからインフルエンザ定点報告数と負の相関がある。
- humid_w1 (一週間前にずらして湿度): 回帰係数の 95%信頼区間が-0.283 から-0.132 で、平均が-0.065 であることからインフルエンザ定点報告数と負の相関がある。
- google (Google における検索数): 回帰係数の 95%信頼区間が-0.145 から-0.310 で、平均が 0.225 であることからインフルエンザ定点報告数と正の相関がある。

モデルが妥当であることと、以上のことから仮説は正しいと判断できる。さらに、十分なデータがあれば、ベイズ推定の回帰分析は従来の OLS 回帰分析に近くと言われている。実際、以前に求めた従来の回帰分析式と、ベイズ推定の回帰分析から導き出した回帰分析式を比較すると回帰係数が非常に近いことがわかる。

Ordinary Least Square Regression

$$\hat{y} = -0.2 * temp_{w2} - 0.25 * humid_{w2} - 0.23 * temp_{w1} - 0.09 * humid_{w1} + 0.24 * google - 0.01$$

Bayes Regression

$$\hat{y} = -0.21 * temp_{w2} - 0.26 * humid_{w2} - 0.23 * temp_{w1} - 0.07 * humid_{w1} + 0.23 * google - 0.02$$

5.4 他のモデルとの比較

最後に、ベイズ推定モデルの精度を他の機械学習モデルの精度と比較する。図 11 と表 5 はそれぞれのモデルの精度をまとめたものである。どの誤差関数を利用して評価するかによって順位に違いはあるが、ベイズ推定モデルは一般的な線形回帰モデルと同じくらいの精度であることが読み取れる。MAE を用いて評価する場合ベイズ推定モデルの精度はナイーブベースラインの 67.83%優れている。

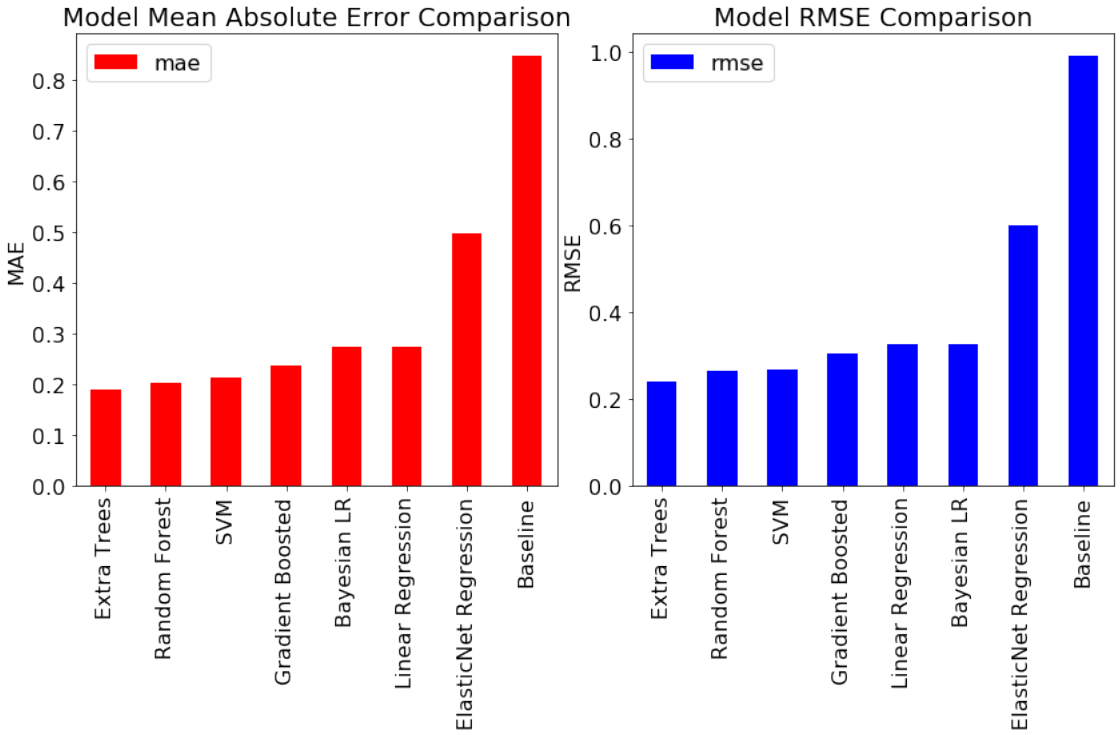

図 11

	mae	rmse
Extra Trees	0.190917	0.241776
Random Forest	0.203898	0.265387
SVM	0.213961	0.269012
Gradient Boosted	0.237296	0.30428
Bayesian LR	0.273645	0.327369
Linear Regression	0.273938	0.32653
ElasticNet Regression	0.499299	0.598888
Baseline	0.849795	0.991545

表 5

6. 分析に用いたコード

個人的に R より Python の方が熟練度が高いため、この課題のための分析は全て Python で行った。そのソースコード、Jupyter Notebook、Python の環境設定、使用したデータは全て以下の Github ベージにアップロードしてある。

https://github.com/ShozenD/Bayes-Statistics.git