論文

岡山県の気象データを用いた重回帰分析に関する研究

岩本 隆志 1)

キーワード：重回帰分析、決定係数、平均二乗誤差、PPDCA、L1 正則化、L2 正則化

1 はじめに

2025 年には、SAP社のR/3 に代表されるソフトウェアのサポート終了が相次ぐ。更に、レガシィシステムへの対応において、IT技術者の大量定年退職等により改定・改修や運用保守業務対応が困難になるといった問題がある。この問題を解消しなければ、2025 年以降で最大 12 兆円/年の経済損失の可能性があり、それは「2025 年の崖」とも呼ばれている。このような社会情勢下において、DXを推進できる人材の確保が急がれる。大学の教育現場においても、ありとあらゆる学部においてDXを推進できる人材の育成が急務であり、既にデータサイエンス教育が始まっている。このような状況下、気象庁が提供している岡山県の気象データを用いて重回帰分析を行い、PPDACサイクルを回すことにより、新たな知見を見出すことを本研究の目的とする.

1.1 重回帰分析

説明変数をx_1、x_2、…、x_M、M個の重みをw_1、w_2、…、w_Mとすると重回帰分析は、

$$y = w_1x_1 + w_2x_2 + \cdots + w_Mx_M + b$$

と定義される。総和の記号を使って表記すると、

$$y = \sum_{m=1}^{M} w_m x_m + b$$

となる。重回帰分析では、 w_1、w_2、…、w_Mとバイアスb があり、合わせて M+1 個のパラメータが存在する。これらのパラメータの定式化を考える。ここで仮に$x_0 = 1$、$w_0 = b$とすると、

$$y = w_1x_1 + w_2x_2 + \cdots + w_Mx_M + b$$
$$= w_0x_0 + w_1x_1 + \cdots + w_Mx_M$$

となる。上式をベクトルの内積を用いて表記すると、

1) 山陽学園大学総合人間学部生活心理学科

$y = w_0 x_0 + w_1 x_1 + \cdots + w_M x_M$
$= \mathbf{w}^{\mathrm{T}}\mathbf{x}$
と表現できる。次に、目的関数を求める。ここで、単回帰分析と同じ目的関数を
$(t_1 - y_1)^2 + (t_2 - y_2)^2 + \cdots + (t_N - y_N)^2$
とし、ベクトルの内積を用いて表記し直すと、
$L(t_1 - y_1)^2 + (t_2 - y_2)^2 + \cdots + (t_N - y_N)^2$
$= (\mathbf{t} - \mathbf{y})^{\mathrm{T}}(\mathbf{t} - \mathbf{y})$
となる。モデルの方程式$\mathbf{y} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$を変形し、展開すると、$y = \mathbf{X}\mathbf{w}$と表記できる。次に、パラメータを最適化するために、目的関数L を最小化するモデルのパラメータベクトル$\mathbf{w}$ を求める。目的関数をパラメータで微分して 0 とおき、$\mathbf{w}$ について解く。目的関数の予測値yを、$\mathbf{w}$を用いた表記に置き換えると、

$(\mathbf{t} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{t} - \mathbf{X}\mathbf{w})$
$= (\mathbf{t} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{t} - \mathbf{X}\mathbf{w})$
$= \{\mathbf{t}^{\mathrm{T}} - (\mathbf{X}\mathbf{w})^{\mathrm{T}}\}(\mathbf{t} - \mathbf{X}\mathbf{w})$
$= (\mathbf{t}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}})(\mathbf{t} - \mathbf{X}\mathbf{w})$
となる。次に、分配法則を用いて展開すると、
$\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{t} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}$
となる。この目的関数に対し$\mathbf{w}$について偏微分すると、
$(\mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w})^{\mathrm{T}} = \mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w}$
となる。目的関数L は、
$\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}$
と変形でき、上式をw 以外の定数項を一つにまとめると、
$\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbf{t}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}$
$= \mathbf{t}^{\mathrm{T}}\mathbf{t} - 2(\mathbf{X}^{\mathrm{T}}\mathbf{t})^{\mathrm{T}}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}$
$= c + \mathbf{b}^{\mathrm{T}}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$
となる。次に、目的関数を最小にするパラメータ$\mathbf{w}$の求め方について考える。目的関数値は、
$\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} + \mathbf{b}^{\mathrm{T}}\mathbf{w} + c$
となる。目的関数L をw_1とw_2の二次関数と見た場合、目的関数が任意のM個の変数により特徴づけられており、目的関数がそれぞれのパラメータについて二次形式になっており
$M + 1$個の連立方程式、
$\frac{\partial}{\partial w_0} L = 0$
$\frac{\partial}{\partial w_1} L = 0$
$\vdots$
$\frac{\partial}{\partial w_M} L = 0$
の解となる。上式をベクトルによる微分を用いて表記すると、

$$\Rightarrow \frac{\partial}{\partial \mathbf{w}} L \quad = \mathbf{0}$$

上式をw について解くため、左辺について整理をすると、

$$\frac{\partial}{\partial \mathbf{w}}(c + \mathbf{b}^{\mathrm{T}}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w})$$

$$= \frac{\partial}{\partial \mathbf{w}}(c) + \frac{\partial}{\partial \mathbf{w}}(\mathbf{b}^{\mathrm{T}}\mathbf{w}) + \frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w})$$

$$= \mathbf{0} + \mathbf{b}^{\mathrm{T}} + \mathbf{w}^{\mathrm{T}}(\mathbf{A} + \mathbf{A}^{\mathrm{T}})$$

となる。上式を 0 とおき、$\mathbf{A}$、$\mathbf{b}$ を展開すると、

$$\begin{aligned} -2(\mathbf{X}^{\mathrm{T}}\mathbf{t})^{\mathrm{T}} + \mathbf{w}^{\mathrm{T}}\{\mathbf{X}^{\mathrm{T}}\mathbf{X} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\} \quad &= \mathbf{0} \\ -2\mathbf{t}^{\mathrm{T}}\mathbf{X} + 2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} \quad &= \mathbf{0} \\ \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} \quad &= \mathbf{t}^{\mathrm{T}}\mathbf{X} \end{aligned}$$

と変形できる。ここで両辺を転置すると、

$$\begin{aligned} (\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}} \quad &= (\mathbf{t}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}} \\ \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} \quad &= \mathbf{X}^{\mathrm{T}}\mathbf{t} \end{aligned}$$

となる。新しい入力変数の値

$\mathbf{x}_q = [x_1、 \ldots、 x_M]^{\mathrm{T}}$

に対し、目標値y_qを予測するために、訓練により決定されたパラメータ$\mathbf{w}$を用いて、

$y_q = \mathbf{w}^{\mathrm{T}}\mathbf{x}_q$

となる［1-6］。

1.2 正則化

機械学習モデルは未知データへの予測精度を高めるために訓練データを学習する。機械学習モデルが訓練データを過剰に学習すると未知データへの予測精度が落ちることがある。これはモデルが訓練データに対して過剰に学習したため、はずれ値やノイズまで学習してしまったと考えることができる。このような現象を過学習と呼ぶ。この過学習が起きる原因としてデータ数が少ない、変数が多い、パラメータが大きすぎるといったことがある。機械学習モデルは損失関数を最小にするパラメータを学習することにより算出することができる。損失関数を$f(x)$、パラメータλとした場合、正則化を行わない損失関数は、次式で表される。

$minf(x)$

次にL1 正則化は、損失関数とL1 正則化項を追加したものであることにより、パラメータをw_iとおくと、

$$minf(x) + \lambda \sum_{i=1}^{n} |w_i|$$

となる。L2 正則化は、損失関数にL2 正則化項を追加したものであることにより、

$$minf(x)+\frac{\lambda}{2}\sum_{i=1}^{n}|w_i|^2$$

と表される。正則化はパラメータが小さいモデルをより高く評価することが理解できる。このパラメータの大きさを評価する項はペナルティ項と呼ばれ、正則化に用いるパラメータλを大きくすると過学習に陥りにくくなる。L1 正則化はパラメータの絶対値の和を、L2 正則化はパラメータの二乗和を罰則項としている。L1 正則化は余分な説明変数を省くことを目的とした手法であり、説明変数を省くために用いられる。L1 正則化を用いた学習ではパラメータw_iが 0 と評価されやすいことをうまく利用して次元圧縮を実現している。L2 正則化は、モデルの過学習を防ぐことで精度を高めるために用いられることとなる。また、L1 正則化、L2 正則化、それぞれを施した線形回帰をLasso回帰［注 1］、Ridge回帰［注 2］と呼ぶ［7-10］。

2　評価指標

重回帰分析で得られたモデルの適合の良さを評価する指標として、平均二乗誤差（MSE）、平均絶対誤差（MAE）や決定係数などが用いられている。解析目的に応じて、これらを使い分ける必要がある。例えば、観測値と回帰モデル出力値の差を出すサンプルを出来るだけ少なくしたい場合、平均二乗誤差（MSE）を使用し、全サンプルの誤差を平等に評価して、サンプル全体の誤差をできるだけ小さくしたい場合、平均絶対誤差（MAE）を使用する等の評価方法を用いる。

2.1 決定係数

重相関係数Rを 2 乗した値であり、重回帰分析における実測値の分散に対する予測値の分散の割合で、重回帰式の適合性を評価する指標となる。絶対的な基準ではないが、0.5 以上であれば適合度が高く、精度が高いと言える。複数の重回帰分析の結果を比較して、適合度の優劣を決めたい場合、決定係数の大きい方を選ぶべきである。ただし、比較する重回帰分析の結果は、説明変数の数が等しい時に限る。決定係数は説明変数の数が多くなると自動的に大きくなるという欠点を補うため、説明変数の数で調整した決定係数を自由度調整済み決定係数［注 3］と呼ぶ。数学的には、観測値を $\mathbf{y}_i(i=1、2、3、\ldots、n)$、モデルから計算した計算値（予測値）を$\hat{y_i}$、観測値の平均を$\overline{y}$とすると、

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

と定義される［11-12］。

2.2　平均二乗誤差

それぞれのデータに対して、実際の値と予測値の差の 2 乗を計算し、その総和をとり、データの総数で割った値である。 平均二乗誤差 (MSE)では、値が小さいほど誤差の少ないモデルと言える。MSEはMean Squared Errorの略であり、最も一般的な損失関数として使われるだけでなく、主に回帰問題における出力層の評価関数としても用いられる。いずれの関

数から出力される値も、0 に近いほどより良い。数学的には、観測値を $\mathbf{y}_i(i=1、2、3、\ldots、n)$、モデルから計算した計算値（予測値）を$\hat{y_i}$とすると、

$$MSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y_i})^2}$$

で定義される［13-14］。

2.3　平均絶対誤差

MAEは残差変動の絶対値の和であり、MSEとは異なり、残差変動を二乗しないので全サンプルの誤差を平等に評価すると言える。サンプル全体に対する誤差を小さくしたい時に、用いることが多い。数学的には、観測値を $\mathbf{y}_i(i=1、2、3、\ldots、n)$、モデルから計算した計算値（予測値）を$\hat{y_i}$とすると、

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

で定義される［15-16］。

3　実験

岡山県の気象データを、気象庁の「過去の気象データ検索」［17］をもとに抽出し、重回帰分析を実施した。使用した説明変数と目的変数を表 1 に示す。

表 1　目的変数と説明変数

目的変数	**平均気温**
説明変数（条件 1）	年間総降水量、年間平均大気圧、最低気温、平均湿度、平均大気量、年間降水量、年間の霧の日、年間平均地圧、年間日照時間、平均年間雲量
説明変数（条件 2）	年平均気圧、最低気温、平均湿度、平均風量、年降水量、年平均地気圧、年日照時間、年平均雲量
説明変数（条件 3）	年間平均気圧、最低気温、平均湿度、年間降水量、年間日照時間
説明変数（条件 4）	最低気温、平均湿度、年間降水量、年間日照時間
説明変数（条件 5）	最低気温、平均湿度、年間日照時間

次に、処理手順を以下に示す。

① 前処理：元データは、年・月・日単位となっていたため、日単位に加工し、分析を実施することとした。

② データの可視化：各変数間のペアプロットを表示し、データの相関を確認した結果を表 2 に示す。

・モデルの作成：表 1 の条件 1 に示した目的変数と説明変数を指定したモデルを作成し、重回帰分析を実装し、モデルを適応させる。但し、本研究ではL2 正則を用いて実装した。

モデルの検証：表 1 の条件 1 において、訓練データとテストデータの比率を変化させて処理を実行した結果を表 3 に示す。次に、目的変数は固定し、説明変数を表 1 の条件 2 から 5 と変えて処理を実行し、平均二乗誤差と決定係数を測定し、検証を実施した結果を表 4 に示す。

・外れ値がないかの確認：残差プロットを用いて外れ値がないかを確認した結果を表 5 に示す。

表 2　分析データのペアプロット

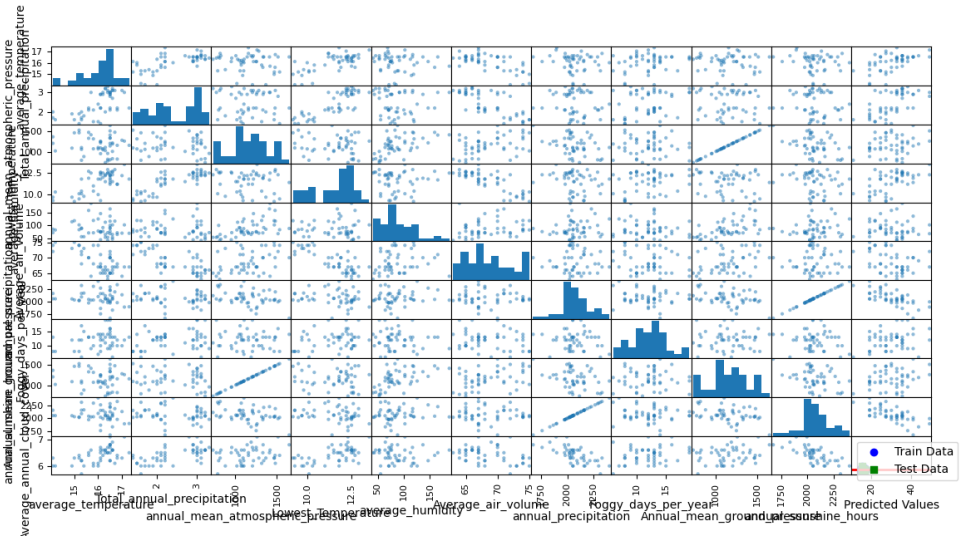


表 3　表 1 の条件 1 でのR^2とMSE

Training set rate	Test set rate	MSE Training	MSE Test	R^2 Training	R^2 Test
0.1	0.9	0.20	0.31	0.61	0.43
0.2	0.8				
0.3	0.7				
0.4	0.6				

表 4 Conditions2 からConditions5 の条件でのR^2とMSE

	MSE Training	MSE Test	R^2 Training	R^2 Test
Conditions2	0.211	0.263	0.591	0.52
Conditions3	0.278	0.264	0.461	0.517
Conditions4	0.277	0.269	0.463	0.507
Conditions5	0.278	0.267	0.46	0.511

表 5　残差プロット

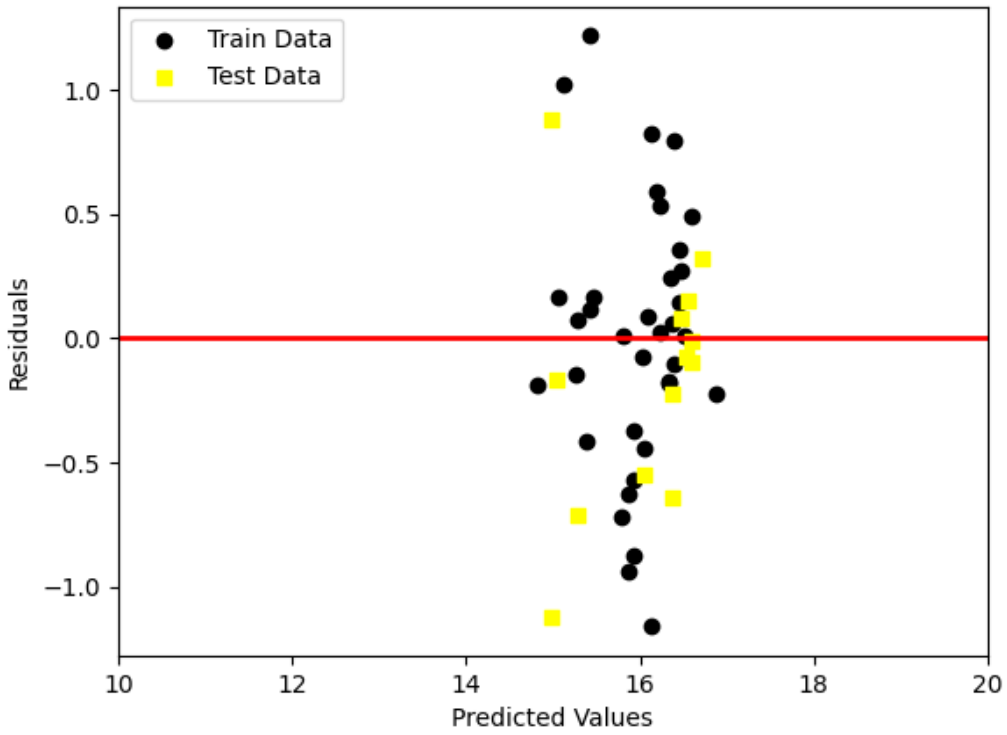


表 6　Conditions2 の散布行列図

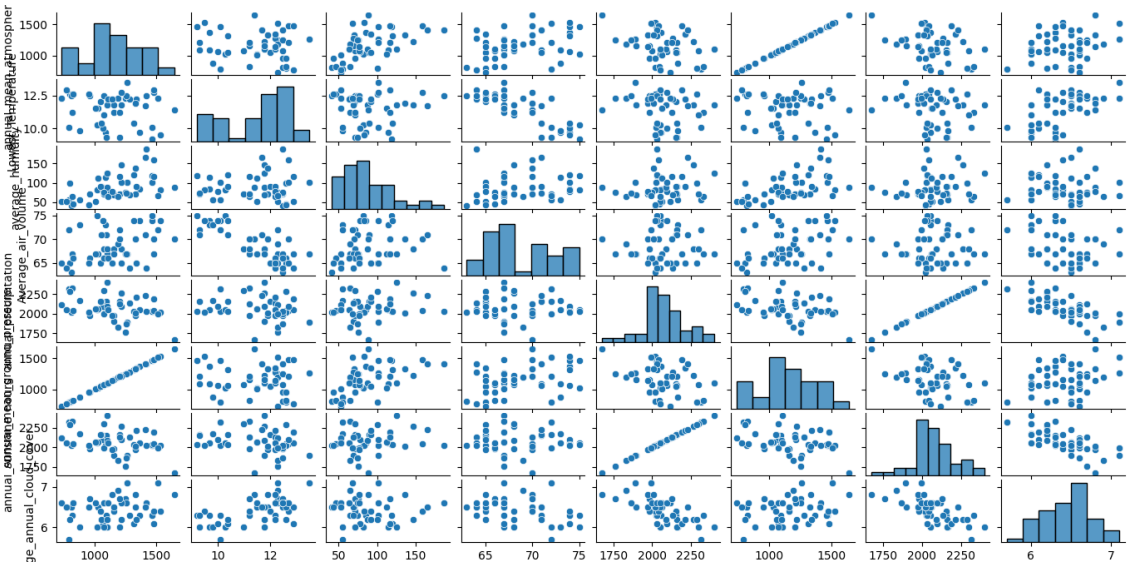


表 7　Conditions3 の散布行列図

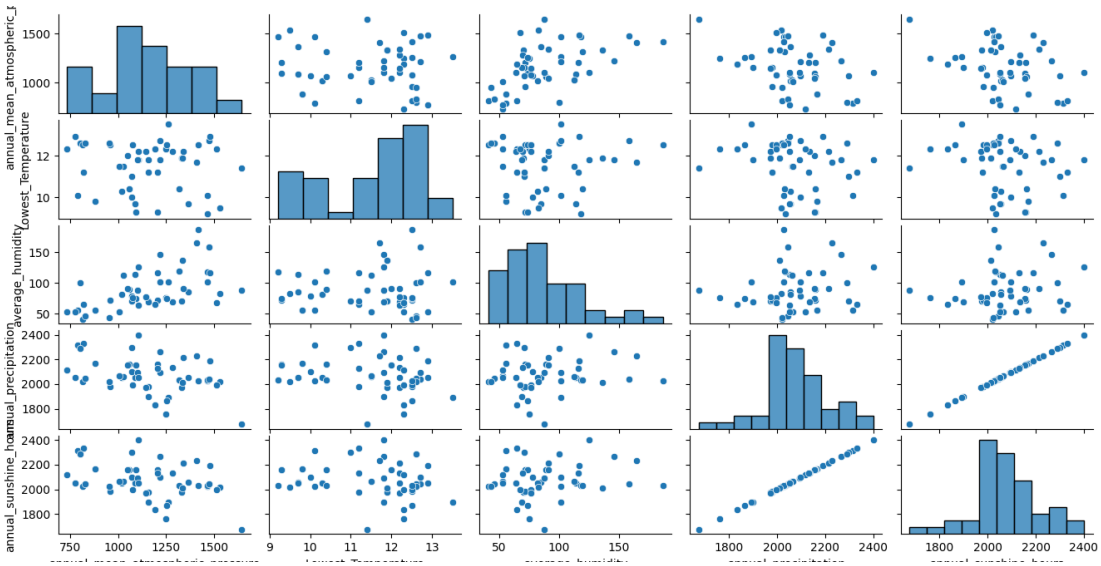


表 8　Conditions4 の散布行列図

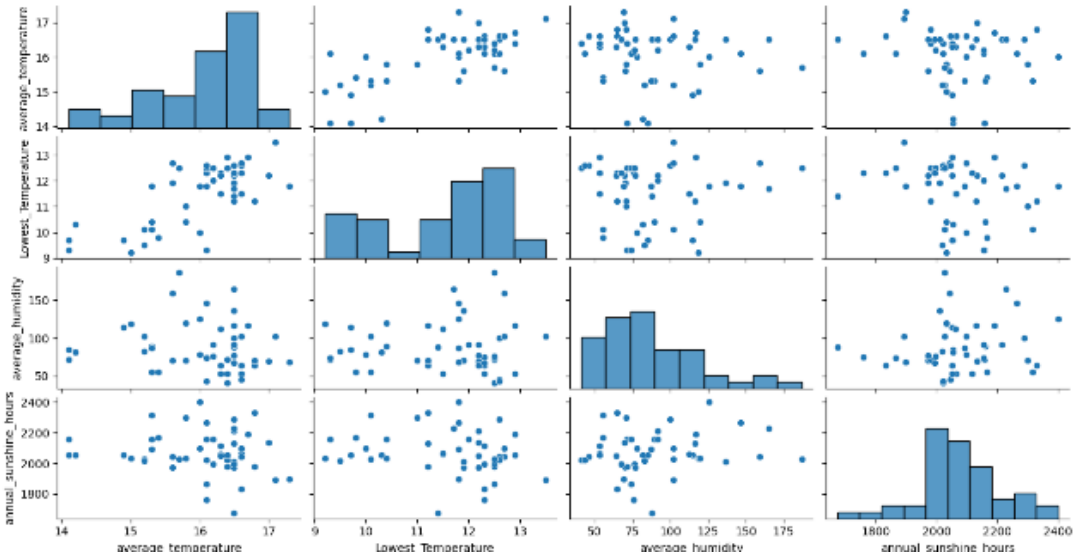
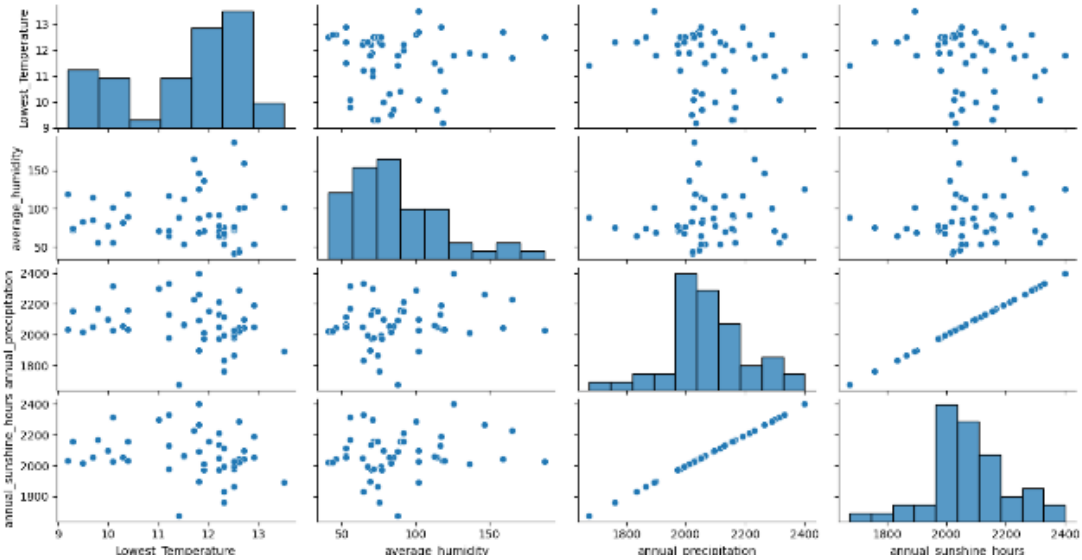


表 9　Conditions5 の散布行列図



4　考察

まず、表 1 条件 1 で、重回帰分析を行い、表 2　分析結果のデータのペアプロット表示を行い、可視化した。次に、表 3 に表 1 の条件 1 で、訓練データの比率を 1 割から 4 割まで変化させてみたが、決定係数と平均二乗誤差に変化が見られなかった。この結果より、訓練データの比率を 3 割としてそれ以降の実験を実施した。表 3 の結果は、説明変数を全て用いており、チューニングを実施していない。訓練データにおける決定係数が 0.61 であるのに対し、テストデータの決定係数は 0.43 となっており、過学習が疑われる。本研究においてモデルの作成で、過学習対策としてL2 正則を用いているが、過学習が全く起こらないということではないことがご理解いただけるのではないであろうか。また、説明変数が最大で 13 個であることによりL1 正則を用いての次元圧縮は、全く意味がないと考えられる。次に、

表 4 のConditions2 からConditions5 と説明変数を変えて、重回帰分析を実施した。その結果、Conditions2 のテストデータに対するMSEが 0.263、決定係数が 0.52 であり、説明変数の調整後としては最適であると判断した。この根拠としては、MSEが 0.263 であり最小値となり、決定係数が 0.52 で 0.5 を超えていることから妥当である。表 5 で誤差プロットを表示させ、外れ値がないか確認を実施した。この結果についてであるが、政府系データは、外れ値を削除するという親切な対応がなされているため当然の結果である。表 6 から表 9 にConditions2 からConditions5 の散布行列図を表示させた。これらの表示に異常は見受けられなかった。

5　まとめ

岡山県の気象データを、気象庁の「過去の気象データ検索」[17] をもとに抽出し、重回帰分析を実施した。目的変数を平均気温とし、説明変数を変化させながら、決定係数と、平均二乗誤差を算出し、最適なモデルを探ることを行った。結果として、本研究では、Conditions2 が最適であると判断した。過学習対策として、一般的ではあるがL2 正則を利用したものの、パラメータのチューニング段階において過学習が発生している。過学習のそもそもの定義としては、トレーニング段階では高スコアを出すが、テスト段階においては、スコアが下がるというものである。これを人間に置き換えてみれば、模擬テストでは、高得点をだすが、本番の入試で実力が発揮できない状況に例えることができるのでないであろうか。コンピュータも学習過程において、人間的な「不器用さ」が存在する。このような「不器用さ」を人間の心理や学習プロセスに適用することにより、人の新たな「改善」につながるのではないであろうか。いずれにしても、応用範囲の広さに機械学習の今後の可能性を感じる。

（注記）

1. 変数選択と正則化の両方を実行し、生成する統計モデルの予測精度と解釈可能性を向上させる回帰分析手法。
2. 独立変数が強く相関している場合に、重回帰モデルの係数を推定する方法。
3. 重回帰分析で抽出する変数の数に応じて決定係数が小さく補正されるようにしたもの。

（引用文献）

[1] Preferred Networks、 Inc.「単回帰分析と重回帰分析」、https://tutorials.chainer.org/ja/07_Regression_Analysis.html（2022/11/12 閲覧）。

[2] 関西学院大学「重回帰(説明変数が複数個)の場合」、 https://www.kwansei.ac.jp/hs/z90010/sugakuc/toukei/rp9/rp9.htm（2022/11/12 閲覧）。

[3] 立命館大学「重回帰分析による予測モデル」、 https://www.ritsumei.ac.jp/se/rv/dse/jukai/MRA.html（2022/11/12 閲覧）。

[4] 慶応義塾大学「回帰分析（重回帰）」、 http://fs1.law.keio.ac.jp/~aso/ecnm/pp/reg2.pdf（2022/11/12 閲覧）。

[5] 拓殖大学「重回帰分析 1（単回帰と重回帰）」、 http://www.ner.takushoku-u.ac.jp/masano/class_material/waseda/keiryo/R18_reg1.html（2022/11/12 閲覧）。

[6] 名古屋市立大学「MS-Excel による回帰分析」、 http://www.econ.nagoya-cu.ac.jp/~kamiyama/siryou/regress/EXCELreg.html（2022/11/12 閲覧）。
[7] 株式会社ライトコード「【機械学習】過学習を防ぐ「正則化」」、 https://rightcode.co.jp/blog/information-technology/regularization-to-prevent-overtraining（2022/11/14 閲覧）。
[8] 東京都市大学「深層学習を用いた組込み OSS の移植可能性評価に関する一考察」、 https://orsj.org/wp-content/nc-abstract/nc2021f/2021f-2-D-1.pdf（2022/11/14 閲覧）。
[9] 筑波大学「機械学習(4)特徴選択とL1 正則化」、 https://ocw.tsukuba.ac.jp/wp/wp-content/uploads/2019/10/fa67c4f9c50e67abbe3c9a684bea594c.pdf（2022/11/16 閲覧）。
[10] 東京工業大学「データ解析第六回「L1 正則化法 : 高次元データ解析」」、 http://ibis.t.u-tokyo.ac.jp/suzuki/lecture/2015/dataanalysis/L6.pdf（2022/11/16 閲覧）。
[11] 一般社団法人日本理学療法学会連合「coefficient of determination」、 https://www.jspt.or.jp/ebpt_glossary/coefficient-of-determination.html（2022/11/14 閲覧）。
[12] iStat、 Inc.「重回帰分析（２/３）」、 https://istat.co.jp/ta_commentary/multiple_0（2022/11/14 閲覧）。
[13] アイティメディア株式会社「残差平方和（RSS : Residual Sum of Squares）／［損失関数］二乗和誤差（SSE : Sum of Squared Error）とは？」、 https://atmarkit.itmedia.co.jp/ait/articles/2111/22/news011.html（2022/11/14 閲覧）。
[14] 日本アイ・ビー・エム株式会社「平均二乗誤差」、 https://www.ibm.com/docs/ja/cloud-paks/cp-data/3.5.0?topic=overview-mean-squared-error（2022/11/16 閲覧）。
[15] 東京大学船津研究室「精度評価指標と回帰モデルの評価」、 https://funatsu-lab.github.io/open-course-ware/basic-theory/accuracy-index/（2022/11/16 閲覧）。
[16] 早稲田大学「心理データ解析　第６回(2)」、http://www.f.waseda.jp/oshio.at/edu/data_b/06_folder/da06_02.html（2022/11/16 閲覧）。
[17] 気象庁 :「過去の気象データ検索」、http://www.data.jma.go.jp/obd/stats/etrn/index.php（2022/11/8閲覧）。