JUSE-StatWorks/V5 機械学習編R2 のご紹介

株式会社 日本科学技術研修所
統計ソリューション事業部
データサイエンス部
犬伏秀生

本資料の説明項目

1. StatWorks/V5機械学習編R2の概要
2. 搭載手法の概要

StatWorks/V5機械学習編R2の概要

JUSE-StatWorks/V5 機械学習編

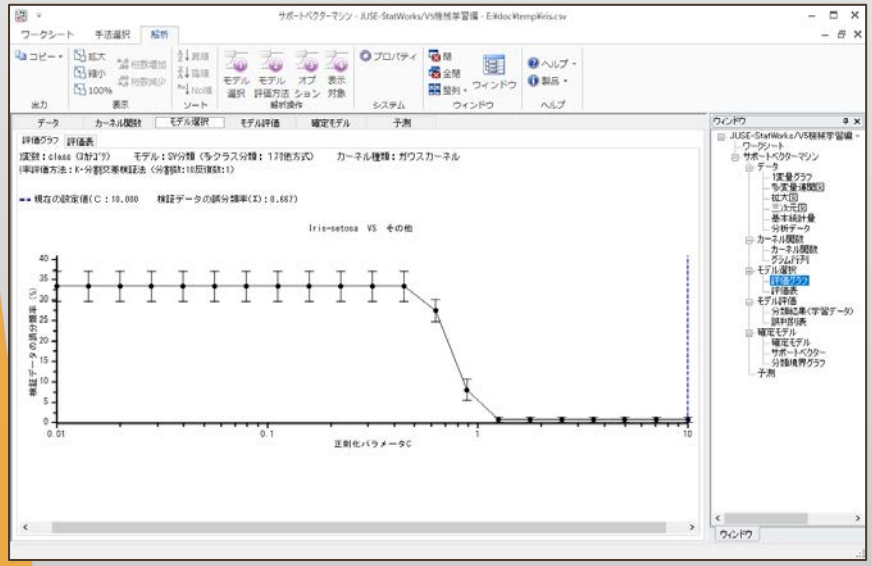
スタンドアロン版　2019年4月25日発売

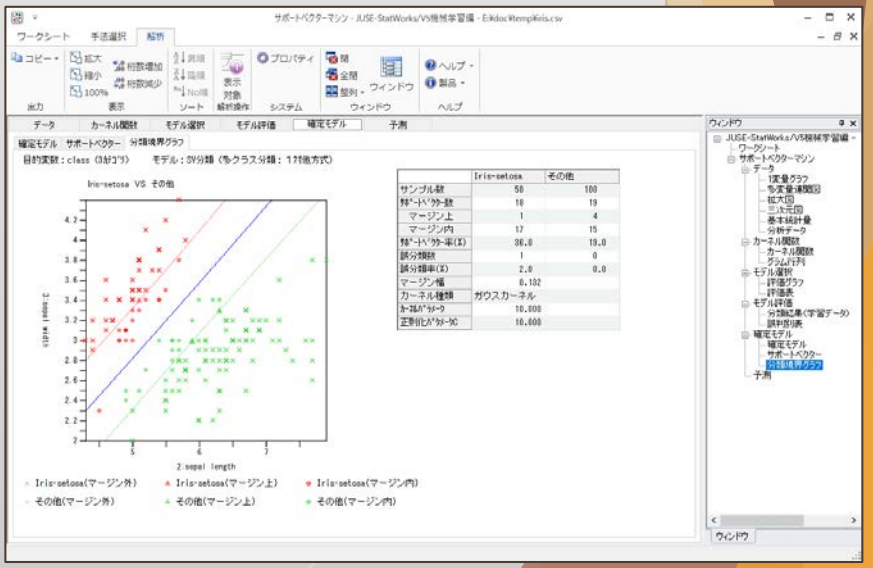
ネットワーク版　2019年6月27日発売

JUSE-StatWorks/V5 機械学習編R2

スタンドアロン版・ネットワーク版

2020年6月18日発売予定





StatWorks/V5の製品構成

製品　手法群

総合編プレミアム

SEM因果分析編

因果分析

総合編

多変量解析

時系列解析

信頼性解析

品質工学編

品質工学

品質管理手法編

実験計画法

回帰分析

QC七つ道具編

QC七つ道具

新QC七つ道具

工程分析

機械学習編R2

機械学習

機械学習編R2開発の背景

1. モノづくりの問題解決における
   データの多様性の増加
   - 情報システムの整備や工場等でのIoTの導入により，多様性に富んだデータを利用できる環境が整いつつある
   - データによっては，従来のSQC手法ではうまく対処できなかったり，手間がかかる等の問題が生じることがある．その対処の一つが機械学習の活用.
2. 現場の技術者が使い易い機械学習
   ツールに対する強いニーズ
   - TQMの理念は全員参加であり，現場の技術者が自分でデータ解析を行う
   - 既存の機械学習ツールは専門家向けが多い
   - StatWorks/V5に機械学習手法を搭載する強いニーズ

システムコンセプト

1. 現場の技術者のためのツール
   - 現場の技術者が手軽に使える機械学習ツール
   - マウスで全て操作可能
   - 単体PC上で動作（サーバー・クラウド不要）
2. 全て自社開発
   - 計算処理も自社開発
   - ユーザー要望への対応可
3. StatWorksのコンセプトを引き継ぐ
   - 豊富なグラフ機能
   - 解析ストーリーに沿ったタブ・機能構成

機能・操作性の特徴

1. 既存製品と一体で使用可能
   - 総合編等と機械学習編R2を同じPCにインストールするとメニューが統合される
2. 既存製品と同じ操作性
   - ワークシートは共通
   - 画面レイアウトも共通
3. 1,000変数(列) ×100,000サンプル(行)以内のデータを解析可能

機械学習編R2 搭載手法一覧

赤文字がR2での新規搭載手法

No	手法分類	ｸﾞﾙｰﾌﾟ	解析手法
1	データ前処理	データ前処理	データクリーニング
2			データ分割
3		データ可視化	モニタリング，濃淡散布図，密度プロット，等高線図
4			時系列グラフ
5	教師なし学習	情報要約	カーネル主成分分析
6		層別	混合ガウス分布
7		外れ値検出	1クラスSVM
8		因果分析	glasso
9	教師あり学習	正則化回帰	リッジ回帰，lasso回帰，Elastic Net
10			正則化ロジスティック回帰
11		分類・予測	k-近傍法
12			ｻﾎﾟｰﾄﾍﾞｸﾀﾏｼﾝ(SVM)
13			ランダム・フォレスト
14	その他	モデル評価	予測判定グラフ，誤判別表

搭載手法の特徴

1. モノづくりの問題解決で有効
   - 従来のSQC手法ではうまく対処できなかった状況で特に有効
   - 変数の数がサンプル数よりも多い状況（多重共線性が発生している状況）やデータの分布が正規分布から大きく乖離している状況などにも対応
2. 要因の検討でも使用可
   - 変数選択が自動的に行われる，重要度が出力される　など
3. 多変量解析の拡張として理解可能

搭載手法と多変量解析手法との対応

No	機械学習手法(A)	対応する多変量解析手法(B)	B→Aの主なメリット
1	ｶｰﾈﾙ主成分分析	主成分分析	複雑な構造のデータに対応
2	混合ガウス分布	非階層的クラスター分析(k-means)	所属確率の取得
3	リッジ回帰，lasso回帰，Elastic Net	重回帰分析・数量化Ⅰ類	変数の数>サンプル数，の場合（多重共線性）に対応
4	正則化ﾛｼﾞｽﾃｨｯｸ回帰	ﾛｼﾞｽﾃｨｯｸ回帰分析	
5	ｻﾎﾟｰﾄﾍﾞｸﾀﾏｼﾝ(SVM)	判別分析・数量化Ⅱ類	汎化能力（新たなデータに対する予測能力）の向上
6	ﾗﾝﾀﾞﾑ・ﾌｫﾚｽﾄ	多段層別分析(AID)	
7	1クラスSVM	MT法	正規分布に従わないデータに対応
8	glasso	ｸﾞﾗﾌｨｶﾙﾓﾃﾞﾘﾝｸﾞ(GM)	変数の数>サンプル数，の場合（多重共線性）に対応

有効な活用場面

1. 中規模データの分析

- R，Pythonなどに比べ，手軽に分析を実行可能
- 1,000変数×100,000サンプル以内のデータを分析でき，モノづくりの問題解決で使用される数値データの多くをカバー

2. 機械学習手法の教育

- 教育を行うための環境構築が容易
- マウスで操作できるので演習を自由に実施可能
- トヨタグループ様の機械学習セミナー，日科技連の機械学習セミナー（2020年度開講）でも使用

3. ビッグデータの分析モデルの検討

- ビッグデータから抽出したデータを分析
- 分析の試行錯誤を行い易い

稼働環境

【スタンドアロン版・ネットワーク版クライアント】

製品	OS	ハードウェア
機械学習編R2	【64bit版】 Windows 10	【CPU】Intel Core i5-4xxx以上 【メモリ】8GB以上 【HDD】1GB以上の空き容量 【ディスプレイ】1024×768(XGA)以上 【ディスク装置】DVD-ROMドライブ

【ネットワーク版サーバー】

製品	OS	ハードウェア
機械学習編R2 を含む全製品	【64bit版】 Windows Server 2016 Windows Server 2012 R2 【32bit版，64bit版】 Windows Server 2008 R2	【CPU】1GHz以上(x86), 1.4GHz以上(x64) 【メモリ】2GB以上(x86，x64) 【HDD】 約130MB以上の空き容量

製品価格

【スタンドアロン版】

製品	標準価格(税別)
機械学習編R2	270,000円
総合編＋機械学習編R2セット	394,000円
総合編	168,000円

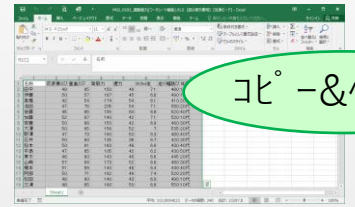
【ネットワーク版】

※同時起動数・インストール台数は5単位で増減

製品	同時起動数	インストール台数	標準価格(税別)
総合編＋機械学習編R2セット	5	25	3,380,000円
総合編	5	25	1,450,000円

※製品価格は弊社ホームページ上でご確認いただける予定です.

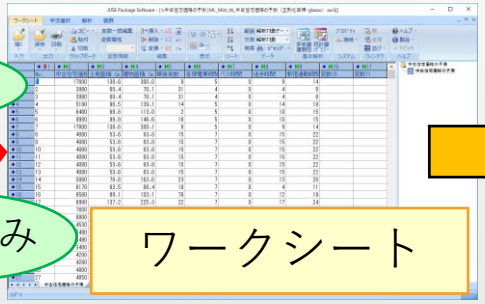
基本的な分析手順(lasso回帰)



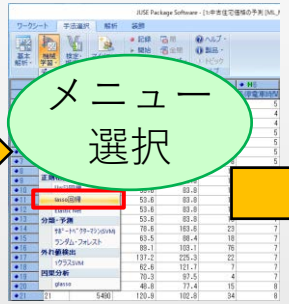
コピー&ペースト

データ
ファイル

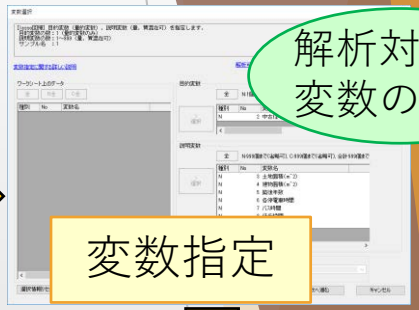
読み込み



ワークシート

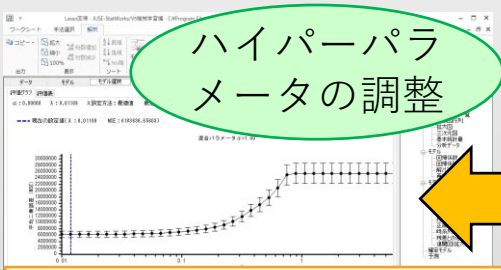


メニュー
選択



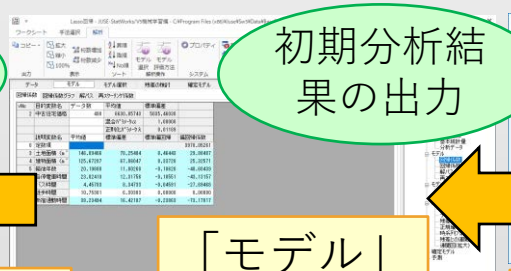
変数指定

解析対象の
変数の指定



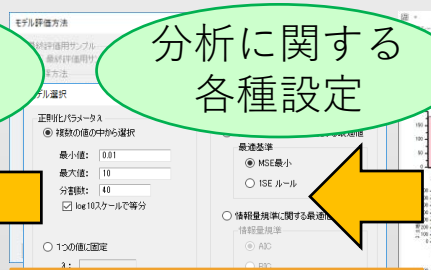
ハイパーパラ
メータの調整

「モデル選択」グループ



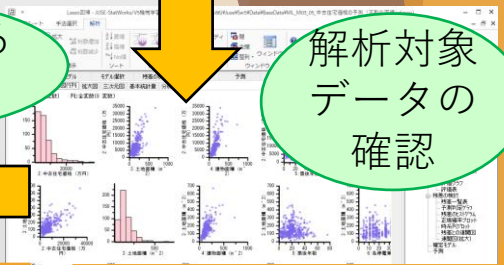
初期分析結
果の出力

「モデル」
グループ



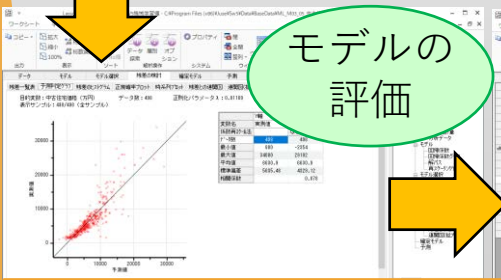
分析に関する
各種設定

設定ダイアログ



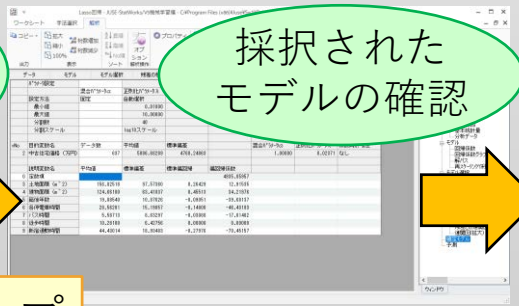
解析対象
データの
確認

「データ」グループ

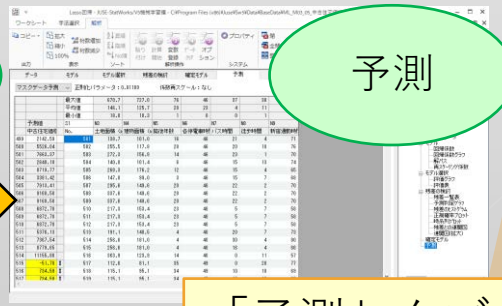


モデルの
評価

「残差の検討」グループ



採択された
モデルの確認



予測

「予測」タブ

搭載手法の概要

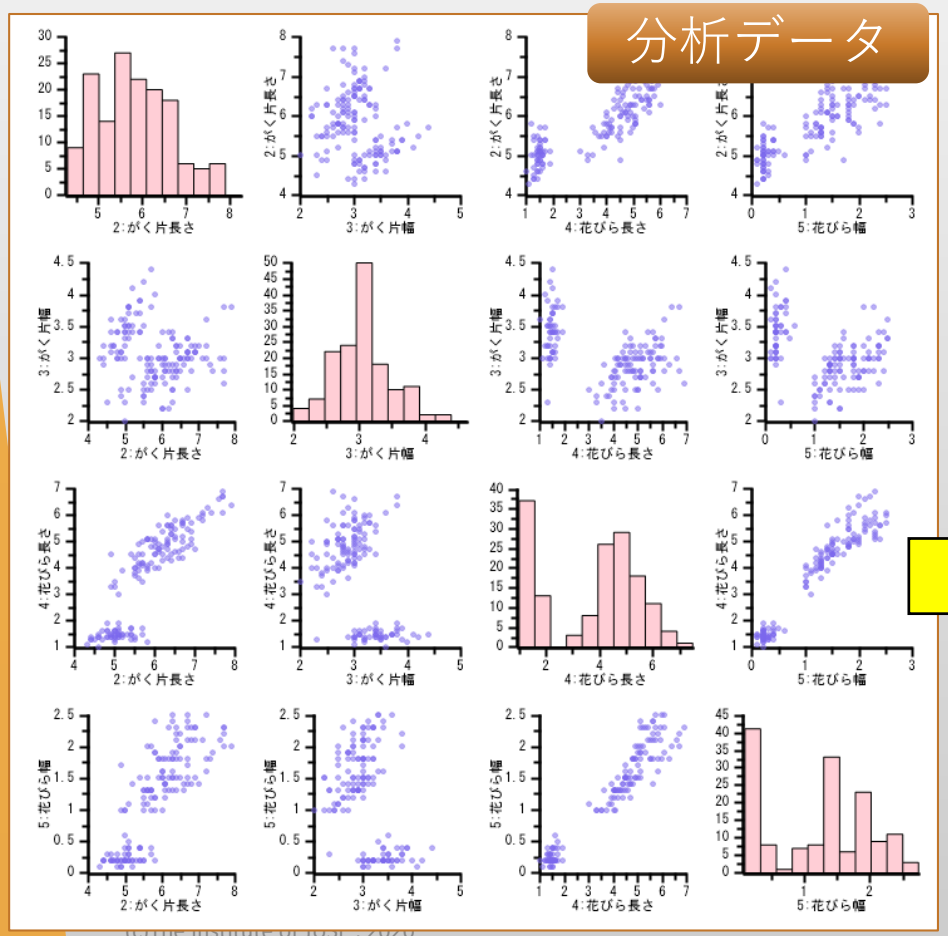
機械学習編R2 搭載手法一覧（再掲）

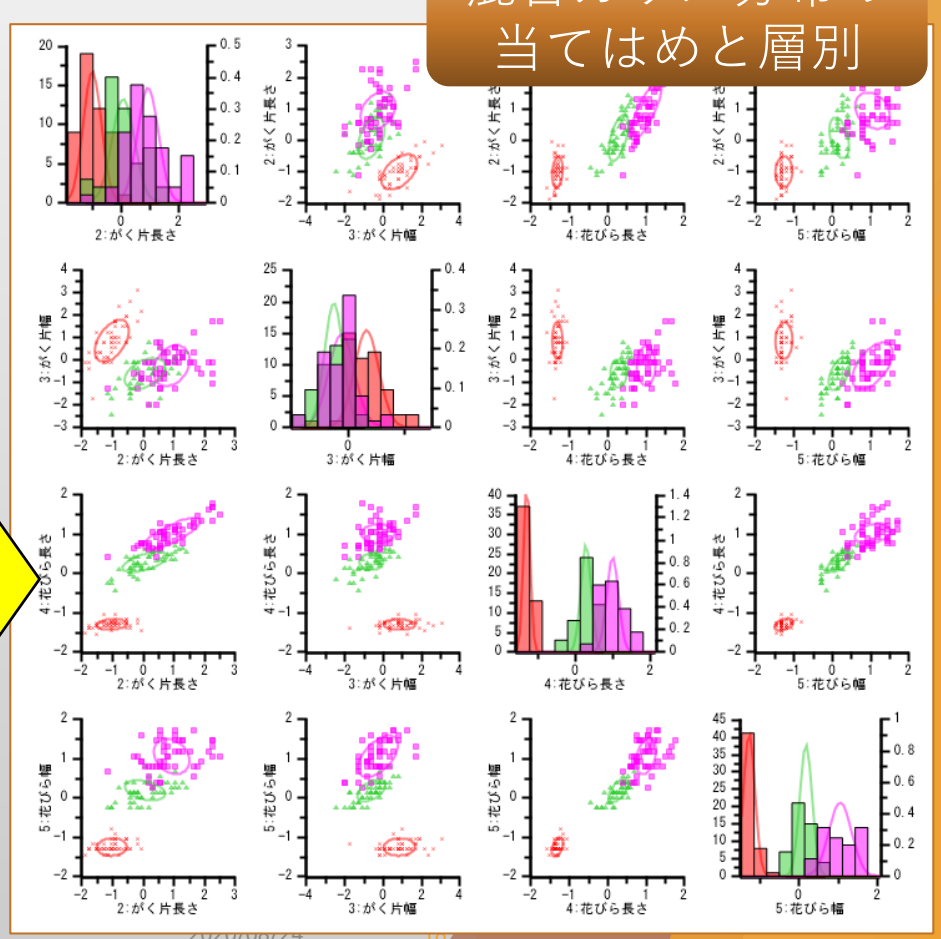
赤文字がR2での新規搭載手法

No	手法分類	ｸﾞﾙｰﾌﾟ	解析手法
1	データ前処理	データ前処理	データクリーニング
2			データ分割
3		データ可視化	モニタリング，濃淡散布図，密度プロット，等高線図
4			時系列グラフ
5	教師なし学習	情報要約	カーネル主成分分析
6		層別	混合ガウス分布
7		外れ値検出	1クラスSVM
8		因果分析	glasso
9	教師あり学習	正則化回帰	リッジ回帰，lasso回帰，Elastic Net
10			正則化ロジスティック回帰
11		分類・予測	k-近傍法
12			ｻﾎﾟｰﾄﾍﾞｸﾀﾏｼﾝ(SVM)
13			ランダム・フォレスト
14	その他	モデル評価	予測判定グラフ，誤判別表

混合ガウス分布

- 解析手法「混合ガウス分布」では，データの分布を複数の正規分布の重ね合わせで近似し，重ね合わせた正規分布に基づいてクラスタリングを行うことができます．





正則化回帰

- 手法群「正則化回帰」では，目的変数（量的変数）を複数の説明変数（量的変数・質的変数）から予測するための式を得ることができます．

データ | モデル | モデル選択 | 残差の検討 | 確定モデル

回帰係数 | 回帰係数グラフ | 解パス | 再スケーリング係数

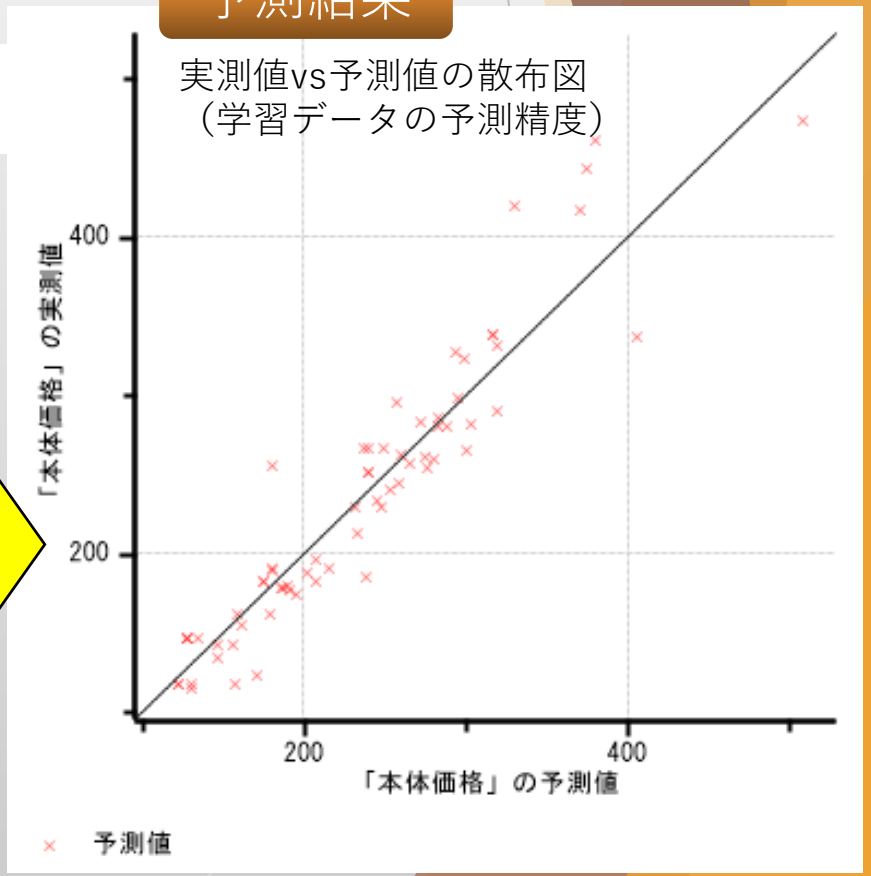
予測式

lasso回帰による予測式

vNo	目的変数名	データ数	平均値	標準偏差	
18	本体価格	67	235.36119	86.35089	
			混合ﾊﾟﾗﾒｰﾀα	1.00000	
			正則化ﾊﾟﾗﾒｰﾀλ	0.03981	
	説明変数名	平均値	標準偏差	標準偏回帰	偏回帰係数
0	定数項				-447.40270
2	全長	4425.29851	388.49776	0.00000	0.00000
3	全幅	1756.94030	70.76319	0.16440	0.20062
4	全高	1621.04478	159.99659	0.00000	0.00000
5	室内長	2220.89552	488.35692	0.00000	0.00000
6	室内幅	1487.98507	62.81969	0.01202	0.01652
7	室内高	1262.08955	80.39153	0.00000	0.00000
8	ﾎｲｰﾙﾍﾞｰｽ	2679.62687	159.07924	0.09412	0.05109
9	最低地上高	155.14925	23.89295	0.00000	0.00000
10	車両重量	1375.82090	313.60280	0.05992	0.01650
11	総排気量	1758.11940	577.72631	0.00000	0.00000
12	燃料ﾀﾝｸ容量	53.38806	13.11483	0.00000	0.00000
13	最高出力	143.85075	51.80827	0.15524	0.25875
14	最大トルク	19.32090	7.56497	0.49372	5.63562
15	ﾊﾟﾜｰｳｪｲﾄﾚｼｵ	10.14730	2.07913	0.00000	0.00000
16	最小回転半径	5.22090	0.35641	0.00000	0.00000
17	燃費	17.05075	4.18057	0.00000	0.00000

予測結果

実測値vs予測値の散布図
（学習データの予測精度）



正則化ロジスティック回帰

- 解析手法「正則化ロジスティック回帰」では，目的変数（質的変数）を複数の説明変数（量的変数・質的変数）から予測するための式を得ることができます．

予測式

データ | モデル | モデル選択 | モデル評価

回帰係数 | 回帰係数グラフ | 解パス

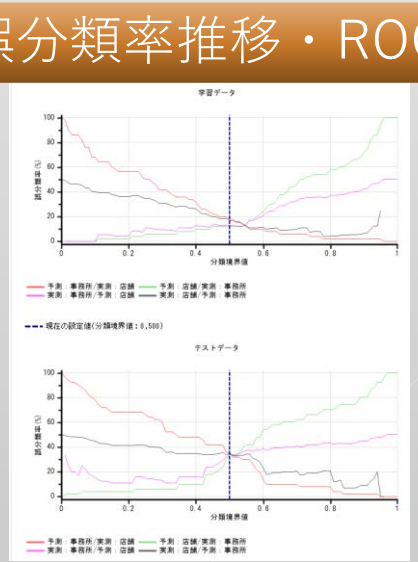
分析データ：規準化データ　モデル：2項ロジスティック回帰

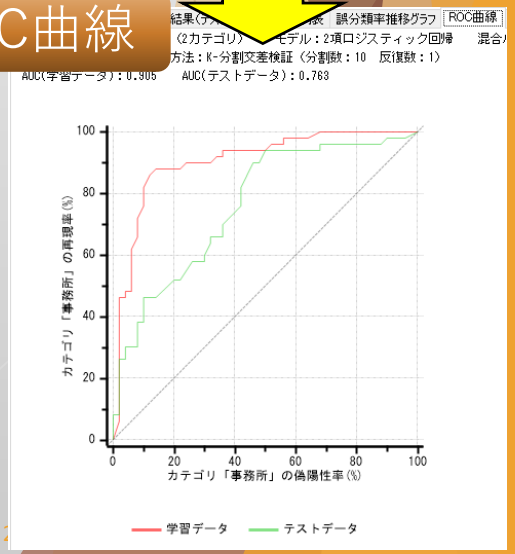
vNo	目的変数名	データ数		店舗	事務所	
2	今後の利用目的	100		50	50	
	混合ﾊﾟﾗﾒｰﾀα	0.80000				
	正則化ﾊﾟﾗﾒｰﾀλ	0.01000			事務所	
	説明変数名	カテゴリ名	平均	標準偏差	標準偏回帰	偏回帰係数
0	定数項				-0.09788	4.27727
3	地域	住宅地	0.43000	0.49508	0.08305	0.16776
		商業地	0.56000	0.49639	-0.76507	-1.54128
		工業地	0.01000	0.09950	0.17479	1.75667
4	都道府県名	北海道	0.15000	0.35707	-0.19742	-0.55288
		東京都	0.35000	0.47697	0.00000	0.00000
		愛知県	0.21000	0.40731	0.56191	1.37956
		大阪府	0.29000	0.45376	-0.09101	-0.20057
5	最寄駅：距離（分）		12.15000	11.92340	0.00000	0.00000
6	取引価格（100万円）		179.94300	334.93822	-0.53634	-0.00160
7	面積（m^2）		307.45000	415.88490	-0.47967	-0.00115
9	間口		13.34500	9.70349	0.00000	0.00000
10	延床面積（m^2）		368.35000	381.97713	0.59352	0.00155
12	前面道路：種類	市道	0.48000	0.49960	0.00000	0.00000
		道道	0.03000	0.17059	0.00000	0.00000
		国道	0.06000	0.23749	0.00000	0.00000
		区道	0.24000	0.42708	0.63461	1.48593
		私道	0.06000	0.23749	-0.80231	-3.37833
		都道	0.04000	0.19596	-0.43154	-2.20221
		県道	0.07000	0.25515	0.00000	0.00000
		府道	0.02000	0.14000	-0.34611	-2.47224
13	前面道路：幅員（m）		13.24600	9.11905	-0.36775	-0.04033
14	建ぺい率（%）		71.20000	10.60943	-0.47119	-0.04441
15	容積率（%）		341.40000	170.57561	0.00000	0.00000

誤判別表

			予測結果 店舗	事務所	合計	誤分類数	誤分類率(%)	再現率(%)
学習データ	実測結果	店舗	41	9	50	9	18.000	82.000
		事務所	6	44	50	6	12.000	88.000
		合計	47	53	100			
		誤分類数	6	9		15		
		誤分類率(%)	12.766	16.981			15.000	
		適合率(%)	87.234	83.019				85.000
テストデータ	実測結果	店舗	32	18	50	18	36.000	64.000
		事務所	17	33	50	17	34.000	66.000
		合計	49	51	100			
		誤分類数	17	18		35		
		誤分類率(%)	34.694	35.294			35.000	
		適合率(%)	65.306	64.706				65.000

誤分類率推移・ROC曲線

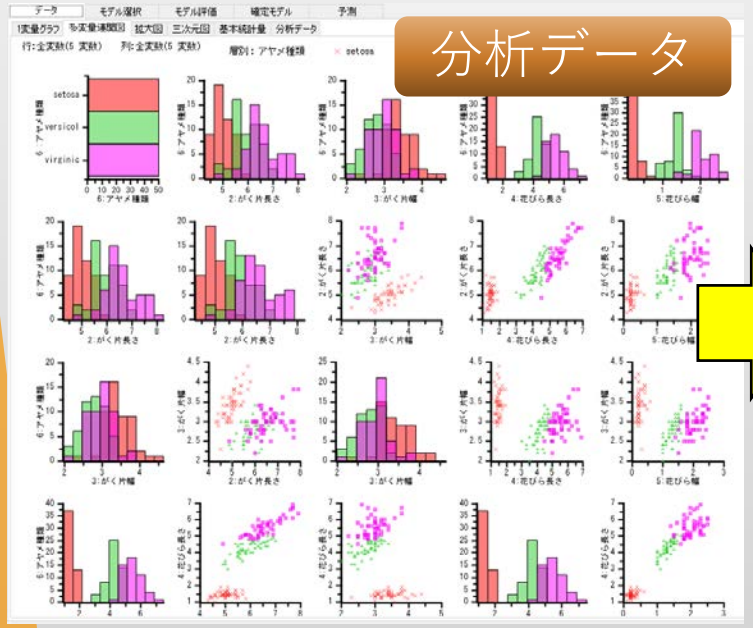




k-近傍法

- 解析手法「k-近傍法」では，各サンプルの質的な目的変数をそのサンプルの近傍の情報から予測することができます．
- k-近傍法は最もシンプルな分類手法であり，他の分類手法のベンチマークとしても利用できます．

分析データ



分析設定・予測結果

データ　モデル選択　モデル評価　確定モデル　予測

予測結果(学習データ)　誤判別表

近傍の個数k:3　距離:ユークリッド距離　予測方法:単純多数決

全サンプル

		データ数	150							
		誤分類数	8							
		誤分類率(%)	5.3		カテゴリ比率			第1近傍点		
sNo	サンプル名	実測カテゴリ	予測カテゴリ	正誤	setosa	versicolor	virginica	sNo	分類	距離
1	1	setosa	setosa	正	1.000	0.000	0.000	18	setosa	0.132
2	2	setosa	setosa	正	1.000	0.				
3	3	setosa	setosa	正	1.000	0.				
4	4	setosa	setosa	正	1.000	0.				
5	5	setosa	setosa	正	1.000	0.				
6	6	setosa	setosa	正	1.000	0.				
7	7	setosa	setosa	正	1.000	0.				
8	8	setosa	setosa	正	1.000	0.				
9	9	setosa	setosa	正	1.000	0.				
10	10	setosa	setosa	正	1.000	0.				
11	11	setosa	setosa	正	1.000	0.				
12	12	setosa	setosa	正	1.000	0.				
13	13	setosa	setosa	正	1.000	0.				
14	14	setosa	setosa	正	1	0.				
15	15	setosa	setosa	正	1	0.				

モデル選択

近傍の個数k
- 複数の値の中から選択
  - 最小値: 1
  - 最大値: 5
- 1つの値に固定
  - k: 3

距離
- ☑ ユークリッド距離
- ☐ マハラノビス距離
- ☐ コサイン距離

予測方法
- ☑ 近傍k個の単純多数決
- ☐ 近傍k個の重み付き多数決（重み:距離の逆数）

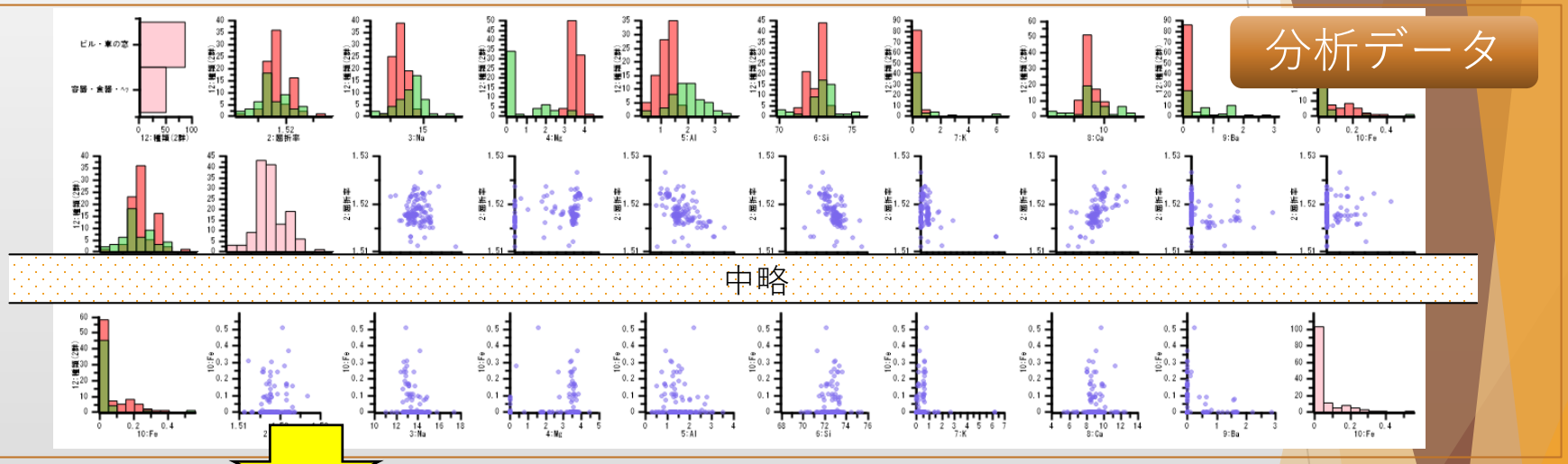
OK　キャンセル　ヘルプ

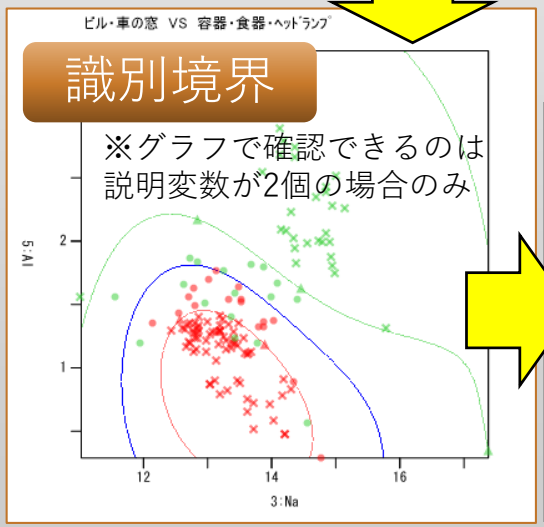
誤判別表

			予測結果						
			setosa	versicolor	virginica	合計	誤分類数	誤分類率(%)	再現率(%)
学習データ	実測結果	setosa	49	1	0	50	1	2.000	98.000
		versicolor	0	47	3	50	3	6.000	94.000
		virginica	0	4	46	50	4	8.000	92.000
		合計	49	52	49	150			
		誤分類数	0	5	3		8		
		誤分類率(%)	0.000	9.615	6.122			5.333	
		適合率(%)	100.000	90.385	93.878				94.667

ｻﾎﾟｰﾄﾍﾞｸﾀｰﾏｼﾝ(SVM)

- 解析手法「ｻﾎﾟｰﾄﾍﾞｸﾀｰﾏｼﾝ(SVM)」では，目的変数(質的変数)を複数の説明変数（量的変数）から予測することができます．





誤判別表

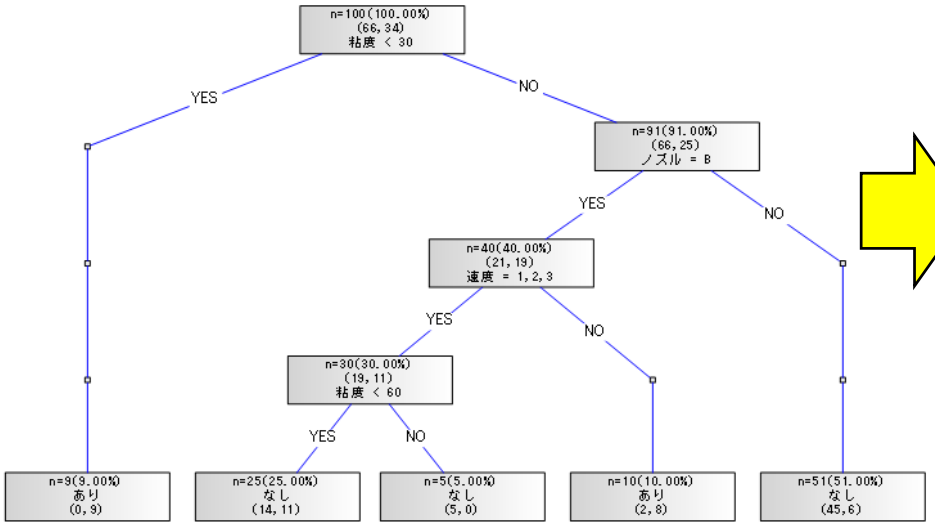
			正答					
			ビル・車の窓	容器・食器・ﾍｯﾄﾞﾗﾝﾌﾟ	合計	誤分類数	誤分類率(%)	適合率(%)
学習データ	予測結果	ビル・車の窓	87	1	88	1	1.136	98.864
		容器・食器・ﾍｯﾄﾞﾗﾝﾌﾟ	0	50	50	0	0.000	100.000
		合計	87	51	138			
		誤分類数	0	1		1		
		誤分類率(%)	0.000	1.961			0.725	
		再現率(%)	100.000	98.039				99.275
テストデータ	予測結果	ビル・車の窓	87	1	88	1	1.136	98.864
		容器・食器・ﾍｯﾄﾞﾗﾝﾌﾟ	0	50	50	0	0.000	100.000
		合計	87	51	138			
		誤分類数	0	1		1		
		誤分類率(%)	0.000	1.961			0.725	
		再現率(%)	100.000	98.039				99.275

ランダム・フォレスト

- 解析手法「ランダム・フォレスト」では，目的変数（量的変数・質的変数）を複数の説明変数（量的変数・質的変数）から予測することができます．
- 　また，手法群「ランダム・フォレスト」では決定木の分析結果も確認することができます．

決定木

※本出力例の目的変数（質的変数）は「塗装たれ不良（あり/なし）」



ランダムフォレストに対する正誤表

木の種類： 分類木　木の本数： 100　使う変数の数： 1　木の最大深さ： 10
学習データの正答率： 87.000　OOBデータの誤り率： 29.000　剪定の有無： 無

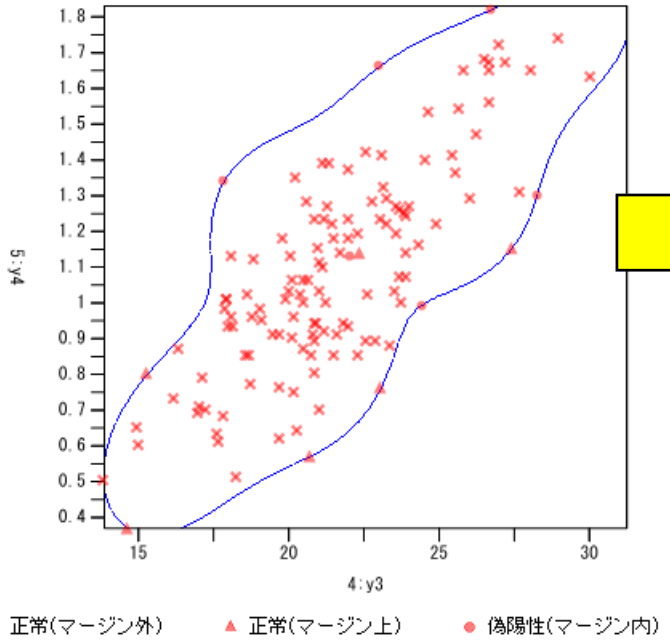
			正答					
			なし	あり	合計	誤分類数	誤分類率(%)	適合率(%)
学習データ	予測結果	なし	66	13	79	13	16.456	83.544
		あり	0	21	21	0	0.000	100.000
		合計	66	34	100			
		誤分類数	0	13		13		
		誤分類率(%)	0.000	38.235			13.000	
		再現率(%)	100.000	61.765				87.000
OOBデータ	予測結果	なし	60	23	83	23	27.711	72.289
		あり	6	11	17	6	35.294	64.706
		合計	66	34	100			
		誤分類数	6	23		29		
		誤分類率(%)	9.091	67.647			29.000	
		再現率(%)	90.909	32.353				71.000
テストデータ	予測結果	なし	59	19	78	19	24.359	75.641
		あり	7	15	22	7	31.818	68.182
		合計	66	34	100			
		誤分類数	7	19		26		
		誤分類率(%)	10.606	55.882			26.000	
		再現率(%)	89.394	44.118				74.000

1クラスSVM

- 解析手法「1クラスSVM」では分析対象データ（学習データ）と異なる傾向を持つサンプルを検出することができます.
- 解析手法「1クラスSVM」の分析対象データには，目的変数（教師）は含みません.

識別境界

※グラフで確認できるのは説明変数が2個の場合のみ



予測結果

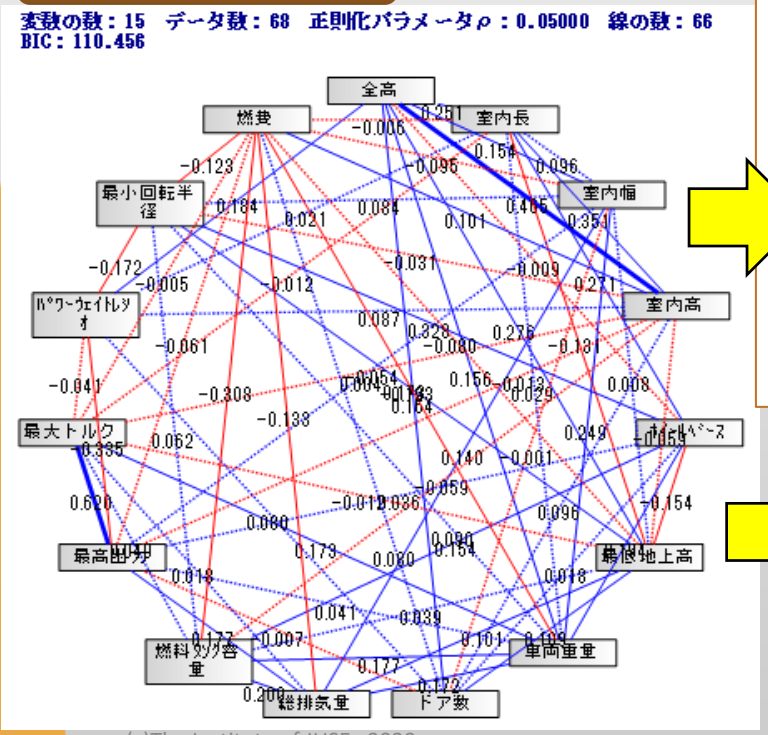
新たなデータに対する予測結果

				最大値	179.8	17
				平均値	95.7	10
				最小値	14.9	3
	目的変数名			S1	N2	N3
	予測カテゴリ	識別関数		サンプル名	y1	y2
136	正常	0.00	#	base-paper136	166.7	15
137	正常	0.01		base-paper137	72.5	8
138	異常	0.00		ex-paper4	60.0	7
139	異常	0.00	#	ex-paper5	98.7	10

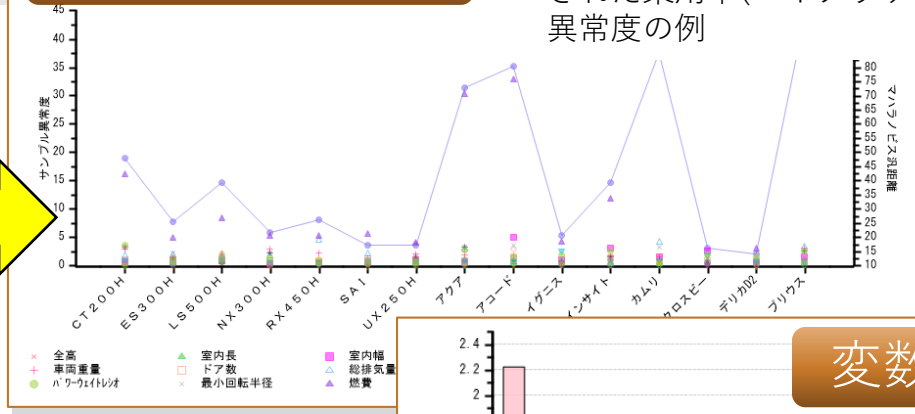
glasso

- 解析手法「glasso」では，変数間の関連(モデル)を視覚的に確認することができます．また，分析対象データ（学習データ）に対して得られた変数間の関連(モデル)を基に，新たなデータに対する「サンプル異常度」や「変数異常度」を確認することもできます．

偏相関グラフ

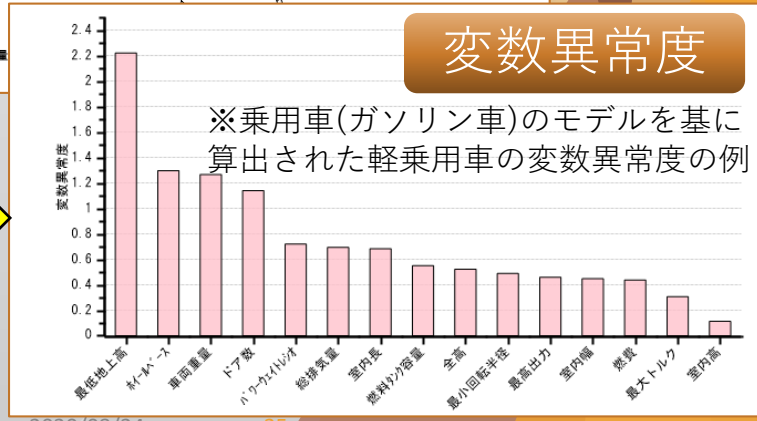


サンプル異常度



※乗用車(ガソリン車)のモデルを基に算出された乗用車(ハイブリッド車)のサンプル異常度の例

変数異常度



※乗用車(ガソリン車)のモデルを基に算出された軽乗用車の変数異常度の例

■各種導入事例資料ページ　https://www.i-juse.co.jp/statistics/jirei/

■お問い合わせフォーム　https:/www.i-juse.co.jp/statistics/support/contact.html