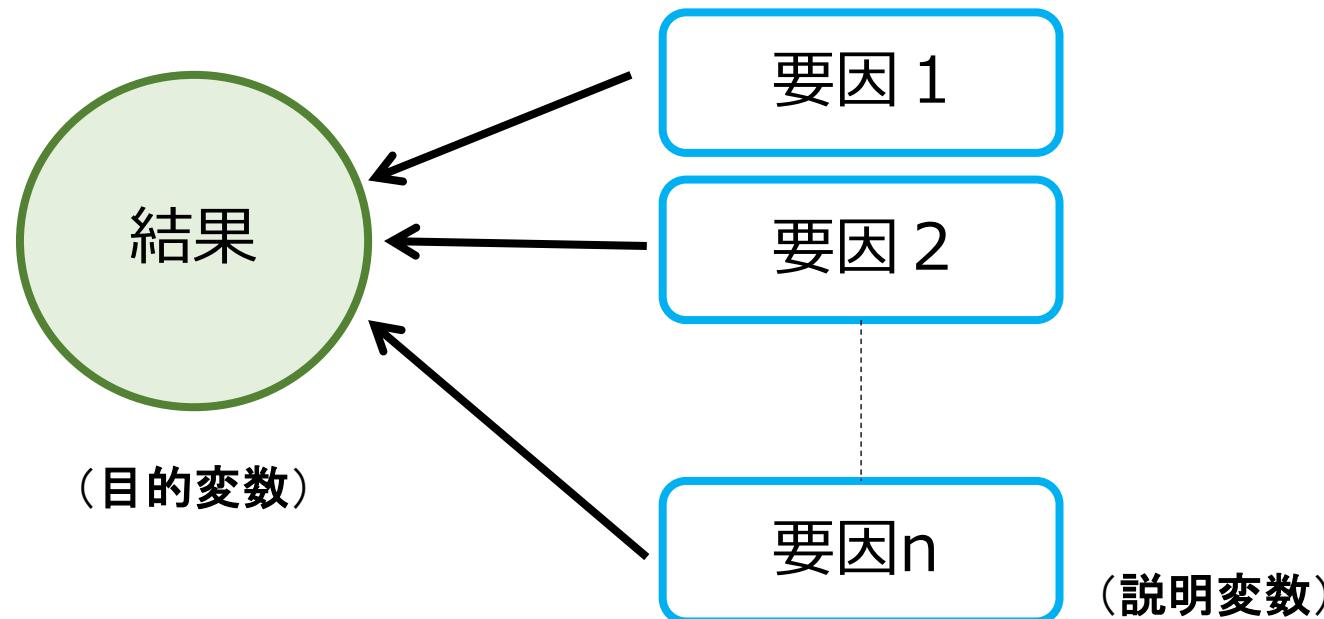


門田塾

-回帰分析応用事例編-

重回帰分析モデル

予測したい変数をいくつかの要因となる変数の関係から予測・説明する方法

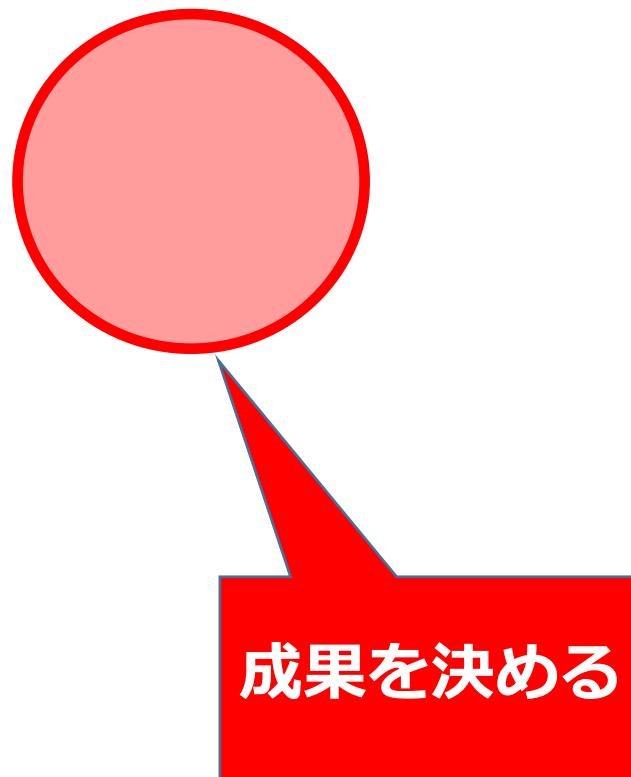


$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

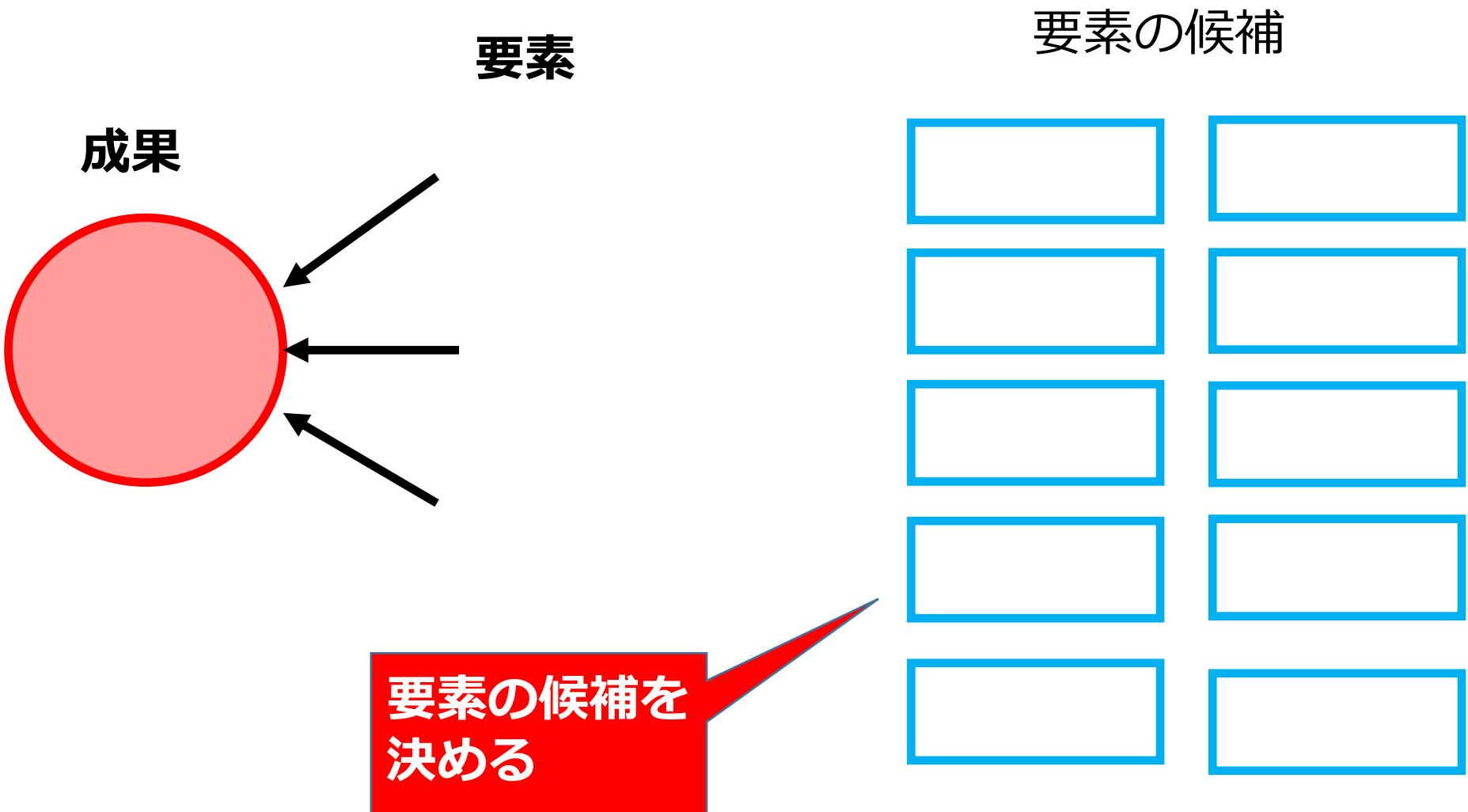
回帰分析（手順 1：成果を選ぶ）

課題を明確に認識すること

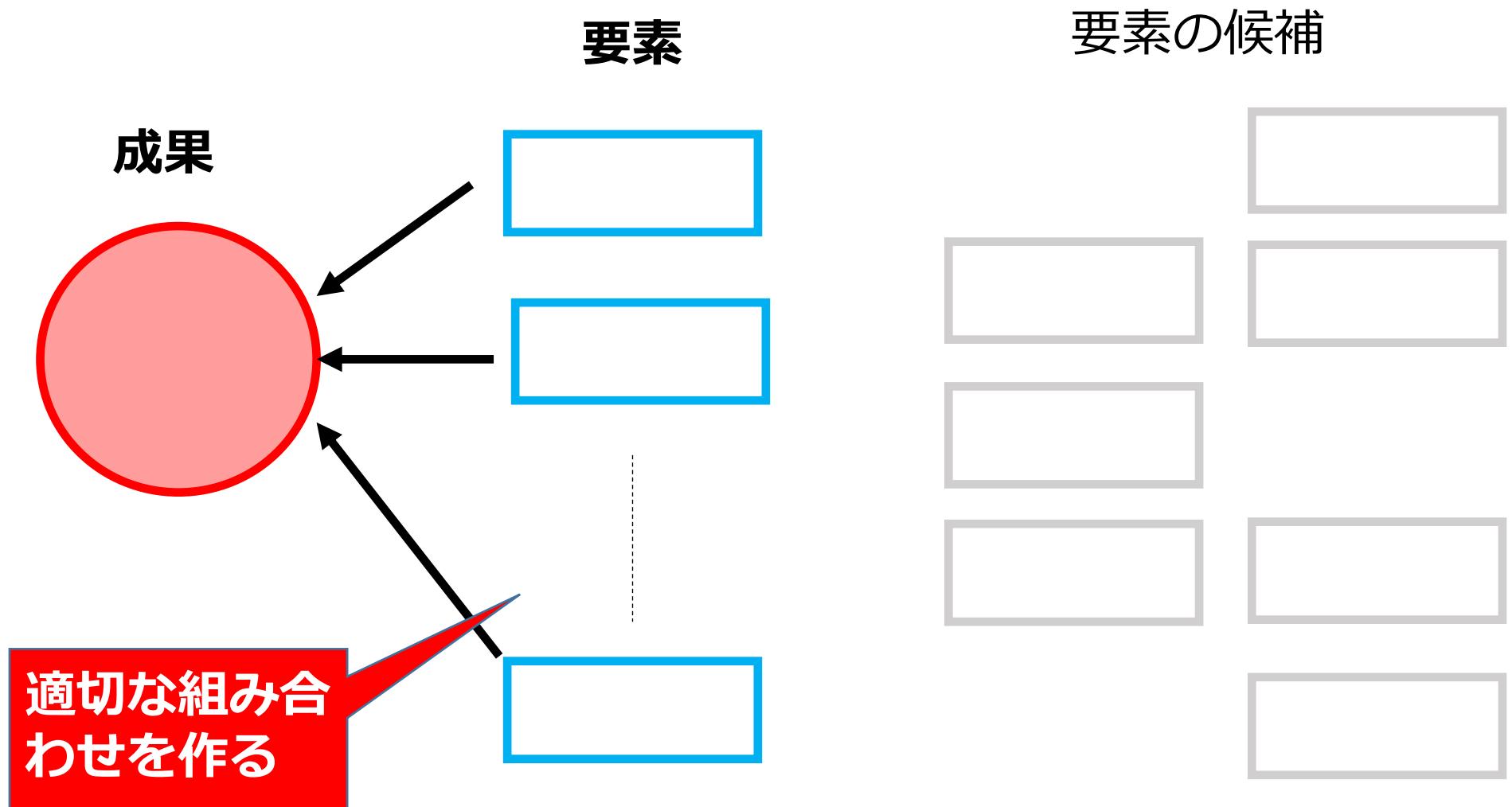
成果



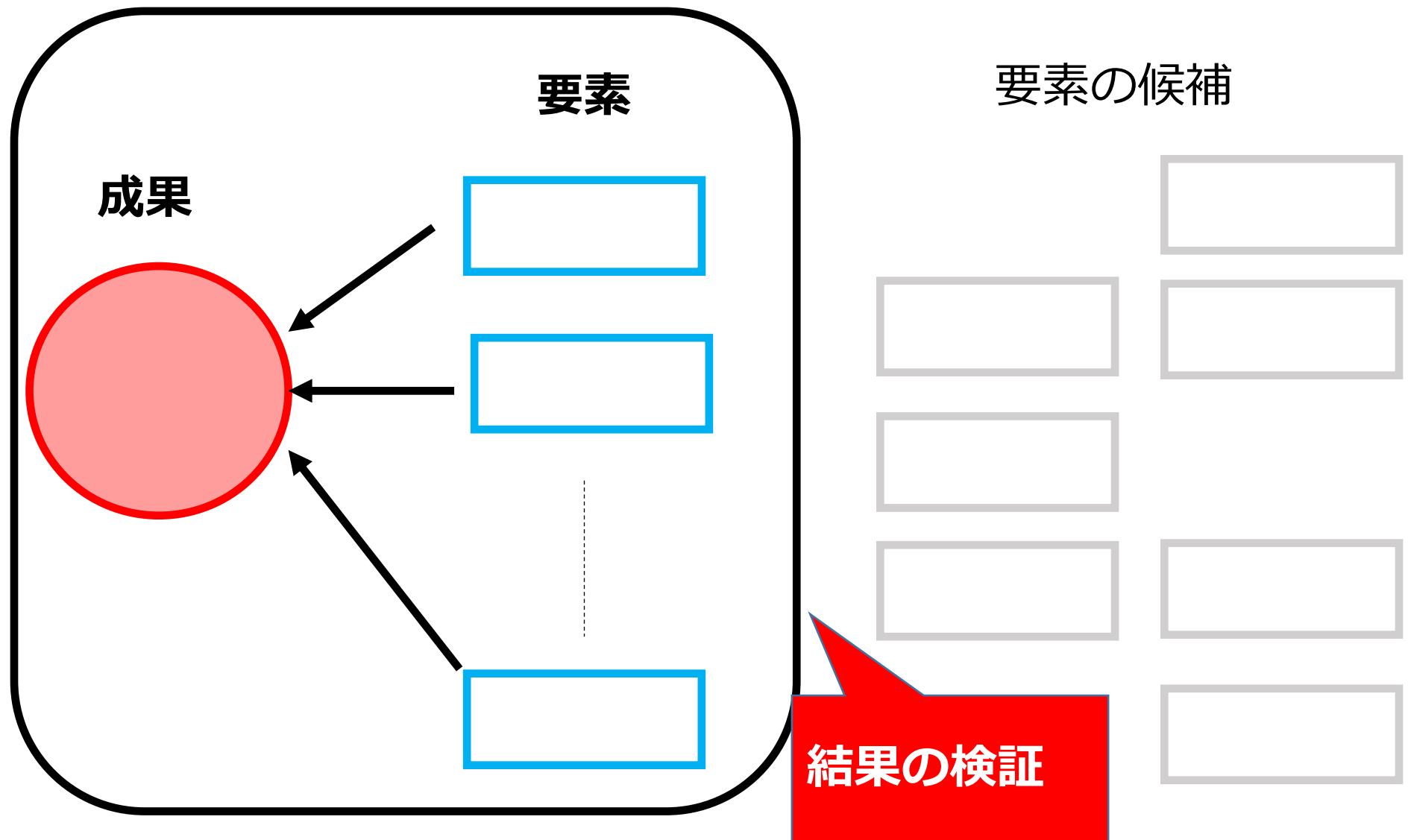
回帰分析（手順2：要素をリストアップ）



回帰分析（手順3：要素の組み合わせ）



回帰分析（手順4：結果の検証）



重回帰分析のデータ

店舗	売上	店舗面積	アクセス数	広告	街の規模	同業者店舗数
1	231	3	294	8.2	8.2	11
2	156	2.2	232	6.9	4.1	12
3	10	0.5	149	3	4.3	15
4	519	5.5	260	12	16.1	1
5	437	4.4	567	10.6	14.1	5
6	487	4.8	237	11.8	12.7	4
7	299	3.1	451	8.1	10.1	10
27	400	8.6	335	7	12	8

成果

要因

重回帰分析のデータ

概要

回帰統計	
重相関 R	0.995085404
重決定 R ²	0.990194961
補正 R ²	0.987860428
標準誤差	21.16130807
観測数	27

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	5	949676.5317	189935.3063	424.1511825	2.4296E-20
残差	21	9403.820141	447.8009591		
合計	26	959080.3519			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	0.296588834	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	3.61064E-05	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	0.979175474	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	9.6594E-08	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	9.15653E-08	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	0.055412839	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R2	0.990194961
補正 R2	0.987860428
標準誤差	21.16130807
観測数	27

帰無仮説 : $b_i = 0$
 対立仮説 : $b_i \neq 0$

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	5	949676.5317	189935.3063	424.1511825	2.4296E-20
残差	21	9403.820141	447.8009591		
合計	26	959080.3519			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	29.66%	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R2	0.990194961
補正 R2	0.987860428
標準誤差	21.16130807
観測数	27

帰無仮説 : $b_1 = 0$
 対立仮説 : $b_1 \neq 0$

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	5	949676.5317	189935.3063	424.1511825	2.4296E-20
残差	21	9403.820141	447.8009591		
合計	26	959080.3519			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	29.66%	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R2	0.990194961
補正 R2	0.987860428
標準誤差	21.16130807
観測数	27

帰無仮説 : $b_1 = 0 \longrightarrow$ 面積が売上に影響なし
 対立仮説 : $b_1 \neq 0 \longrightarrow$ 面積が売上に影響あり

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	5	949676.5317	189935.3063	424.1511825	2.4296E-20
残差	21	9403.820141	447.8009591		
合計	26	959080.3519			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	29.66%	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R2	0.990194961
補正 R2	0.987860428
標準誤差	21.16130807
観測数	27

帰無仮説 : $b_1 = 0$
 対立仮説 : $b_1 \neq 0$

面積が売上に影響なし
 面積が売上に影響あり

分散分析表

	自由度	差	平方和	F値
回帰	5	94967.820141	189935.6402	424.15118
残差	21	9403.820141	447.8009591	
合計	26	959080.3519		

面積の回帰係数 ≠ 0

P値が優位水準 α 以下の場合、
 対立仮説を採択

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	29.66%	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R ²	0.990194961
補正 R ²	0.987860428
標準誤差	21.16130807
観測数	27

分散分析表

	自由度	平方和	平均平方和	F値
回帰	5	94967.8317	18993.5634	424.15118
残差	21	9403.820141	447.8009591	
合計	26	959080.3519		

面積の回帰係数 ≠ 0

帰無仮説 : $b_1 = 0$ → 面積が売上に影響なし
 対立仮説 : $b_1 \neq 0$ → 面積が売上に影響あり

P値が優位水準 α 以下の場合、
 対立仮説を採択

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-79.1953195	36.61749065	-1.07039884	29.66%	-115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

概要

$$(売上) = b_0 + b_1(\text{面積}) + b_2(\text{アクセス}) + b_3(\text{広告}) + b_4(\text{規模}) + b_5(\text{他者数})$$

回帰統計

重相関 R	0.995085404
重決定 R ²	0.990194961
補正 R ²	0.987860428
標準誤差	21.16130807
観測数	27

分散分析表

	自由度	アクセス数の回帰係数 = 0		
回帰	5	949676.5317	189935.3063	424.15118
残差	21	9403.820141	447.8009591	
合計	26	959080.3519		

帰無仮説 : $b_2 = 0$
 対立仮説 : $b_2 \neq 0$

アクセス数が売上に影響なし
 アクセス数が売上に影響あり

P値が優位水準 α 以上の場合、
 帰無仮説仮説を採択

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.1953195	36.61749065	-1.07039884	29.66%	115.34556	36.95492102
店舗面積	20.43522388	3.91825741	5.215385754	0.00%	12.28676152	28.58368623
アクセス数	-0.00073248	0.027729311	-0.02641534	97.92%	-0.05839874	0.05693378
広告	16.966648	2.142086527	7.920617487	0.00%	12.5119352	21.4213608
街の規模	15.68003698	1.973031864	7.947178793	0.00%	11.5768926	19.78318136
同業者店舗数	-4.04959516	1.996639687	-2.02820528	5.54%	-8.2018347	0.102644377

重回帰分析のデータ

店舗	売上	店舗面積	アクセス数	広告	街の規模	同業者店舗数
1	231	3	294	8.2	8.2	11
2	156	2.2	232	6.9	4.1	12
3	10	0.5	149	3	4.3	15
4	519	5.5	260	12	16.1	1
5	437	4.4	567	10.6	14.1	5
6	487	4.8	237	11.8	12.7	4
7	299	3.1	451	8.1	10.1	10
27	400	8.6	335	7	12	8

成果

取り除く

要因

分析結果を読む

概要

回帰統計	
重相関 R	0.99508524
重決定 R2	0.990194635
補正 R2	0.988411842
標準誤差	20.67511966
観測数	27

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	4	949676.2192	237419.0548	555.4174347	9.57991E-22
残差	22	9404.132604	427.4605729		
合計	26	959080.3519			

	斜傾	標準誤差	t	P-値	下限 95%	上限 95%
切片	-39.4600139	34.41056152	-1.14674135	0.263807961	-110.823151	31.90312291
店舗面積	20.4438867	3.814801743	5.359095459	2.21824E-05	12.5324721	28.35530129
広告	16.96614225	2.092787742	8.10695796	4.73186E-08	12.62596611	21.30631838
街の規模	15.67296182	1.909855732	8.206359021	3.85791E-08	11.71216345	19.63376018
同業者店舗数	-4.04330176	1.936828572	-2.08758886	0.048629059	-8.06003837	-0.02656514

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

	係数
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$(売上) = b_0 + b_1(\text{店舗面積})+b_2(\text{広告})+b_3(\text{街の規模})+b_4(\text{同業者店舗数})$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$(売上) = b_0 + b_1(\text{店舗面積}) + b_2(\text{広告}) + b_3(\text{街の規模}) + b_4(\text{同業者店舗数})$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176



$$(売上) = b_0 + b_1(\text{店舗面積}) + b_2(\text{広告}) + b_3(\text{街の規模}) + b_4(\text{同業者店舗数})$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$(売上) = b_0 + b_1(\text{店舗面積})+b_2(\text{広告})+b_3(\text{街の規模})+b_4(\text{同業者店舗数})$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$(売上) = b_0 + b_1(\text{店舗面積})+b_2(\text{広告})+b_3(\text{街の規模})+b_4(\text{同業者店舗数})$$

重回帰モデルの設計

店舗面積= 6, アクセス数= 400, 広告= 10, 街の規模= 15, 同業者店舗数= 10のときの売上を予測せよ

係数	
切片	-39.4600139
店舗面積	20.4438867
広告	16.96614225
街の規模	15.67296182
同業者店舗数	-4.04330176

$$(売上) = -39.4 + 20.4(6) + 16.9(10) + 15.6(15) - 4.0(10) = 447$$

売上の予測値

SUMPRODUCT関数

	係数	変数の値
切片	-39.460014	1
店舗面積	20.4438867	6
広告	16.9661423	10
街の規模	15.6729618	15
同業者店舗数	-4.0433018	10

=SUMPRODUCT(配列 1 , 配列 2)

SUMPRODUCT関数

	係数	変数の値
切片	-39.460014	1
店舗面積	20.4438867	6
広告	16.9661423	10
街の規模	15.6729618	15
同業者店舗数	-4.0433018	10



=SUMPRODUCT(配列 1 , 配列 2)

SUMPRODUCT関数

	係数	変数の値
切片	-39.460014	1
店舗面積	20.4438867	6
広告	16.9661423	10
街の規模	15.6729618	15
同業者店舗数	-4.0433018	10

=SUMPRODUCT(配列 1 , 配列 2)

=447.56261384

【SOTA】アップル 引越し需要予測



【SOTA】アップル 引越し需要予測
過去の引越し実績データから、将来の引越し需要を予測しよう！

■ アップル引越しセンター

SOTA Challenge ⓘ

締切：2020年12月7日（残り21日） 投稿：424件 参加：204人 元コンペはこちら

■ テーブル ■ ランキング：対象外

説明 データ リーダーボード 投稿済みファイル フォーラム

Data

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

■学習用データ(train.csv)

ファイル形式：ヘッダ付き、カンマ区切り

データ数：2,101行

データ説明：2010-07-01から2016-03-31までの引越しデータ

Data

Datetime 日時(YYYY-MM-DD) 文字列

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

Data

Y 引越し数 数値:整数

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

Data

Client : 法人が絡む特殊な引越し日フラグ
数値: 整数(0,1)

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

Data

Close 休業日 数値:整数(0,1)

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

Data

price_am 午前の料金区分(
-1は欠損を表す。5が最も料金が
高い) 数値: 整数(-1,0,1,2,3,4,5)

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

Data

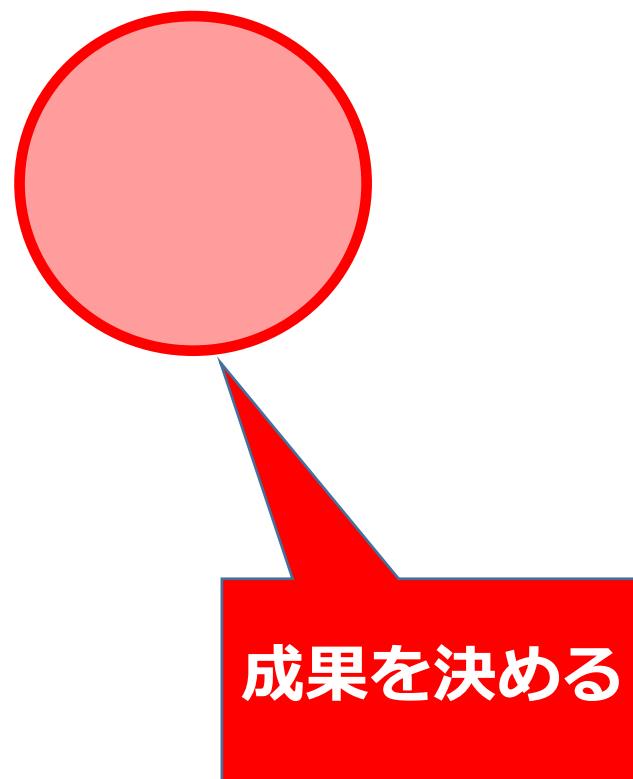
price_pm 午後の料金区分(-1は欠損を表す。5が最も料金が高い) 数値:整数(-1,0,1,2,3,4,5)

datetime	y	client	close	price_am	price_pm
2010/7/1	17	0	0	-1	-1
2010/7/2	18	0	0	-1	-1
2010/7/3	20	0	0	-1	-1
2010/7/4	20	0	0	-1	-1
2010/7/5	14	0	0	-1	-1
2010/7/6	14	0	0	-1	-1
2010/7/7	4	0	0	-1	-1
2010/7/8	10	0	0	-1	-1

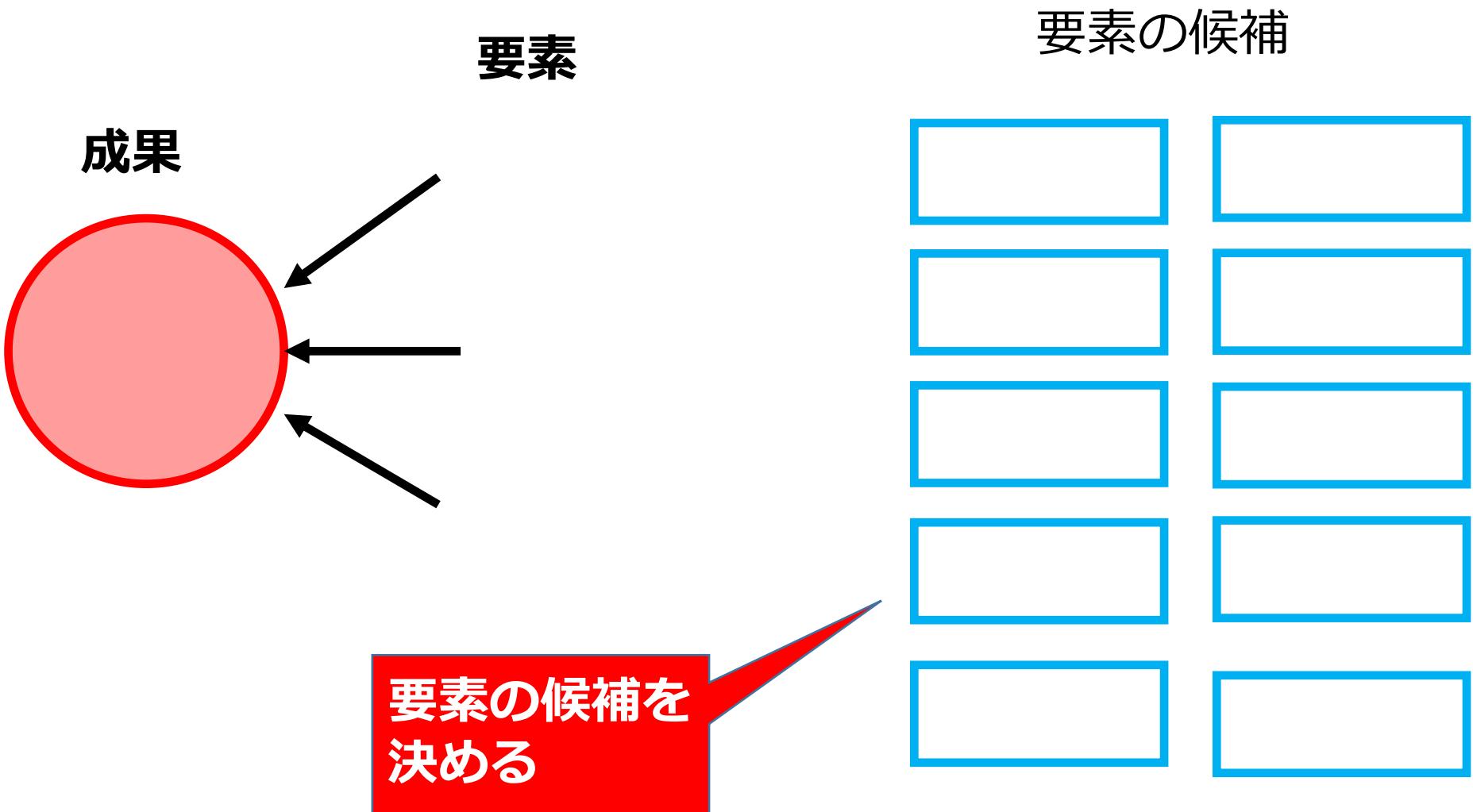
回帰分析（手順 1：成果を選ぶ）

課題を明確に認識すること

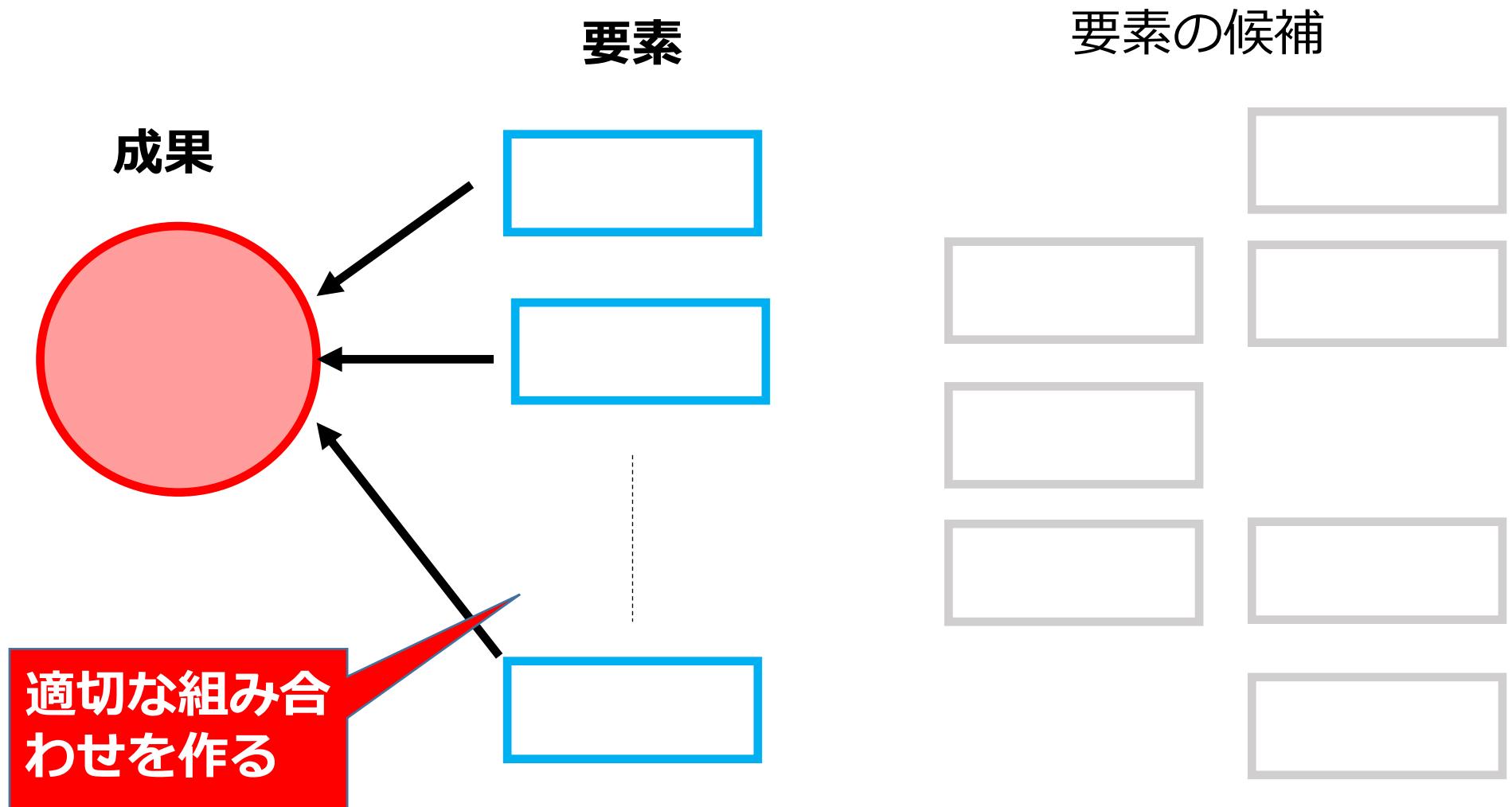
成果



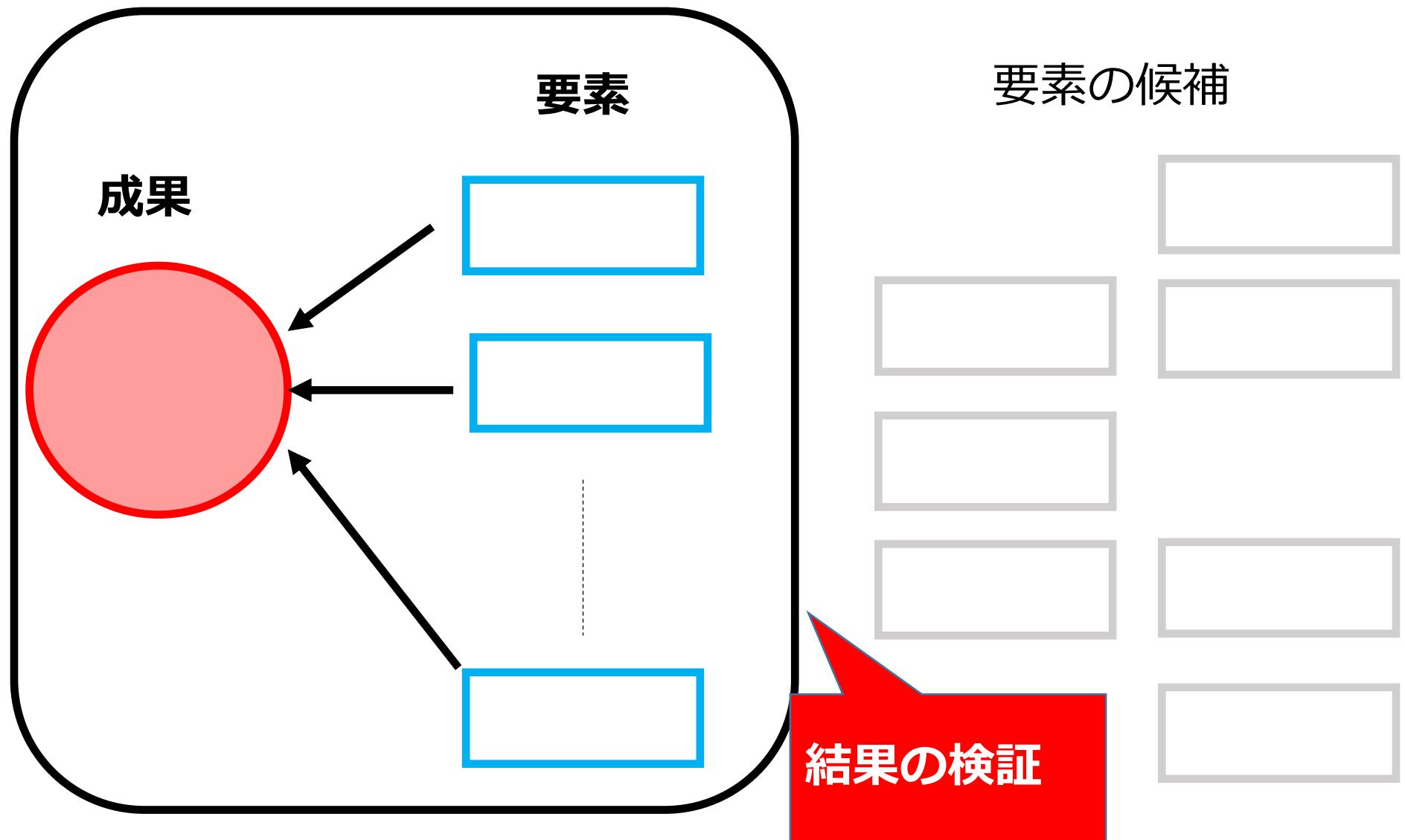
回帰分析（手順2：要素をリストアップ）



回帰分析（手順3：要素の組み合わせ）



回帰分析（手順4：結果の検証）



課題

訓練データ

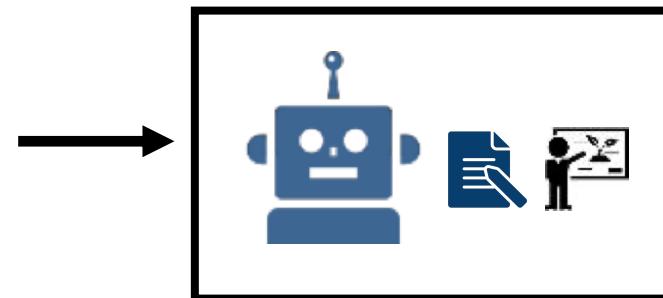
datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成

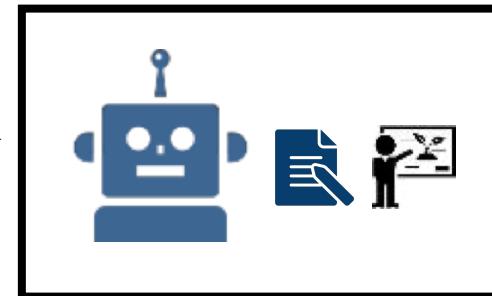


課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成



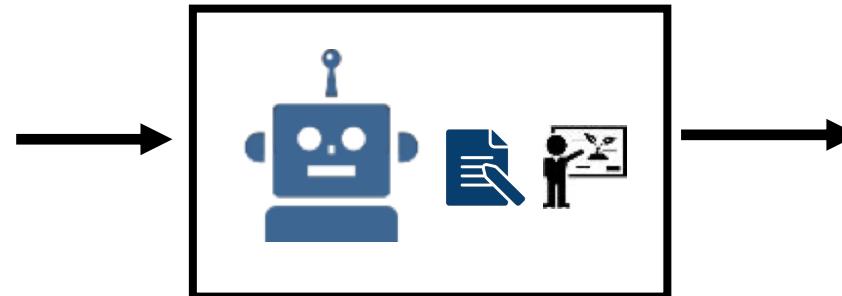
datetime	予測値
2012/7/1	
2012/7/2	
2012/7/3	
2012/7/4	
2013/6/28	
2013/6/29	
2013/6/30	
2013/7/1	

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成



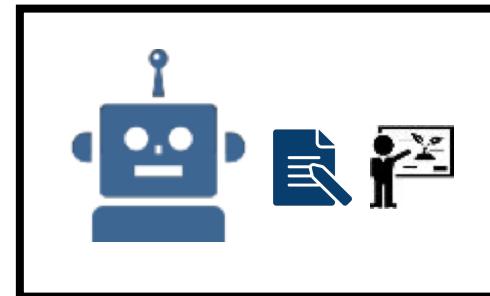
datetime	予測値
2012/7/1	34
2012/7/2	32
2012/7/3	45
2012/7/4	15
2013/6/28	54
2013/6/29	91
2013/6/30	43
2013/7/1	55

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成

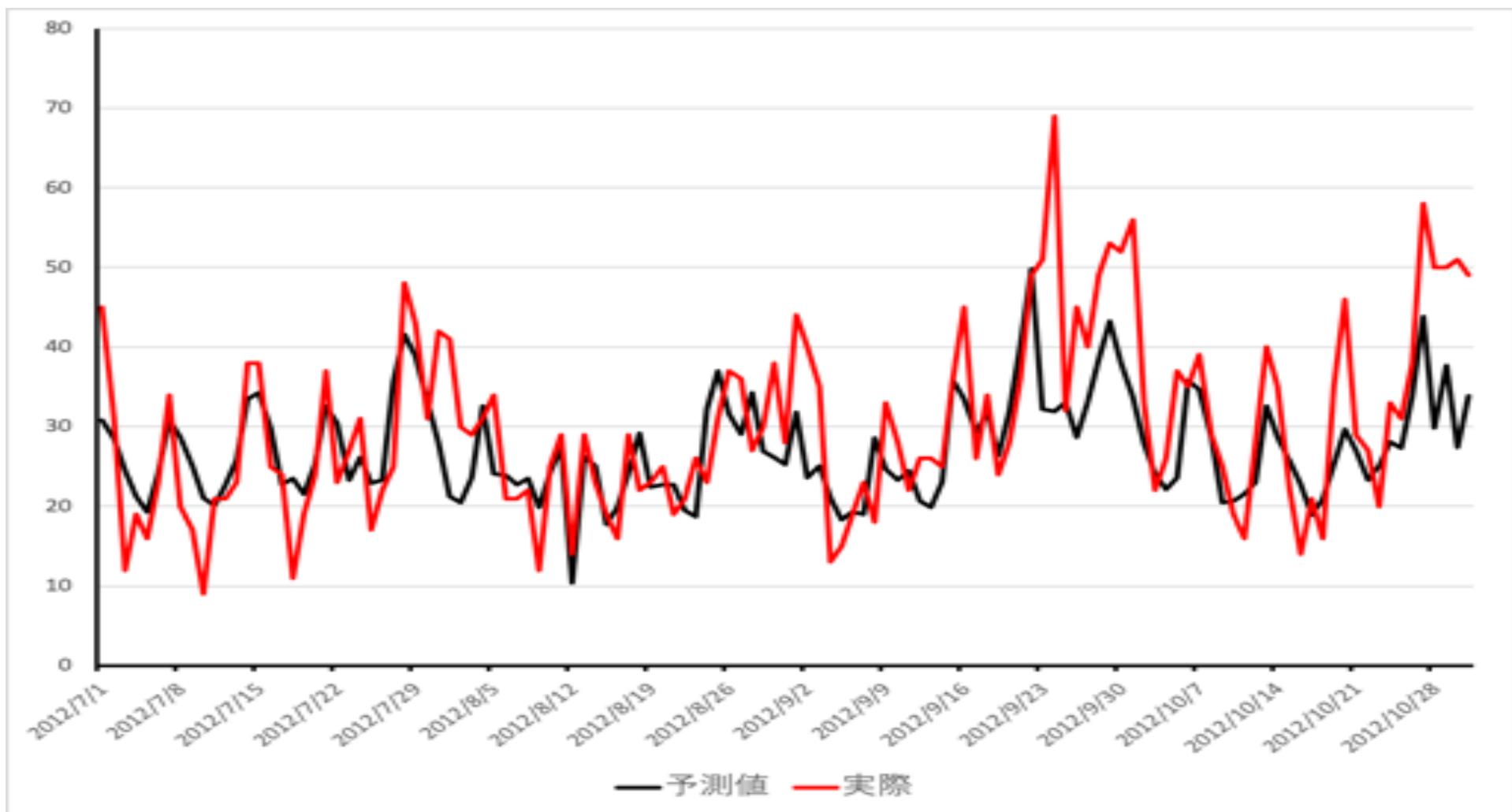


datetime	予測値	y
2012/7/1	34	45
2012/7/2	32	31
2012/7/3	45	12
2012/7/4	15	19
2013/6/28	54	61
2013/6/29	91	84
2013/6/30	43	77
2013/7/1	55	37



比較

結果の比較



二乗平均平方誤差 RMSE

- RMSE は、root mean squared error の略で、回帰モデルの誤差を評価する指標の一つである。RMSE は、観測値を y_i 、モデルから計算した予測値を \hat{y}_i とすると、次の式によって定義される。

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

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

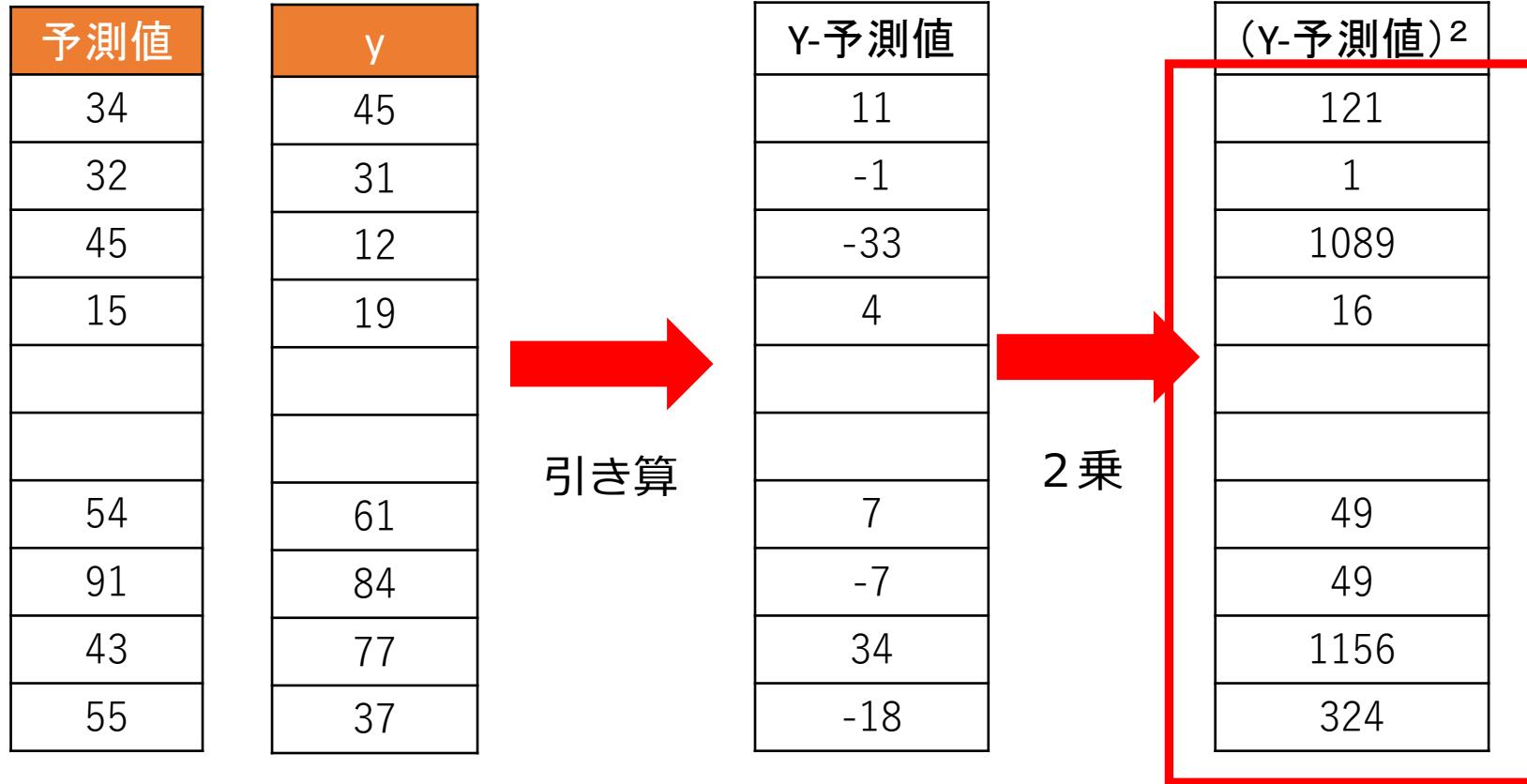
Y-予測値
11
-1
-33
4
7
-7
34
-18



2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

RMSEの計算



合計する
=sum()

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

32611.06801

合計する
=sum()

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

32611.06801

n=366 で割る

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

89.10127872

n=366 で割る

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

89.10127872

平方根をとる
=sqrt()

RMSEの計算

予測値
34
32
45
15
54
61
91
43
55

y
45
31
12
19
54
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

RMSE
9.439347367

モデルを改善する

特徴量の追加

datetime	y	client	close	price_am	price_pm
2011/7/10	20	0	0	1	1
2011/7/11	17	0	0	0	0
2011/7/12	10	0	0	0	0
2011/7/13	9	0	0	0	0
2011/7/14	16	0	0	0	0
2011/7/15	19	0	0	0	0
2011/7/16	17	0	0	2	1
2011/7/17	23	0	0	2	1
2011/7/18	19	0	0	1	0
2011/7/19	13	0	0	0	0

休日

特徴量の追加

datetime	y	client	close	price_am	price_pm
2011/7/10	20	0	0	1	1
2011/7/11	17	0	0	0	0
2011/7/12	10	0	0	0	0
2011/7/13	9	0	0	0	0
2011/7/14	16	0	0	0	0
2011/7/15	19	0	0	0	0
2011/7/16	17	0	0	2	1
2011/7/17	23	0	0	2	1
2011/7/18	19	0	0	1	0
2011/7/19	13	0	0	0	0

休日
0
0
0
0
0
0
0
0
1
0

休日

特徴量の追加

datetime	y
2011/7/10	20
2011/7/11	17
2011/7/12	10
2011/7/13	9
2011/7/14	16
2011/7/15	19
2011/7/16	17
2011/7/17	23
2011/7/18	19
2011/7/19	13

曜日
日曜
月曜
火曜
水曜
木曜
金曜
土曜
日曜
月曜
火曜

特徴量の追加

datetime	y
2011/7/10	20
2011/7/11	17
2011/7/12	10
2011/7/13	9
2011/7/14	16
2011/7/15	19
2011/7/16	17
2011/7/17	23
2011/7/18	19
2011/7/19	13

曜日
日曜
月曜
火曜
水曜
木曜
金曜
土曜
日曜
月曜
火曜

月曜	火曜	水曜	木曜	金曜	土曜
0	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0

特徴量の追加

datetime	y
2011/7/10	20
2011/7/11	17
2011/7/12	10
2011/7/13	9
2011/7/14	16
2011/7/15	19
2011/7/16	17
2011/7/17	23
2011/7/18	19
2011/7/19	13

曜日
日曜
月曜
火曜
水曜
木曜
金曜
土曜
日曜
月曜
火曜

月曜	火曜	水曜	木曜	金曜	土曜
0	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0

日曜: 参照ダミー

特徴量の追加

datetime	y
2012/2/25	49
2012/2/26	47
2012/2/27	41
2012/2/28	40
2012/2/29	39
2012/3/1	36
2012/3/2	45
2012/3/3	37
2012/3/4	30
2012/3/5	29

3 or 4月
0
0
0
0
0
1
1
1
1
1

3月と4月

特徴量の追加

datetime	y
2011/6/30	24
2011/7/1	16
2011/7/2	25
2011/7/3	17
2011/7/4	11
2011/7/5	8
2011/7/6	11
2011/7/7	10
2011/7/8	16
2011/7/9	19

2011年

6月
June (水無月)

日	月	火	水	木	金	土
29	30	31	1 先負	2 大安	3 赤口	4 先勝
5 友引	6 先負	7 仏滅	8 大安	9 赤口	10 先勝	11 友引
12 先負	13 仏滅	14 大安	15 赤口	16 先勝	17 友引	18 先負
19 仏滅	20 大安	21 赤口	22 先勝	23 先負	24 仏滅	25 大安
26 大安	27 赤口	28 先勝	29 友引	30 先負	1	2
3	4	5	6	7	8	9

2010年

6月
June (水無月)

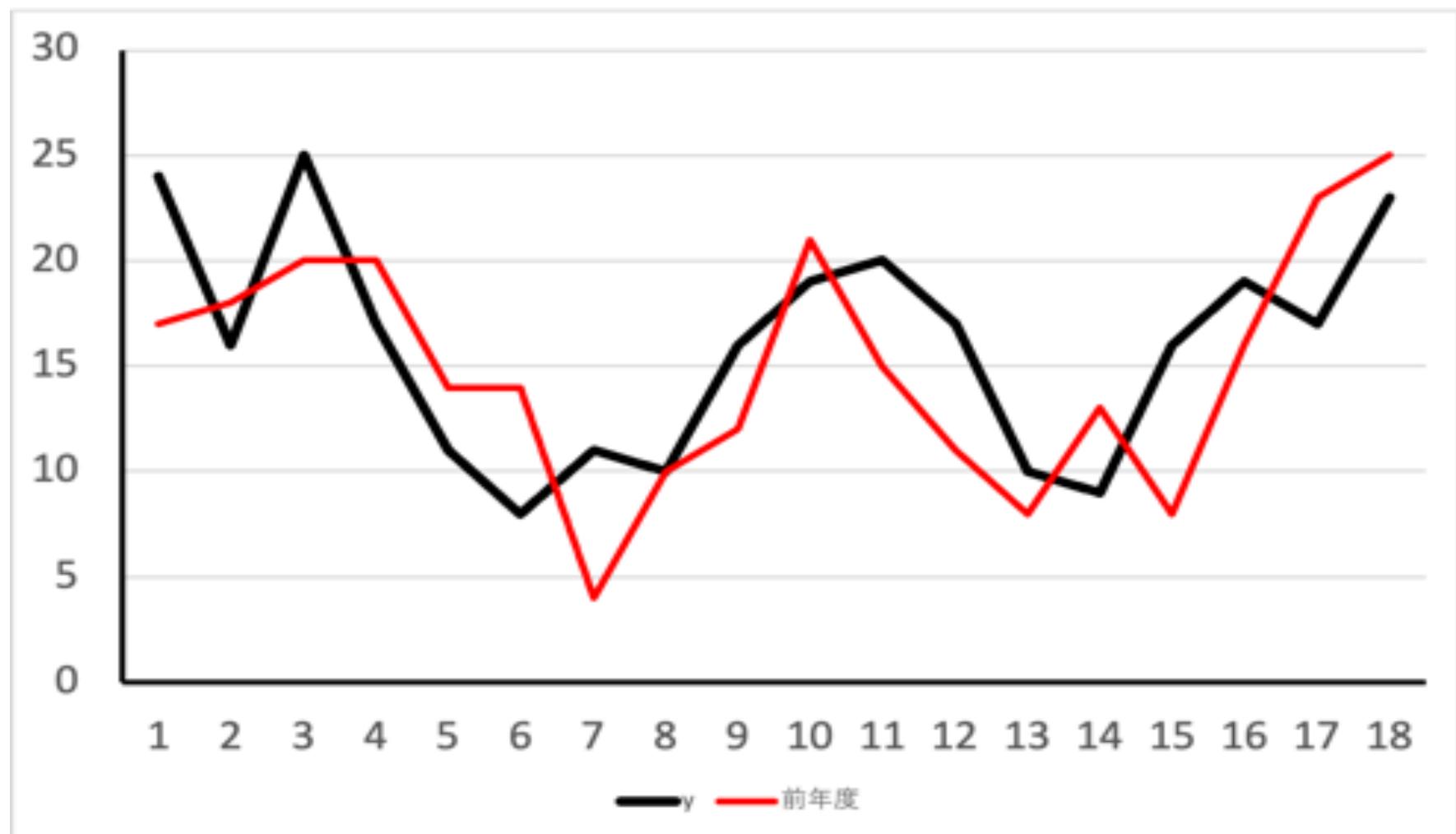
日	月	火	水	木	金	土
30	31	1 仏滅	2 大安	3 赤口	4 先勝	5 友引
6 先負	7 仏滅	8 大安	9 赤口	10 先勝	11 友引	12 大安
13 赤口	14 先勝	15 友引	16 先負	17 仏滅	18 大安	19 赤口
20 先勝	21 友引	22 先負	23 仏滅	24 大安	25 赤口	26 先勝
27 友引	28 先負	29 仏滅	30 大安	1	2	3
4	5	6	7	8	9	10

特徴量の追加

datetime	y
2011/6/30	24
2011/7/1	16
2011/7/2	25
2011/7/3	17
2011/7/4	11
2011/7/5	8
2011/7/6	11
2011/7/7	10
2011/7/8	16
2011/7/9	19

datetime	y
2010/7/1	17
2010/7/2	18
2010/7/3	20
2010/7/4	20
2010/7/5	14
2010/7/6	14
2010/7/7	4
2010/7/8	10
2010/7/9	12
2010/7/10	21

前年度と比較



特徴量の追加

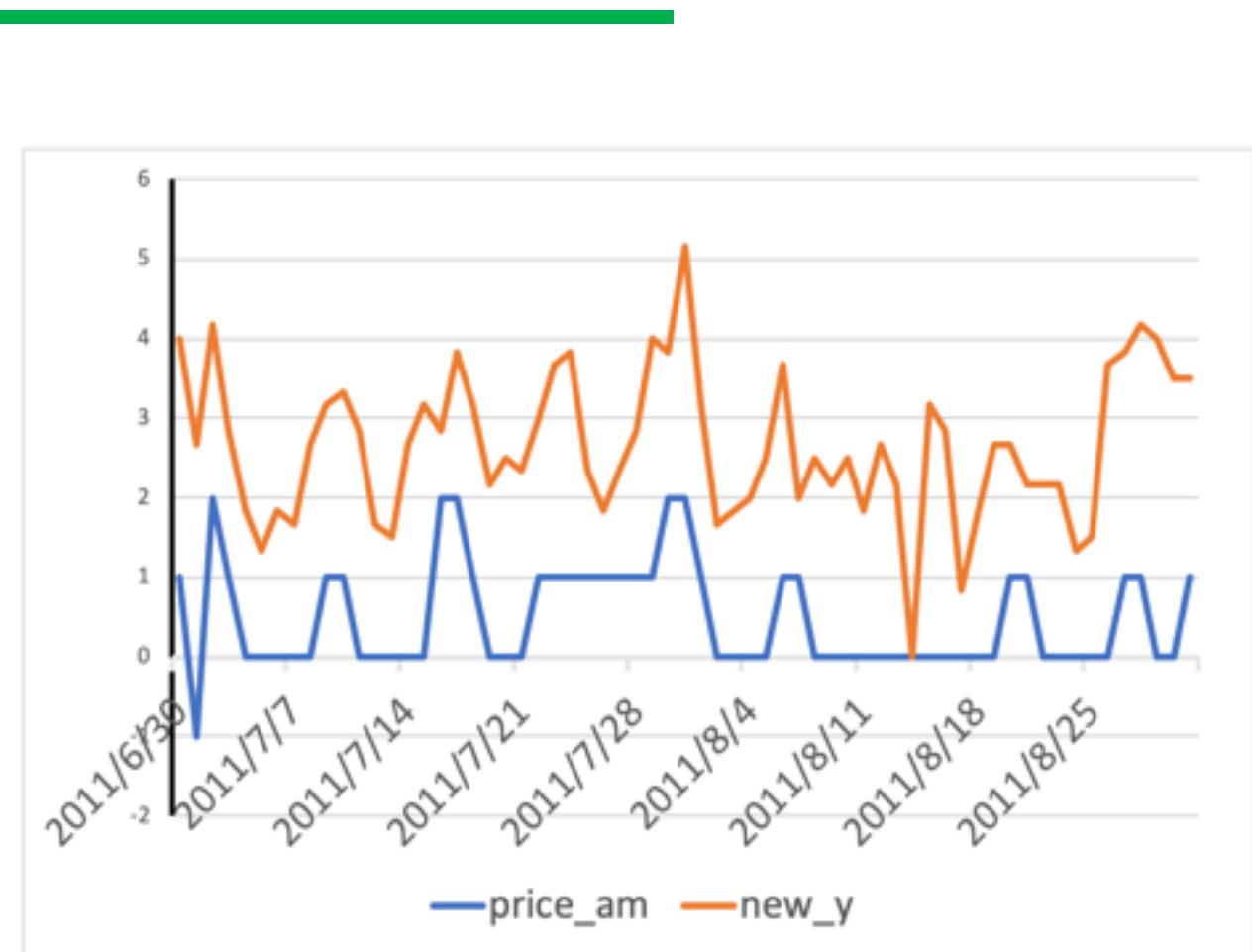
datetime	y
2012/2/25	49
2012/2/26	47
2012/2/27	41
2012/2/28	40
2012/2/29	39
2012/3/1	36
2012/3/2	45
2012/3/3	37
2012/3/4	30
2012/3/5	29

Baseline
17
18
20
20
14
14
4
10
12
21

前年度のY値

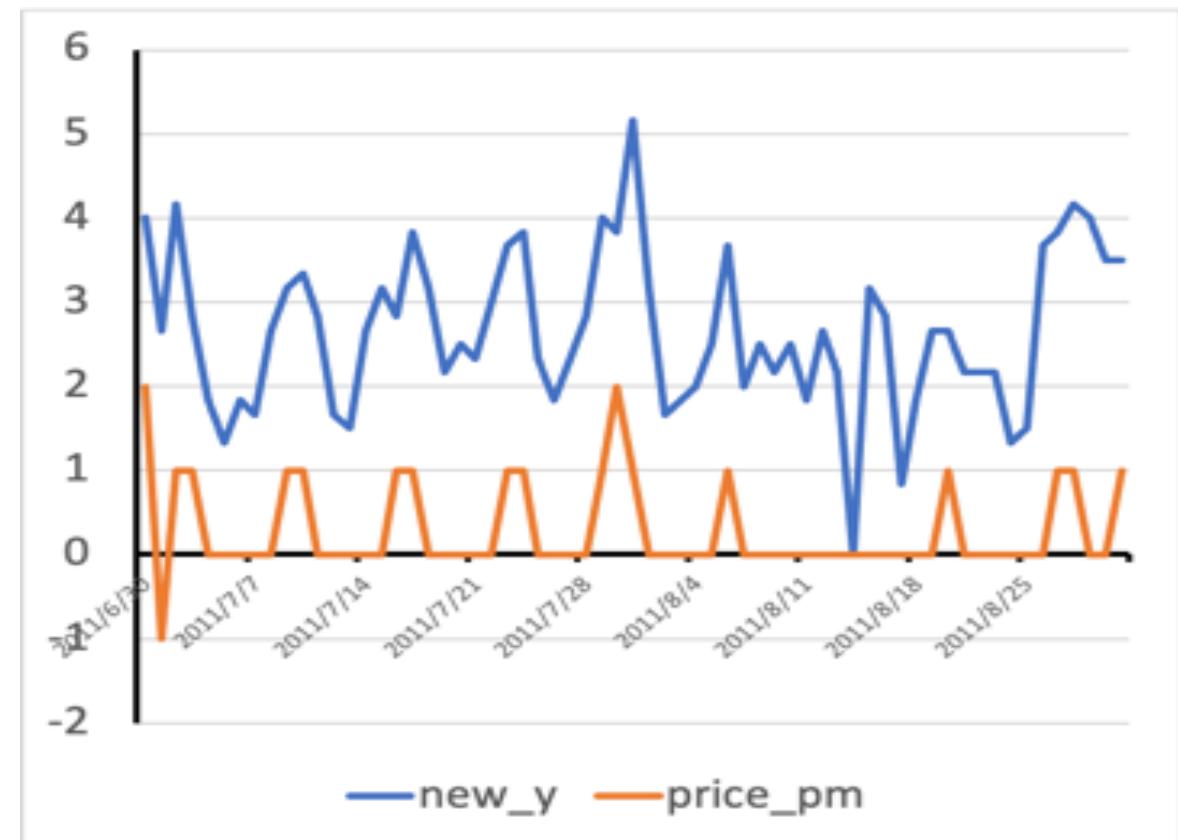
特徴量の追加

datetime	y	price_am
2011/6/30	24	1
2011/7/1	16	-1
2011/7/2	25	2
2011/7/3	17	1
2011/7/4	11	0
2011/7/5	8	0
2011/7/6	11	0
2011/7/7	10	0
2011/7/8	16	0
2011/7/9	19	1



特徴量の追加

datetime	y	price_am
2011/6/30	24	2
2011/7/1	16	-1
2011/7/2	25	1
2011/7/3	17	1
2011/7/4	11	0
2011/7/5	8	0
2011/7/6	11	0
2011/7/7	10	0
2011/7/8	16	0
2011/7/9	19	1



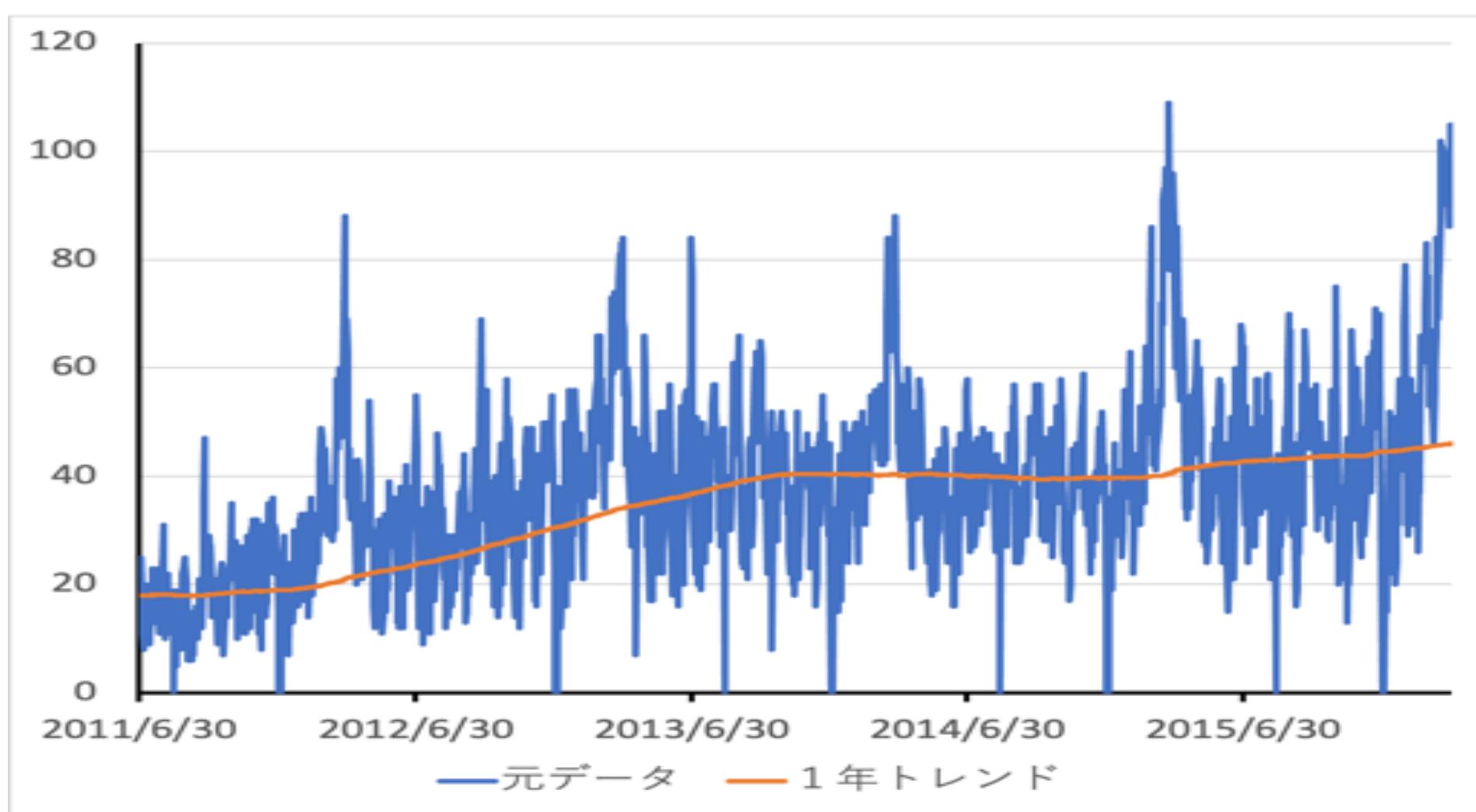
特徴量の追加

datetime	y	price
2011/6/30	24	6
2011/7/1	16	0
2011/7/2	25	6
2011/7/3	17	4
2011/7/4	11	1
2011/7/5	8	1
2011/7/6	11	1
2011/7/7	10	1
2011/7/8	16	1
2011/7/9	19	4



(Price_am+1)*(price_pm+1)

トレンド



課題

訓練データ

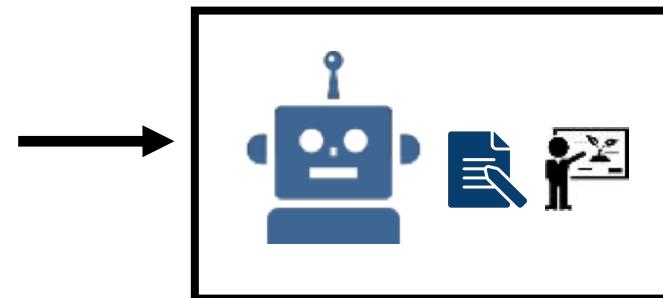
datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成

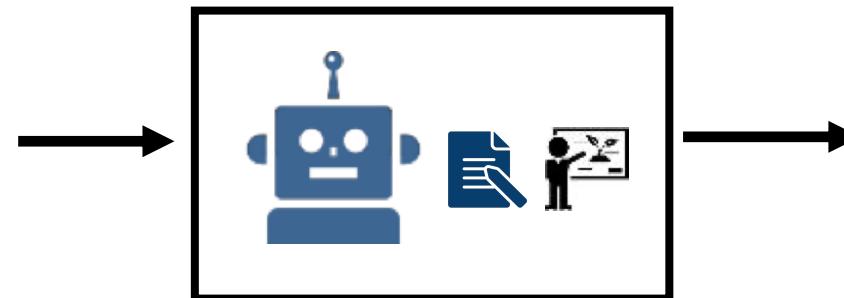


課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成



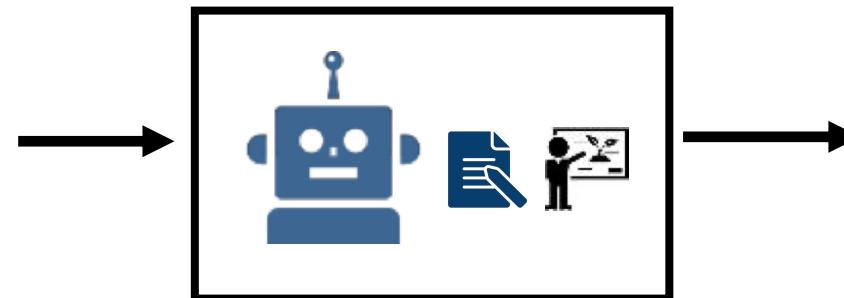
datetime	予測値
2012/7/1	
2012/7/2	
2012/7/3	
2012/7/4	
2013/6/28	
2013/6/29	
2013/6/30	
2013/7/1	

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成



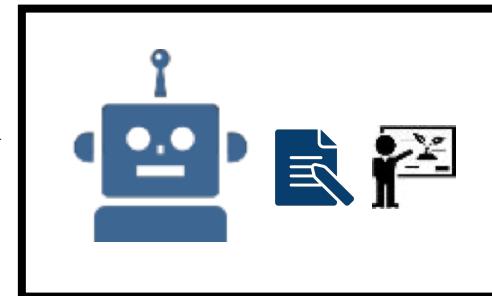
datetime	予測値
2012/7/1	34
2012/7/2	32
2012/7/3	45
2012/7/4	15
2013/6/28	54
2013/6/29	91
2013/6/30	43
2013/7/1	55

課題

訓練データ

datetime	y
2010/7/1	16
2010/7/2	25
2010/7/3	17
2012/6/27	31
2012/6/28	42
2012/6/29	52
2012/6/30	55

モデルの形成

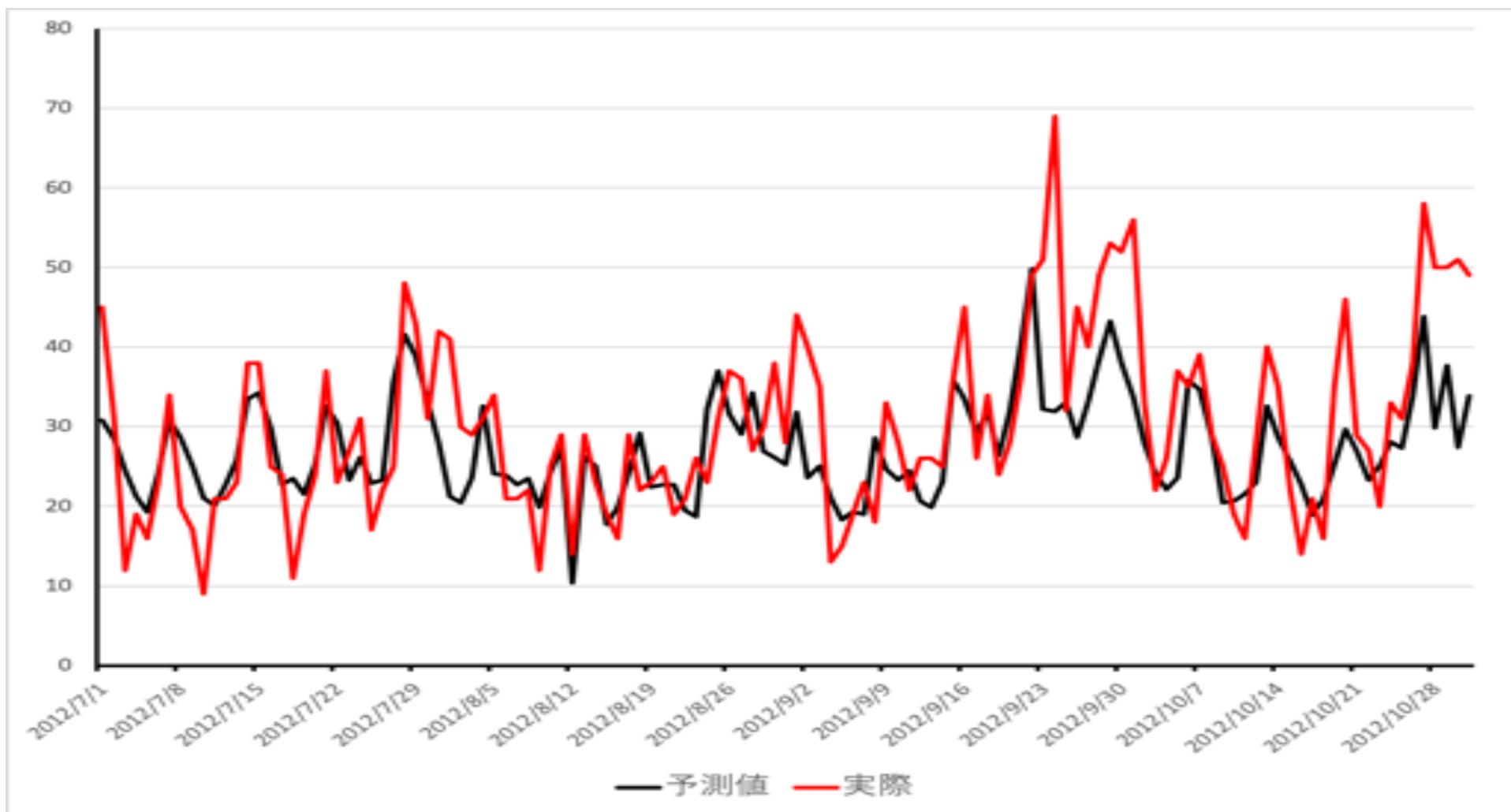


datetime	予測値	y
2012/7/1	34	45
2012/7/2	32	31
2012/7/3	45	12
2012/7/4	15	19
2013/6/28	54	61
2013/6/29	91	84
2013/6/30	43	77
2013/7/1	55	37



比較

結果の比較



二乗平均平方誤差 RMSE

- RMSE は、root mean squared error の略で、回帰モデルの誤差を評価する指標の一つである。RMSE は、観測値を y_i 、モデルから計算した予測値を \hat{y}_i とすると、次の式によって定義される。

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

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

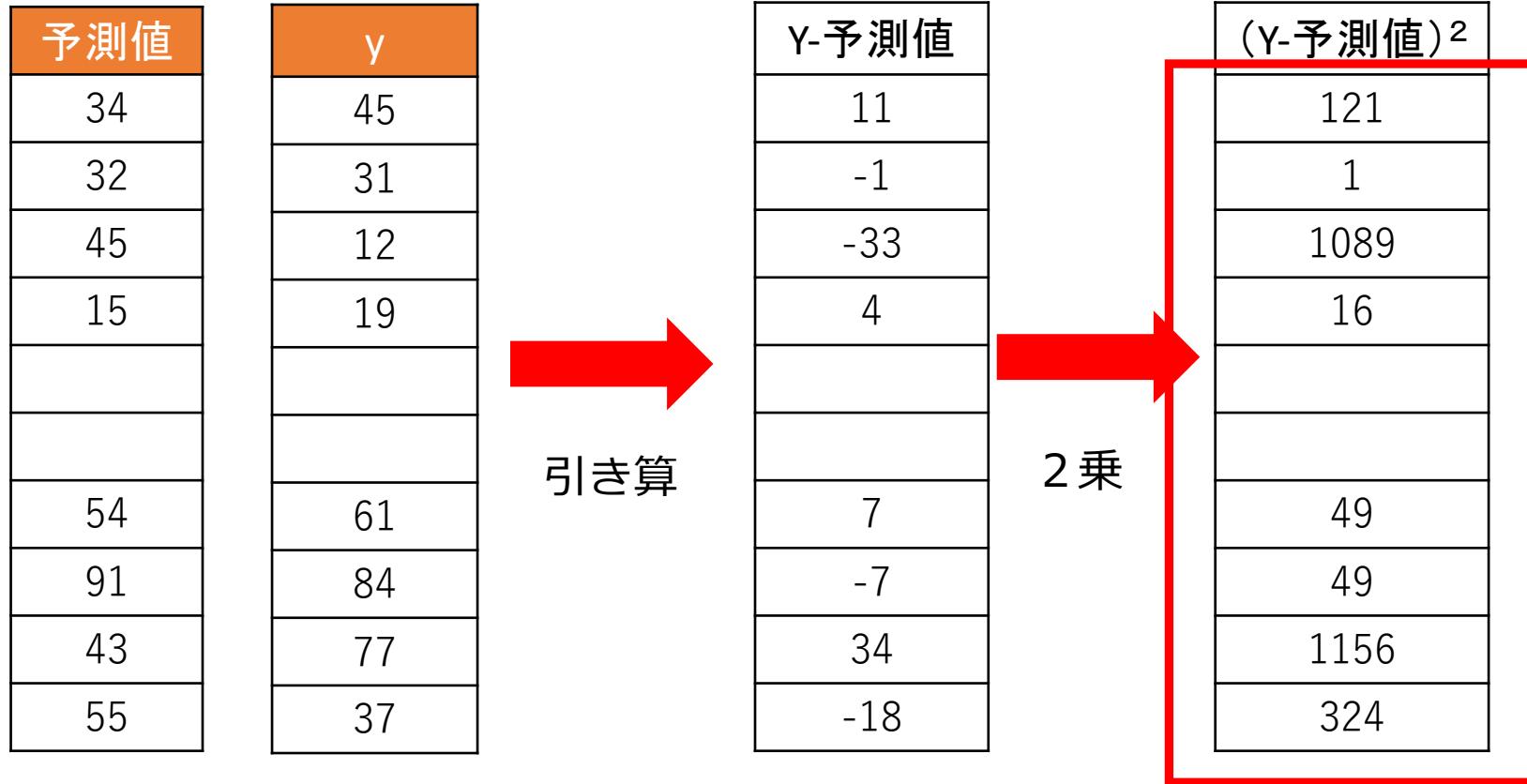
Y-予測値
11
-1
-33
4
7
-7
34
-18



2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

RMSEの計算



合計する
=sum()

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

32611.06801

合計する
=sum()

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

32611.06801

n=366 で割る

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

89.10127872

n=366 で割る

RMSEの計算

予測値
34
32
45
15
54
91
43
55

y
45
31
12
19
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

89.10127872

平方根をとる
=sqrt()

RMSEの計算

予測値
34
32
45
15
54
61
91
43
55

y
45
31
12
19
54
61
84
77
37



引き算

Y-予測値
11
-1
-33
4
7
-7
34
-18

2乗

(Y-予測値) ²
121
1
1089
16
49
49
1156
324

RMSE
9.439347367