専門講座 「データ分析実践」 WEBセミナー

データ分析入門Ⅱ
マーケティングとデータ分析
ー予測と回帰分析ー

森 裕一 **岡山理科大学 経営学部**

- 兵庫県生まれ。
- 大阪教育大学、神戸大学大学院修士課程、岡山大学大学院博士課程修了。博士（学術）。
- 中学校、短期大学の教員を経て、現職。授業では、データサイエンス系の科目を担当。
- 専門は計算機統計学。
- 国内外のジャーナルの編集委員や学会の理事、国際計算機統計協会アジア地区会長（～2021年）など。
- 官公庁や民間のお手伝い、統計相談、統計教育、「統計検定」にもかかわる。

2021/08/17 おかやまデータ活用人材育成講座 WEBセミナー

目　次

- 回帰分析を適用する
- 回帰分析（記述と予測）
  - 単回帰分析
  - 重回帰分析
  - カテゴリカルデータに対する回帰分析
- マーケティングにおける活用
  - 特長と留意点

❗ 回帰分析を適用する…

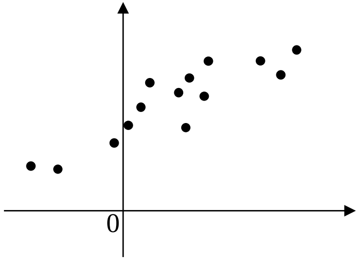
分析をしたい ⇒ 回帰分析

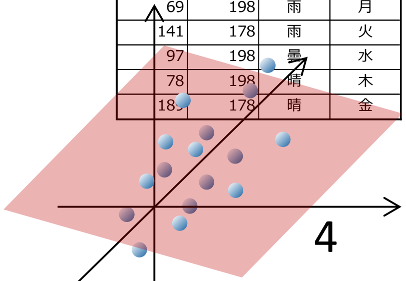
企業や店舗の売上向上
顧客満足度の評価向上
重要な影響を与えている要因

- 成果（結果）の要因（原因）を知りたい。
- 複数の要因を用いて、成果を評価したい。
- それぞれの要因の成果への影響度を判断したい。
- 数値で統計的な予測を立てたい。
- 事象を説明する（プレゼンする）ための根拠となるデータを示したい。

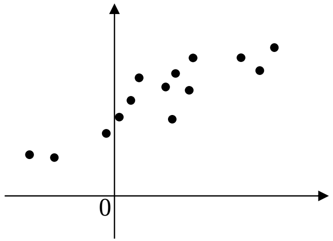
期	売上総額	Web広告費	チラシ宣伝費	雑誌広告費
2018-1	894,884	13,000	101,000	15,000
2018-2	806,523	11,000	66,000	20,000
2018-3	921,544	9,000	95,000	
2018-4	791,293	7,000	84,000	
2018-5	743,659	7,000	76,000	
2018-6	931,192	16,000	109,000	
2018-7	877,314	13,000	99,000	
2018-8	903,283	16,000	75,000	
2018-9	834,065	9,000	72,000	
2018-10	822,309	9,000	90,000	
2018-11	801,600	7,000	75,000	
2018-12	804,382	10,000	93,000	
2018-13	924,139	12,000	87,000	
2018-14	848,518	14,000	73,000	
2018-15	871,736	12,000	100,000	
2018-16	869,016	12,000	85,000	
2018-17	937,214	14,000	103,000	
2018-18	835,121	13,000	86,000	
2018-19	800,103	11,000	56,000	
2018-20	844,836	8,000	82,000	
2018-21	903,386	14,000	84,000	
2018-22	862,425	10,000	89,000	
2018-23	881,652	15,000	77,000	
2018-24	873,669	8,000	89,000	
2018-25	764,985	7,000	67,000	

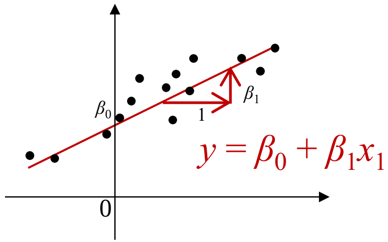
本数	単価	天気	曜日
95	198	晴	木
74	198	晴	金
251	175	曇	土
167	175	曇	日
99	198	曇	月
216	175	曇	火
77	198	晴	水
159	188	晴	木
181	178	晴	金
212	175	晴	土
187	175	晴	日
79	198	晴	月
118	198	晴	火
59	198	晴	水
94	188	曇	金
255	175	晴	土
171	175	晴	日
90	198	曇	月
279	178	曇	金
143	175	雨	土
208	175	曇	日
69	198	雨	月
141	178	雨	火
97	198	曇	水
78	198	晴	木
18[illegible]	178	晴	金





回帰分析の流れ





y を x で説明（記述）
y の予測も

❶関係を見つける → ❷関係式を求める → ❸-1関係式を評価する / ❸-2関係を解釈する / ❸-3予測する

散布図／相関係数 など
⇓
目的変数と**説明変数**を抽出

モデルを**１次式**で作る
⇓
係数の算出
評価指標の算出
||
ExcelやRで計算

関係式の**適切性**

β_0、β_1から**現象を解釈**
変数（x_1）の影響

関係式を使って**予測**

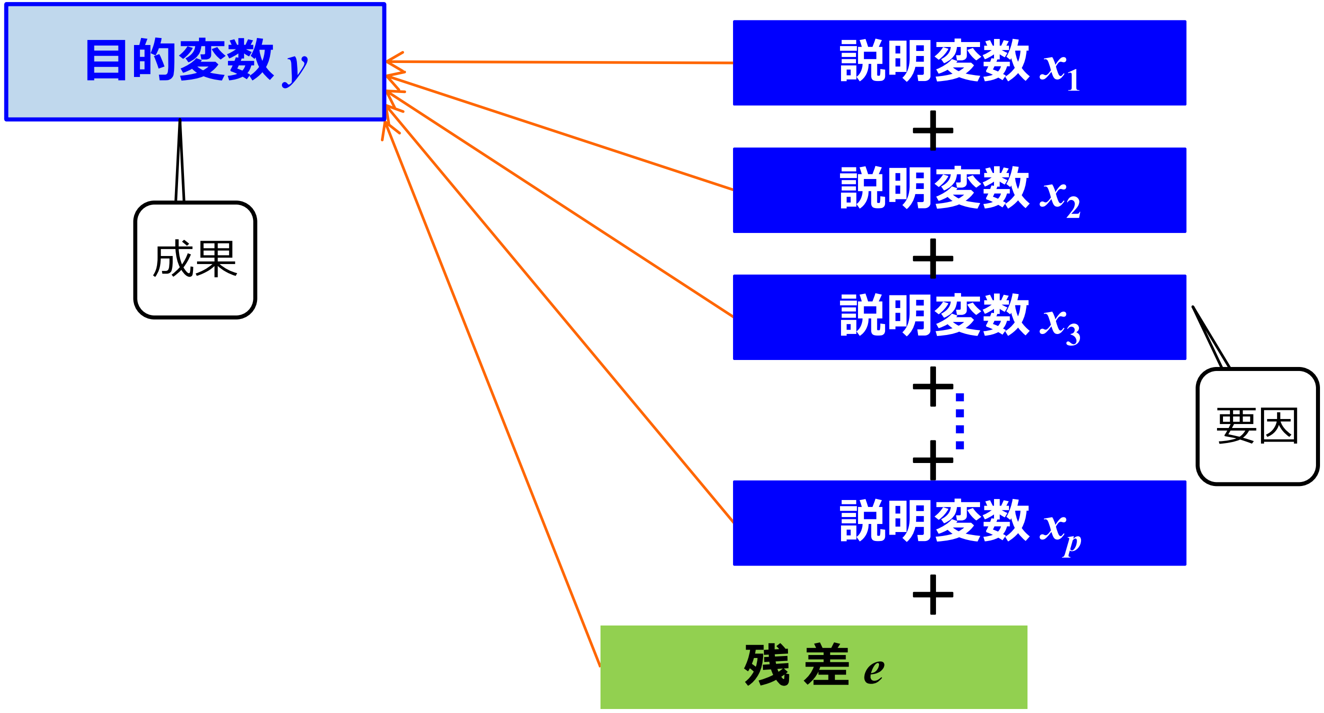
目標達成のための要因操作も

◯**線形関係**
✕非線形関係 直線関係に変換
直線で近似
数式で表せる＝計算ができる＝予測へ

中学校の１次関数
$y=ax+b$ ・・・・・・

- a の意味は？
- b の意味は？
- x が与えられたときの y の値は？

回帰分析のモデル



y アイスクリームの支出金額(円)	x_1 日平均気温の月平均値(℃)	x_2 相対湿度の月平均値(%)	x_3 日照時間の月合計値(h)	x_4 現地気圧の月平均値(hPa)
362	7.0	41	221.9	1010.2
305	6.5	60	118.3	1012.6
383	9.1	61	139.8	1012.7
464	12.4	62	139.9	1013.6
752	19.0	60	198.8	1007.3
841	23.6	67	162.5	1007.1
1211	28.0	70	182.7	1005.9
1451	29.6	67	222.6	1009.7
864	25.1	68	165.3	1008.3
504	18.9	68	81.4	1012.7
351	13.5	56	158.9	1012.3
423	9.9	50	194.9	1005.1
346	5.1	36	243.9	1009.1
289	7.0	52	148.9	1013.6
329	8.1	47	214.8	1010.5
462	14.5	50	204	1008.9
672	18.5	63	146.3	1007.5
791	22.8	71	105.1	1005.3
1265	27.3	67	186.2	1004.5
1241	27.5	71	168.9	1006.4
767	25.1	68	165.8	1007.7
516	19.5	61	141.3	1013.6
393	14.9	58	143.4	1015.4
423	7.5	48	187.6	1013.3

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots\cdots + \beta_p x_p + e$$

1 次関数と回帰分析

1 次関数　2 つの点から直線・式を求める。

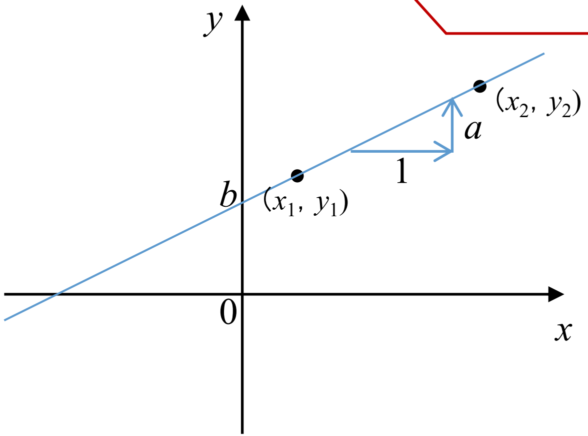
統計　複数の点から最も理想的な超平面・式を求める。

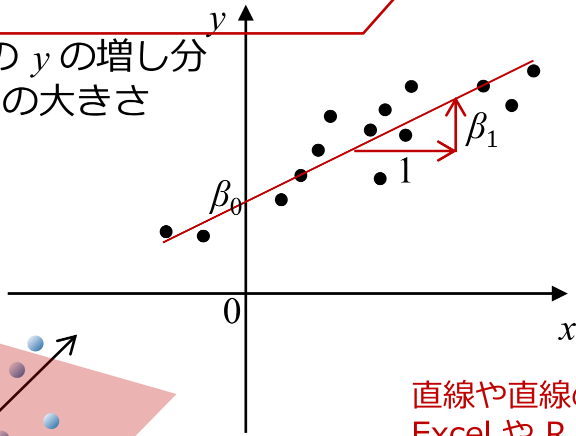
$$y = \underbrace{a}_{\text{傾き}}x + \underbrace{b}_{\text{切片}} \qquad\qquad y = \beta_0 + \beta_1 x$$

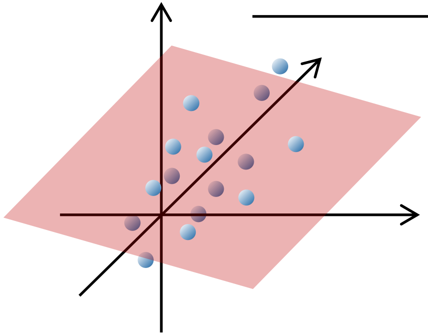
x が 0 のときの y の値

x が 1 増えたときの y の増し分

x の y への影響の大きさ







直線や直線の式は
Excel や R が
求めてくれる。

! 回帰分析（記述と予測）

単回帰分析

1 商品の売れ具合を気象からみる

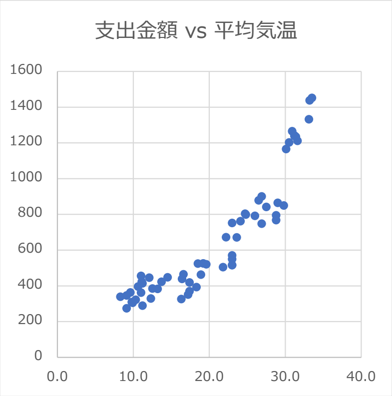
アイスクリームの毎月の支出金額と各月の気象のデータ（月の平均値）があります。

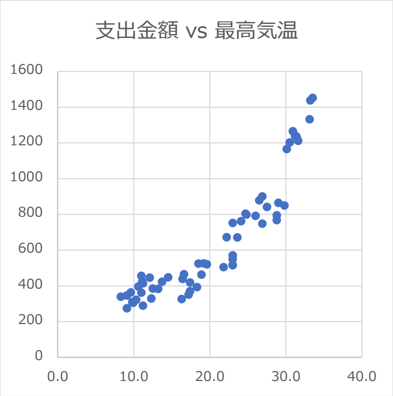
ここから、アイスクリームの支出に最も影響のあると思われる気象要因を特定し、その関係を明らかにしましょう。

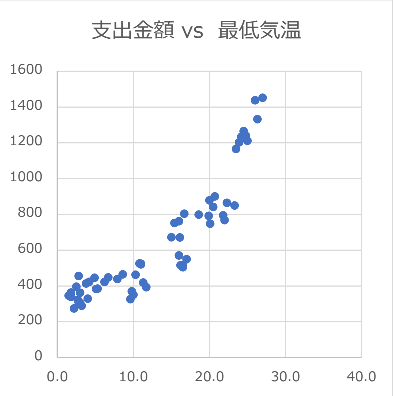
年	月	アイスクリームの支出金額(円)	日平均気温の月平均値（℃）	日最高気温の月平均値（℃）	日最低気温の月平均値（℃）	相対湿度の月平均値（%）	日照時間の月合計値（h）	現地気圧の月平均値（hPa）
2010	1	362	7.0	11.0	3.0	41	221.9	1010.2
2010	2	305	6.5	9.9	3.0	60	118.3	1012.6
2010	3	383	9.1	13.2	5.1	61	139.8	1012.7
2010	4	464	12.4	16.6	8.6	62	139.9	1013.6
2010	5	752	19.0	23.0	15.4	60	198.8	1007.3
2010	6	841	23.6	27.5	20.5	67	162.5	1007.1
2010	7	1211	28.0	31.6	25.0	70	182.7	1005.9
2010	8	1451	29.6	33.5	27.0	67	222.6	1009.7
2010	9	864	25.1	29.0	22.3	68	165.3	1008.3
2010	10	504	18.9	21.8	16.5	68	81.4	1012.7
2010	11	351	13.5	17.2	10.0	56	158.9	1012.3
2010	12	423	9.9	13.7	6.2	50	194.9	1005.1
2011	1	346	5.1	9.1	1.5	36	243.9	1009.1
2011	2	289	7.0	11.2	3.2	52	148.9	1013.6
2011	3	329	8.1	12.3	4.0	47	214.8	1010.5
2011	4	462	14.5	18.9	10.3	50	204	1008.9
2011	5	672	18.5	22.2	15.0	63	146.3	1007.5
2011	6	791	22.8	26.0	19.9	71	105.1	1005.3
2011	7	1265	27.3	30.9	24.5	67	186.2	1004.5
2011	8	1241	27.5	31.2	24.6	71	168.9	1006.4
2011	9	767	25.1	28.8	22.0	68	165.8	1007.7
2011	10	516	19.5	23.0	16.5	61	141.3	1013.6
2011	11	393	14.9	18.3	11.7	58	143.4	1015.4
[illegible]	12	[illegible]	[illegible]	[illegible]	4.2	48	[illegible]	[illegible]

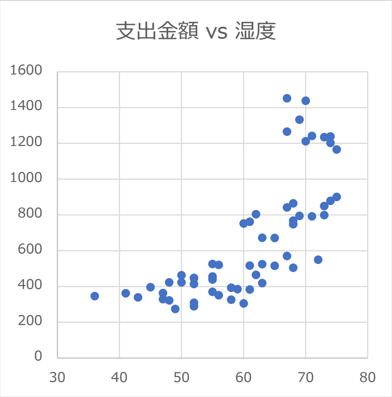
❶関係を見つける

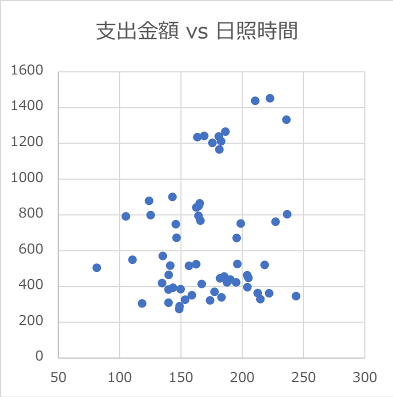
アイスクリームの支出金額を**成果**として、その**要因**となっていそうな気象データを散布図でさがす。

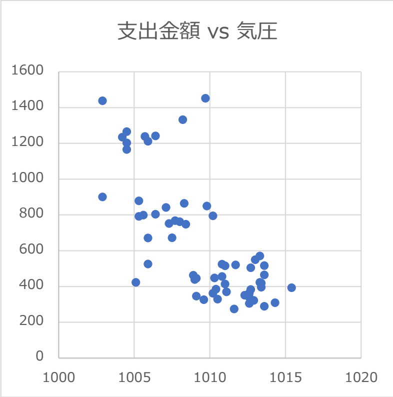








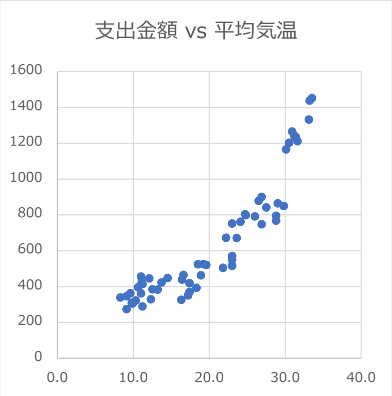


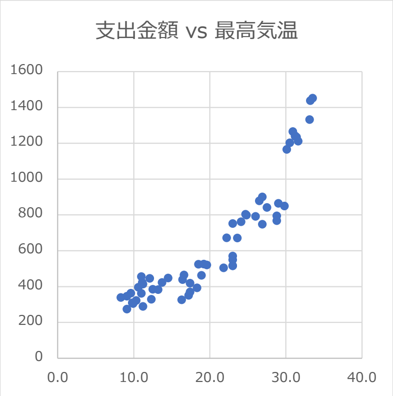


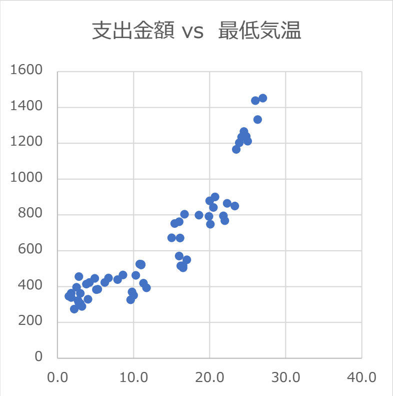
❶関係を見つける

アイスクリームの支出金額を**成果**として、その**要因**となっていそうな気象データを散布図でさがす。

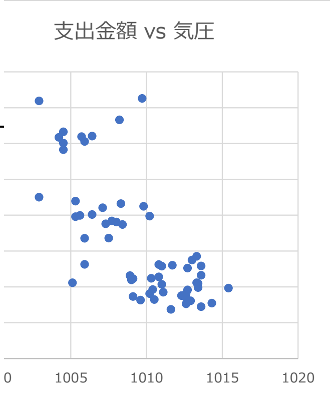
相関係数で確認





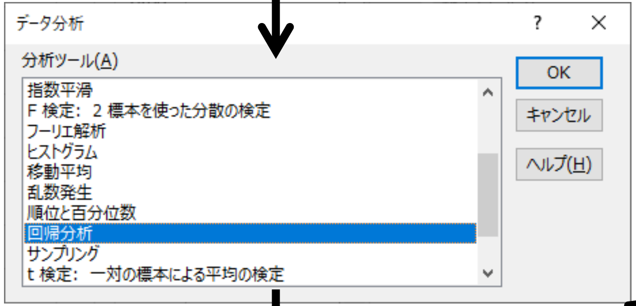


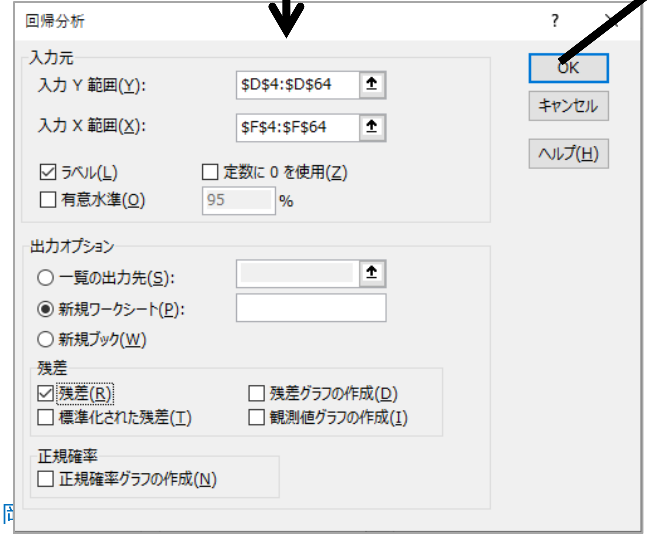
	アイスクリームの支出金額(円)	日平均気温の月平均値（℃）	日最高気温の月平均値（℃）	日最低気温の月平均値（℃）	相対湿度の月平均値（%）	日照時間の月合計値（h）	現地気圧の月平均値（hPa）
アイスクリームの支出金額(円)	1						
日平均気温の月平均値（℃）	0.9051543	1					
日最高気温の月平均値（℃）	0.9098449	0.9991393	1				
日最低気温の月平均値（℃）	0.9044283	0.9990259	0.9968021	1			
相対湿度の月平均値（%）	0.7463064	0.883057	0.8742303	0.8888456	1		
日照時間の月合計値（h）	0.1732483	-0.047349	-0.018293	-0.070142	-0.407772	1	
現地気圧の月平均値（hPa）	-0.719083	-0.642896	-0.64897	-0.632904	-0.518176	-0.221369	1



❷関係式を求める

データ > データ分析





概要

回帰統計	
重相関 R	0.90984486
重決定 R2	0.827817669
補正 R2	0.824849008
標準誤差	139.7817662
観測数	60

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	1	5448477.354	5448477.354	278.8522176	7.97191E-24
残差	58	1133258.646	19538.94217		
合計	59	6581736			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-136.6420492	50.55753595	-2.70270389	0.009006837	-237.843953	-35.4401456	-237.843953	-35.4401456
日最高気温の月平均値（℃）	38.48911904	2.3048938	16.69886875	7.97191E-24	33.87537283	43.10286526	33.87537283	43.10286526

残差出力

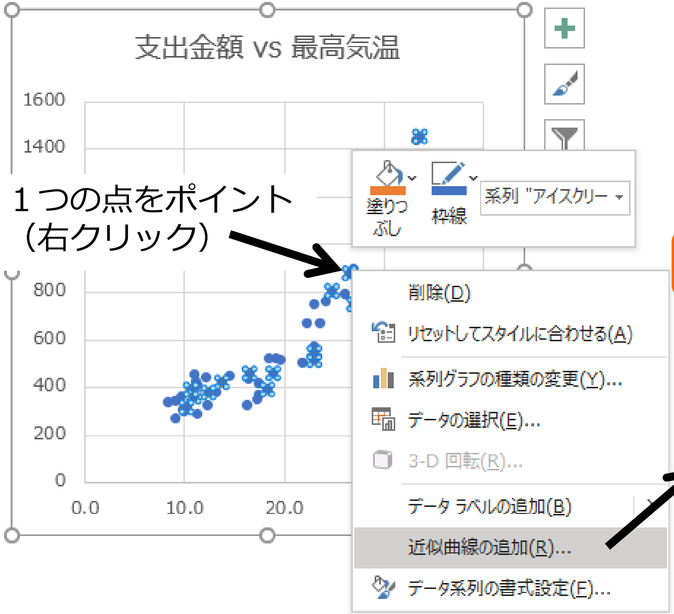
観測値	予測値: アイスクリームの支出金額(円)	残差
1	286.7382603	75.26173973
2	244.4002293	60.59977068
3	371.4143222	11.58567784
4	502.2773269	-38.2773269
5	748.6	

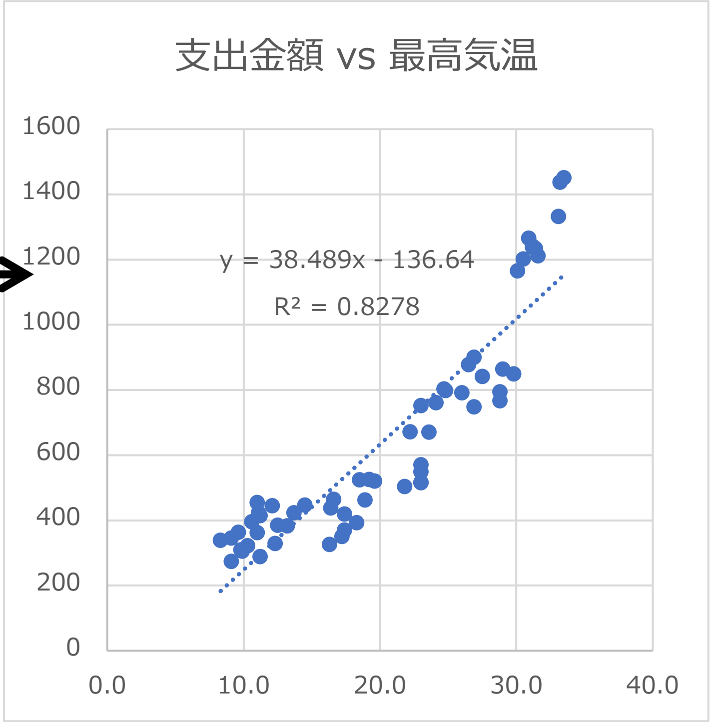
$$\hat{y} = -136.64 + 38.49x$$

アイスの金額 = −136.6 + 38.5 × 最高気温

❷関係式を求める

１つの点をポイント
（右クリック）





$$\hat{y} = -136.64 + 38.49x$$

アイスの金額 ＝ －136.6 ＋ 38.5 × 最高気温

❸-1関係式を評価する

概要

回帰統計	
重相関 R	0.90984486
重決定 R2	0.827817669
補正 R2	0.824849008
標準誤差	139.7817662
観測数	60

相関係数の2乗 ： **決定係数**

自由度調整済みの 決定係数

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	1	5448477.354	5448477.354	278.8522176	[illegible]
残差	58	1133258.646	19538.94217		
合計	59	6581736			

実測値と予測値の **相関係数**

関係式の **あてはまりのよさ**

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 9[illegible]
切片	-136.6420492	50.55753595	-2.70270389	0.00900[illegible]	[illegible]	[illegible]	-2[illegible]7.8[illegible]
日最高気温の月平均値（℃）	38.48911904	2.3048938	16.69886875	7.97191E-24	33.87537283	43.10286526	33.875[illegible]

残差出力

観測値	予測値: アイスクリームの支出金額(円)	残差
1	286.7382603	75.26173973
2	244.4002293	60.59977068
3	371.4143222	11.58567784
4	502.2773269	-38.2773269
5	748.6076888	3.392311198

年	月	アイスクリームの支出金額(円)	日平均気温の月平均値(℃)	日最高気温の月平均値(℃)	日最低気温の月平均値(℃)	相対湿度の月平均値(%)	日照時間の月合計値(h)	現地気圧の月平均値(hPa)
2010	1	362	7.0	11.0	3.0	41	221.9	1010.2
2010	2	305	6.5	9.9	3.0	60	118.3	1012.6
2010	3	383	9.1	13.2	5.1	61	139.8	1012.7
2010	4	464	12.4	16.6	8.6	62	139.9	1013.6
2010	5	752	19.0	23.0	15.4	60	198.8	1007.3
2010	6	841	23.6	27.5	20.5	67	162.5	1007.1
2010	7	1211	28.0	31.6	25.0	70	182.7	1005.9
2010	8	1451	29.6	33.5	27.0	67	222.6	1009.7
2010	9	864	25.1	29.0	22.3	68	165.3	1008.3
2010	10	504	18.9	21.8	16.5	68	81.4	1012.7
2010	11	351	13.5	17.2	10.0	56	158.9	1012.3
2010	12	423	9.9	13.7	6.2	50	194.9	1005.1
2011	1	346	5.1	9.1	1.5	36	243.9	1009.1
2011	2	289	7.0	11.2	3.2	52	148.9	1013.6
2011	3	329	8.1	12.3	4.0	47	214.8	1010.5
2011	4	462	14.5	18.9	10.3	50	204	1008.9
2011	5	672	18.5	22.2	15.0	63	146.3	1007.5
2011	6	791	22.8	26.0	19.9	71	105.1	1005.3
2011	7	1265	27.3	30.9	24.5	67	186.2	1004.5
2011	8	1241	27.5	31.2	24.6	71	168.9	1006.4
2011	9	767	25.1	28.8	22.0	68	165.8	1007.7
2011	10	516	19.5	23.0	16.5	61	141.3	1013.6
2011	11	393	14.9	18.3	11.7	58	143.4	1015.4
[illegible]	12	[illegible]	7[illegible]	11.[illegible]	4.2	48	[illegible]6	[illegible]

$$\hat{y} = -136.64 + 38.49x$$

❸-1関係式を評価する

概要

回帰統計	
重相関 R	0.90984486
重決定 R2	0.827817669
補正 R2	0.824849008
標準誤差	139.7817662
観測数	60

回帰の変動
全体の変動

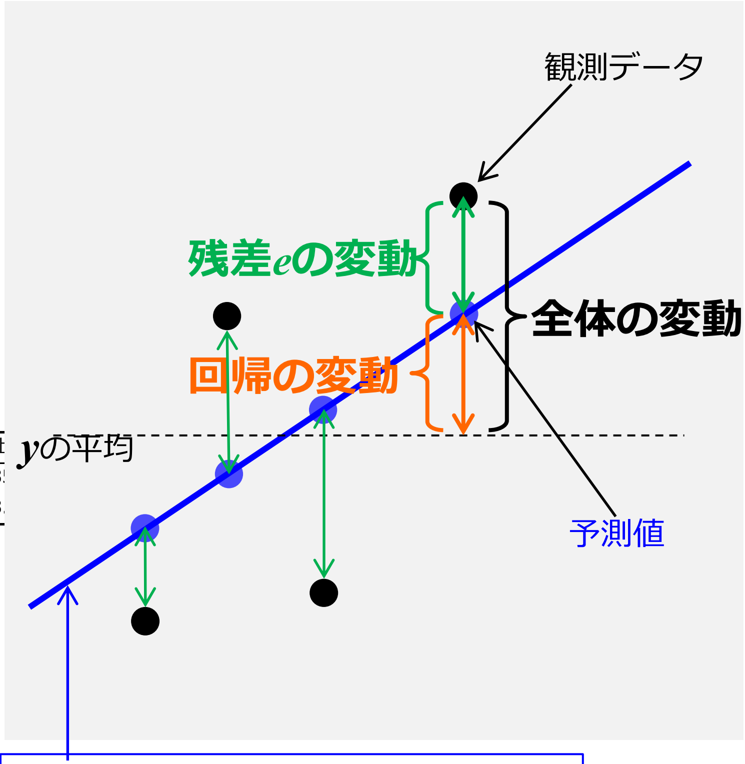
分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	5448477.354	5448477.354	278.8522176	7.97191E-24
残差	58	1133258.646	19538.94217		
合計	59	6581736			

	係数	標準誤差	t	P-値	下限 95%	上
切片	-136.6420492	50.55753595	-2.70270389	0.009006837	-237.843953	-3
日最高気温の月平均値（℃）	38.48911904	2.3048938	16.69886875	7.97191E-24	33.87537283	43.

残差出力

観測値	予測値: アイスクリームの支出金額(円)	残差
1	286.7382603	75.26173973
2	244.4002293	60.59977068
3	371.4143222	11.58567784
4	502.2773269	-38.2773269
5	748.6076888	3.392311198

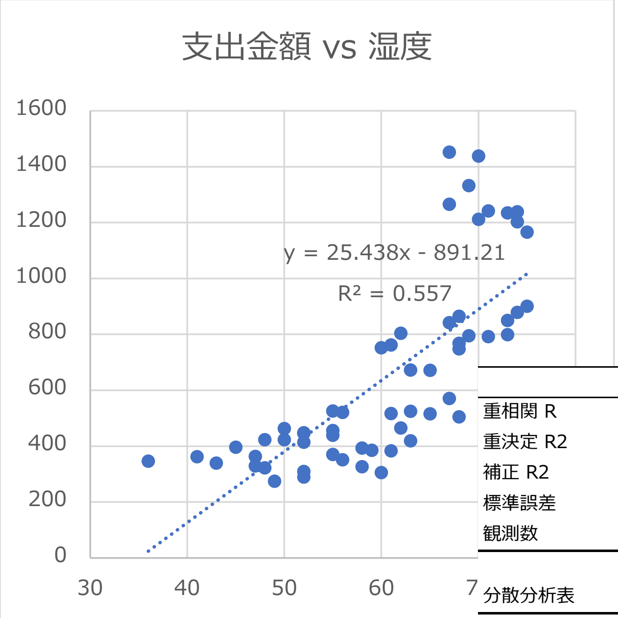


$$\hat{y} = -136.64 + 38.49x$$

$$y_i = -136.64 + 38.49x_i + e_i$$

❸-1関係式を評価する

たとえば、湿度の場合

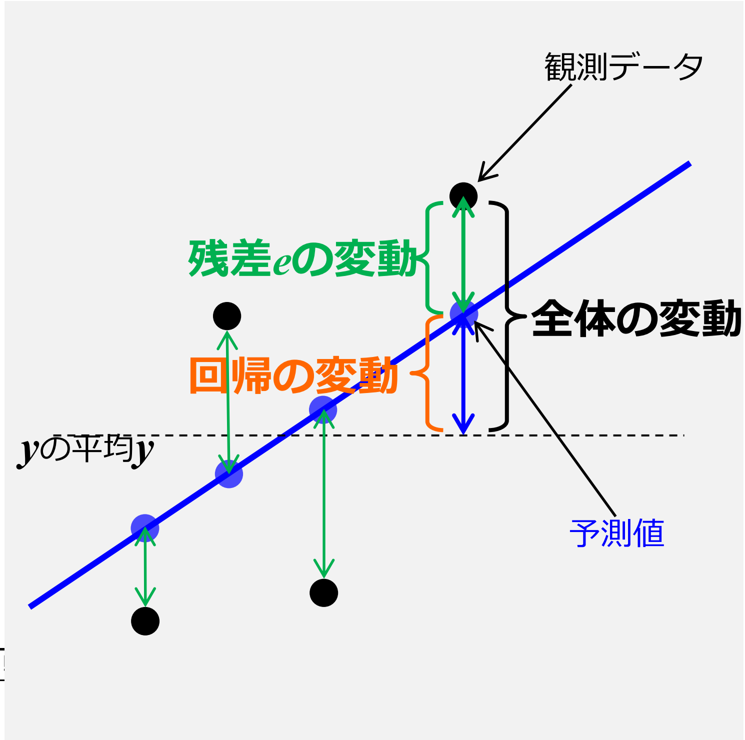


回帰統計	
重相関 R	0.74631
重決定 R2	0.55697
補正 R2	0.54933
標準誤差	224.218
観測数	60

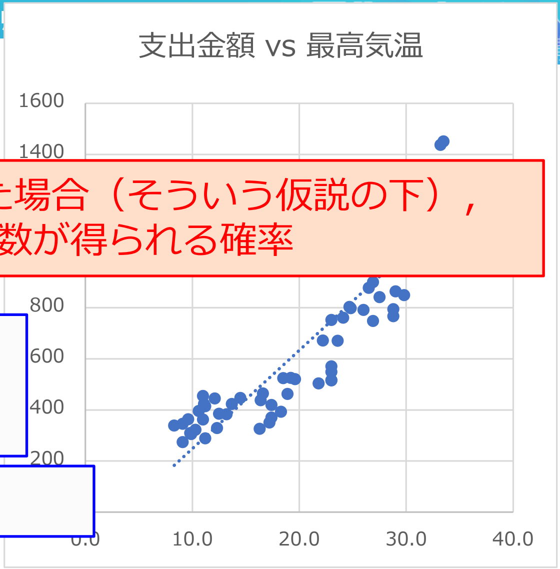
分散分析表

	自由度	変動	分散
回帰	1	3665851	3665851.1
残差	58	2915885	50273.877
合計	59	6581736	

	係数	標準誤差	t	P-値	下限 95%
切片	-891.21	183.025	-4.8693435	9E-06	-1257.6
相対湿度の月平均値（%）	25.4375	2.97892	8.539181	7.7E-12	19.4746



❸-2関係を解釈する



概要

回帰

重相関 R	
重決定 R2	
補正 R2	
標準誤差	139.7817
観測数	

回帰係数の t 値
単位を消す（重回帰のときがわかりやすい）
0 に近いと意味がない

回帰係数＝0であった場合（そういう仮説の下），今回のような回帰係数が得られる確率

~~回帰係数＝0 となる確率~~
0.05より小さいと 0 であることを否定＝回帰係数が有意
0.05より大きいと 0 であることを否定し切れない＝意味がない

分散分析表

	自由度			
回帰	1	5448477.354	5448477.354	2
残差	58	1133258.646	19538.94217	
合計	59	6581736		

真の回帰係数が存在する範囲

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-136.6420492	50.55753595	-2.70270389	0.009006837	-237.843953	-35.4401456	-237.843953	-35.4401456
日最高気温の月平均値（℃）	38.48911904	2.3048938	16.69886875	7.97191E-24	33.87537283	43.10286526	33.87537283	43.10286526

E-24は $\times 10^{-24}$のこと＝$7.97\times \times 10^{-24} \fallingdotseq 0$

残差出力

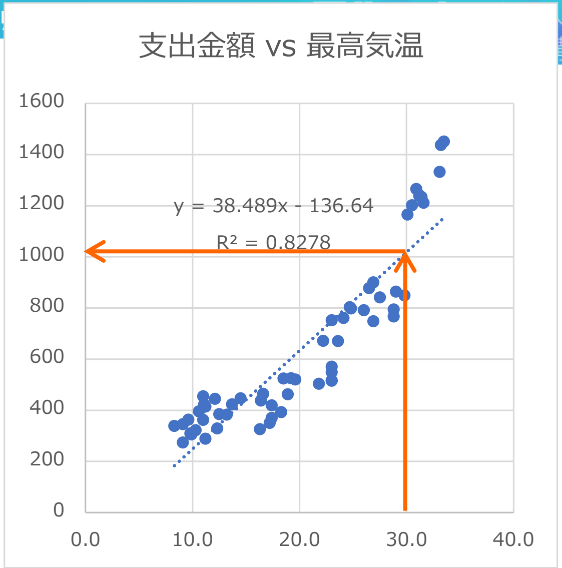
観測値	予測値: アイスクリームの支出金額(円)	残差
1	286.7382603	75.26173973
2	244.4002293	60.59977068
3	371.4143222	11.58567784
4	502.2773269	-38.2773269
5	748.6076888	3.392311198

最高気温が 0 ℃ のとき－136.64円（←理論的な話）
最高気温が 1 ℃ 上がると 38.49円 増える

$$\hat{y} = -136.64 + 38.49x$$

アイスの金額 ＝ －136.6 ＋ 38.5 × 最高気温

❸-3予測する



来月の最高気温が30℃だとすると

$-136.64 + 38.49 \times 30$

より、$y = 1018.03$、すなわち、

1018円と予想できる。

最高気温が 0 ℃ のとき －136.64円（←理論的な話）
最高気温が 1 ℃ 上がると 38.49円 増える

$$\hat{y} = -136.64 + 38.49x$$

アイスの金額 ＝ －136.6 ＋ 38.5 × 最高気温

回帰分析のまとめ

アイスの売り上げを左右する要因をさぐりたい。

回帰分析を実行

回帰統計	
重相関 R	0.90984486
重決定 R2	0.827817669
補正 R2	0.824849008
標準誤差	139.7817662
観測数	60

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	1	5448477.354	5448477.354	278.8522176	7.97191E-24
残差	58	1133258.646	19538.94217		
合計	59	6581736			

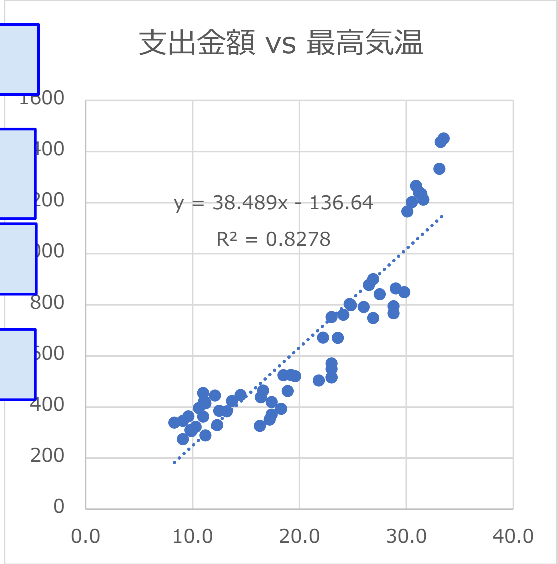
	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-136.6420492	50.55753595	-2.70270389	0.009006837	-237.843953	-35.4401456
日最高気温の月平均値（℃）	38.48911904	2.3048938	16.69886875	7.97191E-24	33.87537283	43.10286526

$$\hat{y} = -136.64 + 38.49x$$

１増えると 38.49 増えるな（０のとき－136.64）

あてはまりはよい。この式で問題なさそうだ。

p値が0.05より小さいので、yの説明において
この変数に意味があるぞ。



アイスの売り上げには最高気温が・・・
最高気温が○℃のときアイスの売り上げは・・・

【練習１】缶コーヒーの好感度

次のデータは、首都圏の中学生から59歳までの男性374人に、14種の缶コーヒーの銘柄それぞれについての意識や経験を聞き、各項目において「そう思う」と答えた人数を集計したものです。
このデータから、買いたいという意欲に最も影響があると思われる要因を特定し、その関係を明らかにしてください。

ブラインド名	購入意向	購入経験	おいしい	近くで買える	デザイン
ボス	94	288	131	206	82
エメラルドマウンテン	86	209	153	165	56
コーヒーオリジナル	67	277	120	191	30
ボスプラスワン	64	168	131	153	37
NOMO缶	60	180	112	131	64
J.O.スペシャルブレンド	56	209	112	157	26
サンタマルタ	56	183	120	127	22
UCC缶コーヒーオリジナル	52	224	108	142	19
アサヒカフェオ缶280	49	116	157	101	11
ポッカコーヒー	49	251	105	165	30
UCCスーパーオリジナル	45	191	120	127	19
UCCブラック無糖	45	194	108	131	15
ジョージアゾット	37	150	97	138	22
モンテアルバン	37	127	94	101	15

購入意向：
買いたいと思いますか
購入経験：
買ったことがありますか
おいしい：
おいしいと思いますか
近くで買える：
近くで買えますか
デザイン：
デザインはよいと思いますか

日経流通新聞（1996/12/14）

上田太一郎 (1998). データマイニング事例集、共立出版.

重回帰分析

2 広告宣伝費をどう調整するか

広告にかける費用が売上総額にどのように影響しているかを調べ、要因の重要度や次期の売上の予測を行ってみることにします。

データは、毎週の３つの広告（Web広告、チラシ宣伝、雑誌広告）にかける費用とその週の売上総額を2018年から139週分を集めたものです。

期	売上総額	Web広告費	チラシ宣伝費	雑誌広告費
2018-1	894,884	13,000	101,000	15,000
2018-2	806,523	11,000	66,000	20,000
2018-3	921,544	9,000	95,000	19,000
2018-4	791,293	7,000	84,000	18,000
2018-5	743,659	7,000	76,000	12,000
2018-6	931,192	16,000	109,000	19,000
2018-7	877,314	13,000	99,000	21,000
2018-8	903,283	16,000	75,000	20,000
2018-9	834,065	9,000	72,000	21,000
2018-10	822,309	9,000	90,000	12,000
2018-11	801,600	7,000	75,000	14,000
2018-12	804,382	10,000	93,000	19,000
2018-13	924,139	12,000	87,000	14,000
201[illegible]	[illegible]8,518	[illegible]	[illegible]3,000	[illegible]
[illegible]26	958[illegible]	[illegible],000	11[illegible]	[illegible],000
2020-27	1,013,900	22,000	121,000	19,000
2020-28	959,693	21,000	97,000	22,000
2020-29	975,074	23,000	102,000	26,000
2020-30	981,555	19,000	83,000	21,000
2020-31	974,691	18,000	92,000	19,000
2020-32	933,922	19,000	76,000	28,000
2020-33	937,789	15,000	103,000	19,000
2020-34	938,267	18,000	95,000	24,000
2020-35	1,001,564	19,000	94,000	22,000

回帰式を求め解釈する

❶関係を見つける

売上総額を広告宣伝費で
評価するとして

回帰統計	
重相関 R	0.899130899
重決定 R2	0.808436374
補正 R2	0.804179405
標準誤差	38929.48301
観測数	139

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	3	8.63423E+11	2.87808E+11	189.9088967	3.02142E-48
残差	135	2.04593E+11	1515504647		
合計	138	1.06802E+12			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	604411.1028	24836.58102	24.33551955	3.27362E-51	555291.9883	653530.2173
Web広告費	11.774537	0.975841021	12.06604022	3.37501E-23	9.844623751	13.70445025
チラシ宣伝費	1.163165652	0.322850929	3.602794812	0.000441466	0.524665863	1.801665441
雑誌広告費	0.708350384	1.043385662	0.67889603	0.49836596	-1.355145382	2.77184615

❷関係式を求める

$$\hat{y} = 604,411 + 11.77x_1 + 1.16x_2 + 0.71x_3$$

あてはまりはよい。

Web広告、チラシ宣伝のp値が0.05より小さいので、
yの説明においてこれらの変数に意味はある。
一方、雑誌広告は0.05より大きい。

❸-2関係を解釈する

各宣伝費に10,000円追加すると、それぞ117,700円、11,600円、7,100円増える。ただし、雑誌広告の売上総額への影響は小さいと考えられる。

Web広告に投下するのが最も効率がよいが、チラシ代理店とはつきあいも長いので･･･

❸-3予測する

売上総額100万円を超えたい。

- チラシ広告と雑誌を過去最低額（56,000円と9,000円）とするとWeb広告費は？
- 広告にかける金額を最も安くするには？

【練習２】新店舗の出店

新しい店舗を建てるため、3つの候補地があります。この３候補地の予測売上から、建設する候補地を絞りたいと思います。
既設15支店について、売上に関係すると思われる通行人数、最寄り駅からの時間、店舗面積、駐車台数、従業員数、品数の6つを調査したデータがあります。３候補地についても同じデータを得たので、これらから売上を予想してください。

支店	通行人（人）	駅からの時間（分）	店舗面積（m^2）	駐車台数（台）	従業員数（人）	品数（種）	売上（千円）
三条	716	25	338	16	5	19,200	680
京都南	2,208	10	315	0	7	20,400	1,200
長岡京	1,880	3	109	2	7	17,700	1,820
生駒	1,416	20	326	50	8	17,700	980
高槻	904	10	217	32	5	19,000	980
枚方	1,850	3	120	10	4	16,900	1,550
池田	1,039	15	219	15	7	19,900	1,200
東大阪	2,394	1	118	0	10	18,400	2,100
堺	711	12	214	12	6	20,400	560
八尾	738	10	202	10	6	19,600	650
和歌山	1,322	11	214	15	6	18,400	820
宝塚	793	18	352	30	4	20,600	320
西宮	1,733	3	111	5	4	20,300	1,900
西神	1,569	4	111	3	4	21,000	1,680
加古川	1,770	6	118	8	6	19,100	1,950

	通行人（人）	駅からの時間（分）	店舗面積（m^2）	駐車台数（台）	従業員数（人）	品数（種）
候補1	1956	3	120	0	6	18000
候補2	1275	12	223	30	8	20000
候補3	1556	8	198	5	6	18000

（改）http://mo161.soci.ous.ac.jp/@d/DoDStat/store/store_st01J.xml

カテゴリカルデータに対する回帰分析

3 発注精度を上げるには

発注ロスをできるだけ最小に抑えるように対策を考えることにします。

牛乳の販売本数は、その日の売値（単価）に加え、天気や曜日が関係していることが経験的にわかっていたので、そのデータを集めました。

このデータを用いて、牛乳の売上本数を予測してみましょう。

❶関係を見つける

本数	単価	天気	曜日
95	198	晴	木
74	198	晴	金
251	175	曇	土
167	175	曇	日
99	198	曇	月
216	175	曇	火
77	198	晴	水
159	188	晴	木
181	178	晴	金
212	175	晴	土
187	175	晴	日
79	198	晴	月
118	198	晴	火
59	198	晴	水
94	188	曇	金
255	175	晴	土
171	175	晴	日
90	198	曇	月
279	178	曇	金
143	175	雨	土
208	175	曇	日
69	198	雨	月
141	178	雨	火
97	198	曇	水
78	198	晴	木
189	178	晴	金

上田太一郎 (1998). データマイニング事例集、共立出版.

質的データを量的データに

本数	単価	天気	曜日
95	198	晴	木
74	198	晴	金
251	175	曇	土
167	175	曇	日
99	198	曇	月
216	175	曇	火
77	198	晴	水
159	188	晴	木
181	178	晴	金
212	175	晴	土
187	175	晴	日
79	198	晴	月
118	198	晴	火
59	198	晴	水
94	188	曇	金
255	175	晴	土
171	175	晴	日
90	198	曇	月
279	178	曇	金
143	175	雨	土
208	175	曇	日
69	198	雨	月
141	178	雨	火
97	198	曇	水
78	198	晴	木
189	178	晴	金

質的データ
定性的データ
カテゴリカルデータ
計数データ

意図的に**数量化**する

① カテゴリー１つに
１列（１変数）
② 該当変数に１、
他に０
（２値データ）

ダミー変数

晴	曇	雨
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
1	0	0
1	0	0
0	1	0
0	1	0
0	0	1
0	1	0
0	0	1
0	0	1
0	1	0
1	0	0
1	0	0

このままだと
情報過多
⇓
１列落とす

雨を落とす
００は雨

晴	曇
1	0
1	0
0	1
0	1
0	1
0	1
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
0	1
1	0
1	0
0	1
0	1
0	0
0	1
0	0
0	0
0	1
1	0
1	0

晴を落とす
００は晴

曇	雨
0	0
0	0
1	0
1	0
1	0
1	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
1	0
0	0
0	0
1	0
1	0
0	1
1	0
0	1
0	1
1	0
0	0
0	0

曇を落とす
００は曇

晴	雨
1	0
1	0
0	0
0	0
0	0
0	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	0
0	0
1	0
1	0
0	0
0	0
0	1
0	0
0	1
0	1
0	0
1	0
1	0

質的データを量的データに

本数	単価	天気	曜日
95	198	晴	木
74	198	晴	金
251	175	曇	土
167	175	曇	日
99	198	曇	月
216	175	曇	火
77	198	晴	水
159	188	晴	木
181	178	晴	金
212	175	晴	土
187	175	晴	日
79	198	晴	月
118	198	晴	火
59	198	晴	水
94	188	曇	金
255	175	晴	土
171	175	晴	日
90	198	曇	月
279	178	曇	金
143	175	雨	土
208	175	曇	日
69	198	雨	月
141	178	雨	火
97	198	曇	水
78	198	晴	木
189	178	晴	金

→

雨を落とす
００は雨

日を落とす
０００００００は日

本数	単価	晴	曇	月	火	水	木	金	土
95	198	1	0	0	0	0	1	0	0
74	198	1	0	0	0	0	0	1	0
251	175	0	1	0	0	0	0	0	1
167	175	0	1	0	0	0	0	0	0
99	198	0	1	1	0	0	0	0	0
216	175	0	1	0	1	0	0	0	0
77	198	1	0	0	0	1	0	0	0
159	188	1	0	0	0	0	1	0	0
181	178	1	0	0	0	0	0	1	0
212	175	1	0	0	0	0	0	0	1
187	175	1	0	0	0	0	0	0	0
79	198	1	0	1	0	0	0	0	0
118	198	1	0	0	1	0	0	0	0
59	198	1	0	0	0	1	0	0	0
94	188	0	1	0	0	0	0	1	0
255	175	1	0	0	0	0	0	0	1
171	175	1	0	0	0	0	0	0	0
90	198	0	1	1	0	0	0	0	0
279	178	0	1	0	0	0	0	1	0
143	175	0	0	0	0	0	0	0	1
208	175	0	1	0	0	0	0	0	0
69	198	0	0	1	0	0	0	0	0
141	178	0	0	0	1	0	0	0	0
97	198	0	1	0	0	1	0	0	0
78	198	1	0	0	0	0	1	0	0
189	178	1	0	0	0	0	0	1	0

→ **重回帰分析**

回帰式を求める

❷関係式を求める

$$\hat{y} = 1157.66 - 5.93\,x_1 + 58.06x_2 + 69.75x_3 + 51.98x_4 + 47.81x_5 + 32.83\,x_6 + 49.95x_7 + 34.72x_8 + 49.44x_9$$

概要

回帰統計	
重相関 R	0.928091656
重決定 R2	0.861354122
補正 R2	0.783365816
標準誤差	30.32261202
観測数	26

あてはまりはよい。

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	9	91396.16567	10155.12952	11.04465739	2.51692E-05
残差	16	14711.37279	919.4607996		
合計	25	106107.5385			

	係数	標準誤差	t	P-値		
切片	1157.657422	207.2318466	5.586291111	4.09487E-05		
単価	-5.933197899	1.192956147	-4.973525567	0.000138051		
晴	58.05667439	22.27240423	2.606664004	0.019080708		
曇	69.74774539	21.79141525	3.200698284	0.005569527		
月	51.97772028	36.18949258	1.436265517	0.170191232		
火	47.80511842	26.70177604	1.790334783	0.09233906	-8.800118105	104.4103549
水	32.82873018	35.55050347	0.923439248	0.369495579	-42.53497052	108.1924309
木	49.94842752	32.53411113	1.535263322	0.144255839	-19.02080706	118.9176621
金	34.71788819	22.88652923	1.516957326	0.148785833	-13.7993864	83.23516279
土	49.43693635	22.12260181	2.234680024	0.040052921	2.539115533	96.33475716

x_1はそのままでよい

x_2(晴)とx_3(曇)は元は同じアイテム
（どちらかが1だと他は0）
雨の係数は？　・・・0

x_4 ～ x_9も同様
日の係数は？　・・・0

$$\text{牛乳の本数} = 1157.66 - 5.93 \times \text{単価} + \begin{array}{c} \begin{pmatrix} 58.06 \\ 69.75 \\ 0 \end{pmatrix} \end{array}\begin{array}{l} \text{晴} \\ \text{曇} \\ \text{雨} \end{array} + \begin{pmatrix} 0 \\ 51.98 \\ 47.81 \\ 32.83 \\ 49.95 \\ 34.72 \\ 49.44 \end{pmatrix}\begin{array}{l} \text{日} \\ \text{月} \\ \text{火} \\ \text{水} \\ \text{木} \\ \text{金} \\ \text{土} \end{array}$$

回帰式を解釈する

❸-2関係を解釈する

概要

回帰統計	
重相関 R	0.928091656
重決定 R2	0.861354122
補正 R2	0.783365816
標準誤差	30.32261202
観測数	26

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	9	91396.16567	10155.12952	11.04465739	2.51692E-05
残差	16	14711.37279	919.4607996		
合計	25	106107.5385			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	1157.657422	207.2318466	5.586291111	4.09487E-05	718.3455328	1596.969312
単価	-5.933197899	1.192956147	-4.973525567	0.000138051	-8.462151958	-3.40424384
晴	58.05667439	22.27240423	2.606664004	0.019080708	10.84128664	105.2720621
曇	69.74774539	21.79141525	3.200698284	0.005569527	23.55200874	115.9434821
月	51.97772028	36.18949258	1.436265517	0.170191232	-24.74057683	128.6960174
火	47.80511842	26.70177604	1.790334783	0.09233906	-8.800118105	104.4103549
水	32.82873018	35.55050347	0.923439248	0.369495579	-42.53497052	108.1924309
木	49.94842752	32.53411113	1.535263322	0.144255839	-19.02080706	118.9176621
金	34.71788819	22.88652923	1.516957326	0.148785833	-13.7993864	83.23516279
土	49.43693635	22.12260181	2.234680024	0.040052921	2.539115533	96.33475716

$$\text{牛乳の本数} = 1157.66 - 5.93 \times \text{単価} + \left(\begin{array}{cc} 58.06 & \text{晴} \\ 69.75 & \text{曇} \\ 0 & \text{雨} \end{array}\right) + \left(\begin{array}{cc} 0 & \text{日} \\ 51.98 & \text{月} \\ 47.81 & \text{火} \\ 32.83 & \text{水} \\ 49.95 & \text{木} \\ 34.72 & \text{金} \\ 49.44 & \text{土} \end{array}\right)$$

単価を**１円安く**すると**約６本多く**売れる。

天候による効果は、雨の日を基準として（０として）、**曇の日**は約70本、晴の日は約58本多く売れる。

曜日による効果は、日曜日を基準として（０として）、**月**52本、火48本、水33本、木50本、金35本、土49本多く売れる。

❸-3予測する

- 明日は日曜日で、晴の予報、単価を180円にすると、予測売上本数は~~154本~~。

148本

【練習３】ペットに最適なリフォームプラン

ペットに最適なリフォームプランを提案するために、8つのリフォーム要因を設定し、各要因の条件をいろいろ変えた18種のプランについて、60人のモニターに10点満点で評価してもらいました。プラン内容とその評価（満足度）の平均点を記したものが下記です。これより、各要因で最も満足度が高くなる条件と、そのときの満足度はいくらになるかを調べてください。

業者選び	環境（有害物質）	プランニング	ペットへの配慮	近隣への配慮	期間と内容	アフターサービス	費用	満足度
ロコミ/チラシ	重要視する	自分で決めたい	衛生	ペットの毛	短期間にうまく	迅速対応を重視	40万円	6.67
ロコミ/チラシ	重要視する	相談して決める	健康	ペットの騒音	長くても丁寧に	納得いくまで	45万円	6.67
ロコミ/チラシ	重要視する	まかせる	リラックス	ペットの臭い	土日限定	有料もやむなし	50万円	1.67
ロコミ/チラシ	配慮する	自分で決めたい	衛生	ペットの騒音	長くても丁寧に	有料もやむなし	50万円	2.5
ロコミ/チラシ	配慮する	相談して決める	健康	ペットの臭い	土日限定	迅速対応を重視	40万円	8.33
ロコミ/チラシ	配慮する	まかせる	リラックス	ペットの毛	短期間にうまく	納得いくまで	45万円	4.17
ロコミ/チラシ	従来通り	自分で決めたい	健康	ペットの毛	土日限定	納得いくまで	50万円	0.83
ロコミ/チラシ	従来通り	相談して決める	リラックス	ペットの騒音	短期間にうまく	有料もやむなし	40万円	3.33
ロコミ/チラシ	従来通り	まかせる	衛生	ペットの臭い	長くても丁寧に	迅速対応を重視	45万円	3.33
インターネット	重要視する	自分で決めたい	リラックス	ペットの臭い	長くても丁寧に	納得いくまで	40万円	5.83
インターネット	重要視する	相談して決める	衛生	ペットの毛	土日限定	有料もやむなし	45万円	5.83
インターネット	重要視する	まかせる	健康	ペットの騒音	短期間にうまく	迅速対応を重視	50万円	5
インターネット	配慮する	自分で決めたい	健康	ペットの臭い	短期間にうまく	有料もやむなし	45万円	4.17
インターネット	配慮する	相談して決める	リラックス	ペットの毛	長くても丁寧に	迅速対応を重視	50万円	2.5
インターネット	配慮する	まかせる	衛生	ペットの騒音	土日限定	納得いくまで	40万円	4.17
インターネット	従来通り	自分で決めたい	リラックス	ペットの騒音	土日限定	迅速対応を重視	45万円	5
インターネット	従来通り	相談して決める	衛生	ペットの臭い	短期間にうまく	納得いくまで	50万円	2.5
インターネット	従来通り	まかせる	健康	ペットの毛	長くても丁寧に	有料もやむなし	40万円	2.5

この条件の組み合わせには直交表の考え方を使っています。このようにして調査をして分析する方法をコンジョイント分析といいます。

http://mo161.soci.ous.ac.jp/@d/DoDStat/ReformPet/ReformPet_st01J.xml

渕上美喜・上田太一郎・古谷都紀子 (2004). 実践ワークショップ・Excel徹底活用・ビジネスデータ分析、秀和ビジネス.

❗マーケティングにおける活用

回帰分析の特長

❶関係を見つける → ❷関係式を求める → ❸-1関係式を評価する / ❸-2関係を解釈する / ❸-3予測する

- 成果（結果）の要因（原因）を知ることができる。
- 複数の要因（原因）を用いて、成果（結果）を評価することができる。
- それぞれの要因（原因）の成果（結果）への影響度が判断できる。

- 数値で統計的な予測を立てることができる。
- 事象を説明する（プレゼンする）ための根拠となるデータを示すことができる。

扱っていないこと

変数選択
説明変数を多くすれば、決定係数が１に近づく。
すべての変数がyの説明に意味をもつわけではない。
変数が多いと解釈に困る。
⇓　説明力が高い少数の説明変数で、適合度の高いモデルを作る。
変数選択

多重共線性（multicollinearity、マルチコ）
説明変数どうしは本来独立であるべき。
説明変数間に高い相関･･･回帰係数の推定精度が極端に落ちる＝**多重共線性**
⇓　多重共線性を回避する。
互いの相関係数をチェック（一方を落とす）／変数選択

留意点

心構え

- 現象を記述することに終わらず、問題の解決や改善に利用する

× 関係式に表すだけ
○ ⇒売上向上、満足度向上、要因の特定

変数の解釈

- 変数が少なすぎる／変数が多すぎる
- 過去の経験も大切（ただ、柔軟に）

十分な要因が取り込まれているか確認
解釈が複雑になるなら減らす

要因の特定

- 見かけの相関にだまされない
- １回で終わらない
- 思った結果が出なくてもやめない
- あたり前が出たからといって捨てない

他の調査で裏をとる
相関関係は必ずしも因果関係ではない
（因果関係があれば相関関係がみられる）
分析をしては要因を取り替えまた分析を
それは根拠となる。勘の裏打ちでもある

計算・数値

- 回帰係数は相関係数ではない

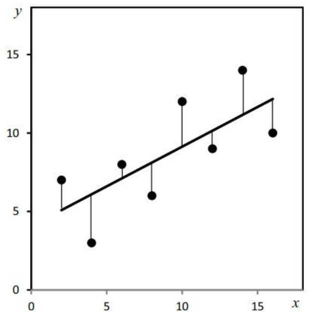
回帰分析・・・

要因の特定、改善、予測に活用を

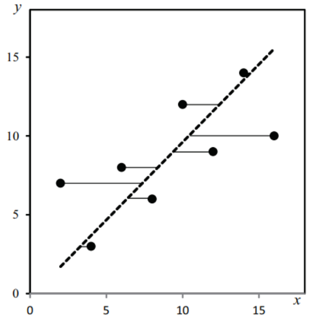
参考

３つの回帰直線

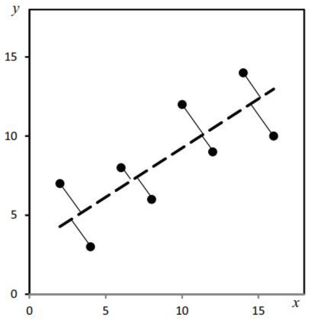
直線までの距離の二乗和を最小にする直線を考える（最小二乗法）場合、次の３通りが考えられる。
y方向のずれを考えているのが普通にいう「回帰分析」（yのxへの回帰）



y 軸方向のズレを最小にする直線（y の x への回帰：点が観測点，実線の直線が求めたい直線，その直線と各点を結ぶ y 軸に平行な線分がズレを表す．）



x 軸方向のズレを最小にする直線（x の y への回帰：点が図 1.9 と同じ観測点，点線の直線が求めたい直線，その直線と各点を結ぶ x 軸に平行な線分がズレを表す．）



直線までの距離を最小にする直線（直交回帰直線：点が図 1.9 と同じ観測点，破線の直線が求めたい直線，その直線に各点からおろした垂線がズレを表す．）

回帰とは、元は「xでyを上手に言い表す」、「点の集まりを直線近似する」ではなかった

「回帰」は「regression」の訳。 regression とは後退とか退化という意味である。19世紀、フランシス・ゴルトンが父親と息子の身長の関係を調べたとき、背の高い父親の息子の平均が父親より低く、背の低い父親の息子の平均は父親より高くなる、すなわち、父親の平均値に戻っていく（後退していく）ことを発見した。これは、x軸に父親の身長、y軸に息子の身長をとったとき、点の集まりが45度の直線より寝て、子孫（y）が親（x）の平均へ先祖返りする（y の x への回帰）ということで、この現象を「平均への回帰（regression to the mean）」とよんだ。 このことから、x と y の（直線）関係の考察で、 「回帰」の用語が用いられるようになった。
（背が高くなるのは遺伝的要素と偶然的要素があり、偶然的要素は全体の平均を凡庸にすることになる。血圧の高い人が食事療法により正常に、売上好調な会社群の次期の売上が下降、成績の悪かった群が次のテストで平均アップ、いずれも偶然的要素を排除しない場合は、平均へ回帰する。）