統計学(基礎)

第12回
重回帰分析

回帰分析(regression analysis)とは

- データをもとに、ある変数(従属変数または目的変数)を別の変数(独立変数または説明変数)で説明(予測)する式を作ること
  - 独立変数が1つだけの場合を単回帰分析
  - 複数の独立変数で1つの従属変数を予測する場合は重回帰分析

線形モデルと非線形モデル

- 一次式:線形モデル
  - 解釈が明確で汎用性が高い
  - モデル構造がシンプル
- 二次以上:非線形モデル
  - 複雑で過学習しやすく、意味づけが難しい
    - モデルの適合と仮説の検証が乖離することがある
  - 次数の決定は研究者次第
  - 調査系ではあまり採用しない

線形回帰(単回帰)

- 線形回帰　一次式

  $y = ax + b \quad (ax + by = c)$

  - aが回帰係数(回帰直線の傾き)、b が切片(x=0 のときのyの値)
  - xが独立変数(説明変数)、yが従属変数(目的変数)
- 一方の値でもう一方が説明(予測)できる
  - yがわからなくてもxがわかっていれば、式に当てはめて値が出る
    - これまでのxとyで式は作ってあるという場合
  - 実測値 － 予測値を「残差」という
  - 残差が小さいほど、予測式の精度は高い(精度が低くても式はできる)
  - 式が作れることと、その式に意味があるかは別問題

回帰式

$y = ax + b$ は以下で表される

$$y = r\frac{s_y}{s_x}x + \left(\bar{y} - \frac{s_{xy}}{{s_x}^2}\bar{x}\right)$$

$$a = r\frac{s_y}{s_x} = \frac{s_{xy}}{{s_x}^2} \quad b = \bar{y} - \frac{s_{xy}}{{s_x}^2}\bar{x} = \bar{y} - \bar{x}a$$

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$ ←共分散

回帰係数(*a*)＝相関係数(*r*)×(*Y*の標準偏差(*Sy*)÷*X*の標準偏差(*Sx*))

ちなみに

- 平均値　$\bar{x} = \frac{\sum_i (x_i)}{n}$　$\bar{y} = \frac{\sum_i (y_i)}{n}$
- 標本標準偏差　$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$　$s_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}$
- 相関係数　$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$
- 偏差積和　$\sum_i (x_i - \bar{x})(y_i - \bar{y})$
- 共分散　$s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

おまけ

- 標本標準偏差　$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$
- 不偏分散　$s_x{}^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$
- 偏差平方和　$\sum_i (x_i - \bar{x})^2$
- 偏差平方　$(x_i - \bar{x})^2$
- 偏差　$x_i - \bar{x}$

復習　散布図

- 2つの量的変数のグラフ
  - 同一のケースの2つの変数

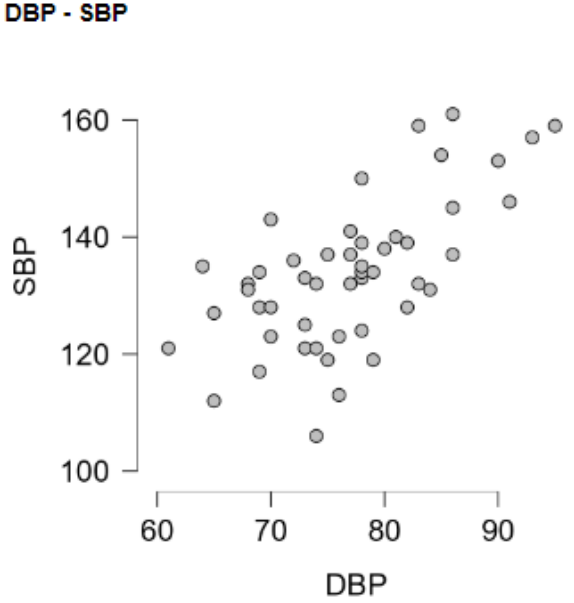
	No	DBP	SBP
1	1	78	134
2	2	86	161
3	3	91	146
4	4	75	119
5	5	86	137
6	6	64	135
7	7	78	124
8	8	73	121
9	9	72	136
10	10	80	138

記述統計 ▼

記述統計 ▼

	DBP	SBP
妥当	50	50
欠損値	0	0
平均値	76.90	133.8
標準偏差	7.541	12.47
最小値	61.00	106.0
最大値	95.00	161.0

散布図

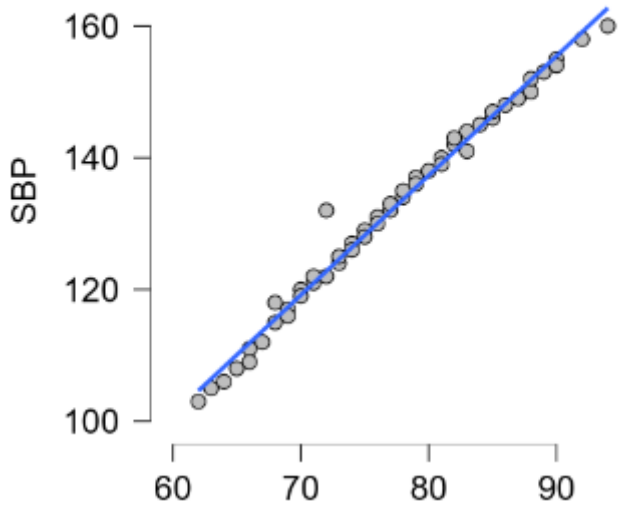


復習　相関

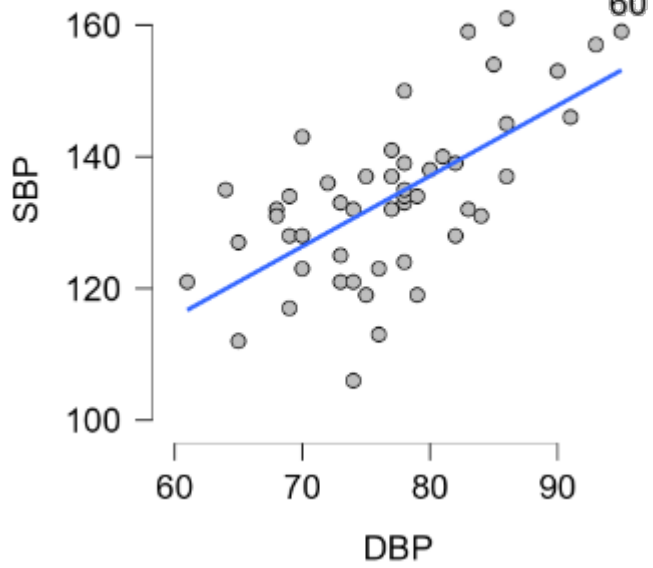
- 2変数間の関係
  - 量的な変数の関係
    - 順序尺度もできないわけじゃない
  - 直線的な関係
  - 一方が大きくなったときに、もう一方の大小がどうなるか
    - 片方が増えるともう片方も増える
    - 片方が増えるともう片方は減る
    - 片方の増減ともう片方の増減は関係ない

相関と回帰直線

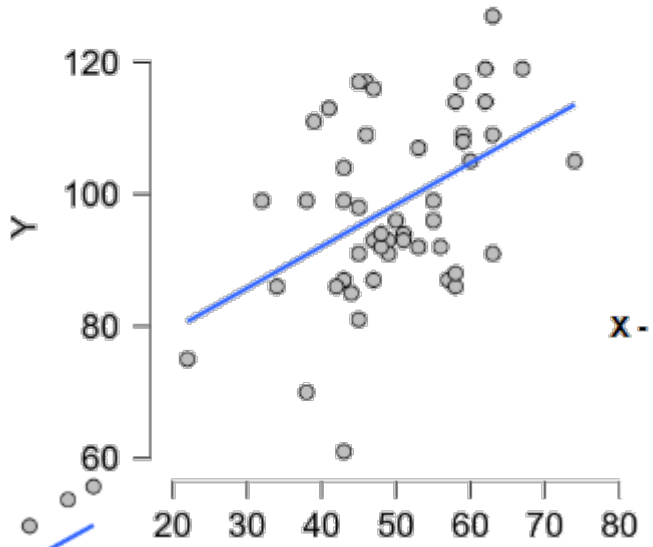
DBP - SBP r=0.93



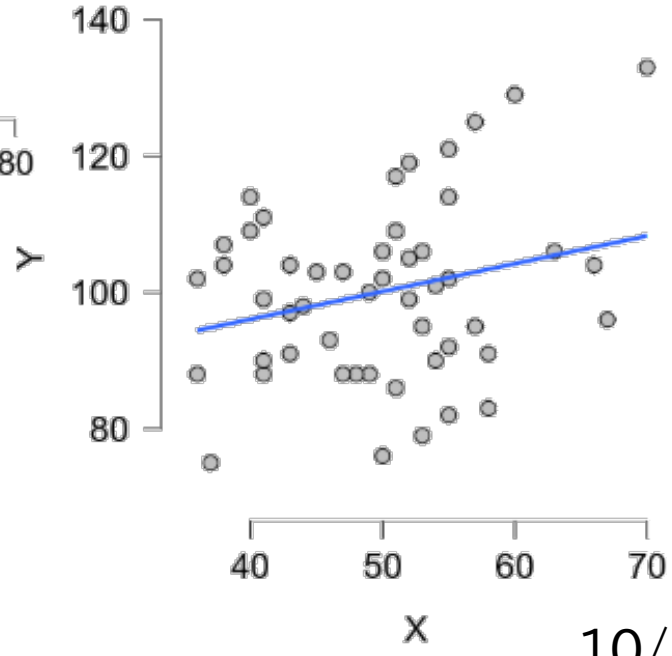
DBP - SBP r=0.65



X - Y r=0.45



X - Y r=0.25



回帰分析の出力

data12_01

線形回帰

モデルの概要 - Y

モデル	R	R^2	調整済み R^2	RMSE
M_0	0.000	0.000	0.000	21.850
M_1	0.980	0.960	0.959	4.425

注 M_1 includes X1

分散分析

モデル		平方和	df	平均平方	F	p
M_1	回帰	13296.4	1	13296.44	679.1	< .001
	残差	548.2	28	19.58		
	合計	13844.7	29			

注 M_1 includes X1

注 意味のある情報を表示できないため、切片モデルは省略されています。

係数

モデル		非標準化	標準誤差	標準化	t	p
M_0	(Intercept)	134.731	3.989		33.774	< .001
M_1	(Intercept)	5.489	5.025		1.092	.284
	X1	2.604	0.100	0.980	26.059	< .001

線形回帰

モデル適合度指標

モデル	R	R^2	調整済R^2
1	0.980	0.960	0.959

注. Models estimated using sample size of N=30

オムニバスANOVA検定

	2乗和	自由度	2乗平均	F	p
X1	13296	1	13296.4	679	< .001
残差	548	28	19.6		

注. タイプ3の2乗和を使用しています

[3]

モデル係数 - Y

予測変数	推定値	標準誤差	t	p	標準化推定値
切片	5.49	5.0249	1.09	0.284	
X1	2.60	0.0999	26.06	< .001	0.980

回帰分析の出力の見方

1. 決定係数(R^2)でモデルの適合度を見る
2. 分散分析表で適合度が統計的に有意かどうか判断する
3. 係数表で各独立変数の係数が統計的に有意かどうか判断する

モデルの概要(適合度)

- 作成されたモデルのあてはまりの良さ
  - R:重相関係数　実測値と予測値の相関
  - R^2:決定係数　独立変数全体の従属変数の説明割合
  - 調整済みR^2:独立変数の個数で調整したR^2
  - RMSE:2乗平均平方根誤差(Root Mean Square Error)

モデルの概要 - Y

モデル	R	R^2	調整済み R^2	RMSE
M_0	0.000	0.000	0.000	21.850
M_1	0.980	0.960	0.959	4.425

注 M_1 includes X1

モデル適合度指標

モデル	R	R^2	調整済R^2
1	0.980	0.960	0.959

注. Models estimated using sample size of N=30

分散分析(ANOVA)表

- 決定係数(R^2)の有意性の検定
  →「モデル全体が有効か」を調べる検定
  - 回帰平方和:説明できた変動
  - 残差平方和:説明できなかった変動
  - 平均平方:各平方和を自由度で除した商
  - F:回帰の平均平方を残差の平均平方で除した商

分散分析 ▼

モデル		平方和	df	平均平方	F	p
M_1	回帰	13296.4	1	13296.44	679.1	< .001
	残差	548.2	28	19.58		
	合計	13844.7	29			

オムニバスANOVA検定

	2乗和	自由度	2乗平均	F	p
X1	13296	1	13296.4	679	< .001
残差	548	28	19.6		

係数(Coefficients)表

- 各説明変数が有意かどうか(0でないかどうか)を検定

係数

モデル		非標準化	標準誤差	標準化	t	p
M_0	(Intercept)	134.731	3.989		33.774	< .001
M_1	(Intercept)	5.489	5.025		1.092	.284
	X1	2.604	0.100	0.980	26.059	< .001

モデル係数 - Y

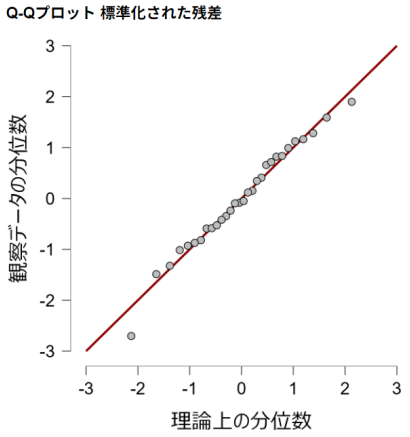
予測変数	推定値	標準誤差	t	p	標準化推定値
切片	5.49	5.0249	1.09	0.284	
X1	2.60	0.0999	26.06	< .001	0.980

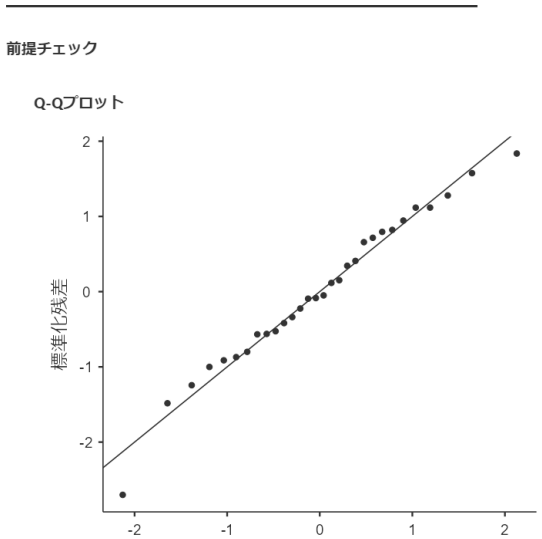
標準化係数(standardized β)

- 標準化回帰係数(標準化偏回帰係数)
  - 単位を揃えたときの(偏)回帰係数
    - 標準化(z変換)した(偏)回帰係数
  - 独立変数の「影響度」の指標
  - 絶対値が大きいほど、目的変数Yへの影響が強い
  - 重回帰では「どの変数が効いているか」を見る

補足

- 回帰分析は残差の正規性が必要といわれてる
  - n>30なら大抵は大丈夫(中心極限定理が成立するから)
  - 必要ならQ-Qプロットで確認するぐらいでいい





重回帰分析で問題になること

重回帰分析の式

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$
  - Y　目的変数(従属変数)
  - $X1, X2, \cdots, Xp$　説明変数(独立変数)
  - $\beta 0$　切片(定数項)
  - $\beta 1, \beta 2, \cdots, \beta p$　各説明変数の回帰係数(偏回帰係数)
    →それぞれが「他の変数が一定のとき、Xjが1単位増えると Yがどれだけ変化するか」を表す
  - ε　誤差項(モデルが説明しきれない部分)

重回帰分析で問題になること

- 切片の指定
- 多重共線性(Multicollinearity:マルチコ)
- 変数の選択

切片(Intercept)の扱い

- 切片は「説明変数がすべて0のときの予測値」
  - 実質的な意味が弱いことが多い
    - 独立変数がゼロの時に、従属変数が絶対的にゼロであるという強力な理論的仮定がある場合にのみ成立
  - 切片有りで式を作るのが普通
    - モデルがゆがむことが多い

多重共線性(Multicollinearity)

- 説明変数の中に、相関係数が高い組み合わせがあると発生する
  - 同じ内容を違う方法で測定している
  - 実質的に同じ内容の変数
- 同じ内容が複数説明変数に入ると、影響がおかしくなる
  - 実質的に同じものに別々の係数が発生する
  - データのわずかな変化が影響しやすい
- 共線性に関係なく、説明変数が多いほど決定係数は大きくなる

多重共線性のチェック

- 検出許容値(tolerance) $1 - R_j^2$
  - R_j^2:その説明変数が他の説明変数でどれだけ説明できるか
    - 1に近い　その説明変数が他の説明変数で説明できる
    - 0に近い　その説明変数は独自成分が高い
- tolerance
  - 1に近い　その説明変数は独自成分が高い
  - 0に近い　その説明変数が他の説明変数で説明できる
    →共線性が高い

多重共線性のチェック

- VIF(Variance Inflation Factor:分散拡大係数)

$$\frac{1}{tolerance}$$

  - 5以上　共線性の傾向が見られる(結構ヤバい)
  - 10以上　共線性が高い(完全にヤバい)
  - 許容されるのは2～3以下と言われている

多重共線性を抑えるために

- 似た変数は代表1つ
  - 尺度の下位項目を全部入れない
- VIFを確認
  - R^2だけを重視しない
  - R^2が高くても多重共線性がある場合もある
    - R^2とマルチコは基本独立

変数選択

- 統計アプリには変数の自動選択機能(ステップワイズ)が備わっている
  - 説明変数が多いときに使われるが
  - 数式的なモデル適合しか考えていない
- 重回帰分析は、変数が多いほどR^2が高くなる
  - R^2は独立変数全体の従属変数の説明割合
  - R^2が高いことと仮定モデルが説明できることは違う

ステップワイズ法(変数選択)

- 方法
  - 前方(Forward):何も入れずにスタートし、有意な変数を追加
  - 後方(Backward):全て投入してスタートし、有意でない変数を削除
  - ステップワイズ(Stepwise):追加と削除を繰り返す
- メリット
  - 変数が多いとき、機械的に「それっぽいモデル」を作ってくれる
  - 手作業での組み合わせ検討が不要になる

ステップワイズ法の本質的な問題

- 見かけ上便利だが、研究目的とズレる危険が大きい
  1. アルゴリズムの目的が研究目的と異なる
     - データへの最適な当てはめが目的(R^2とかを最適にする)
  2. 説明したい変数が残らないことがある
     - 理論的に重要な変数が外れる可能性がある
     - 説明したい変数が入らない → 研究として意味を欠く
  3. 過剰適合や p値の信頼性低下
     - 決定係数が「よく見える」だけのモデルになる
     - 別データで再現しない不安定なモデルになる

ステップワイズをもし使うなら

- 基本は「研究仮説」に基づいて変数を選ぶ
  - ステップワイズは補助的な探索としての利用にとどめる
    - 自動で変数を選んでくれるが、責任は取らない
  - 必要な説明変数は強制的に投入するしかない
- ステップワイズは過剰適合になりやすい
  - 投入したデータへの適合しか考えていないので、仮説は考えてない
  - 理論ではなく「データ任せ」のモデルになる
  - サンプルが少し変わると、選ばれる変数が変わり不安定

過学習(オーバーフィッティング)とは

- 投入した「データのクセ」に合わせすぎてしまう現象
  - 本質ではなく誤差(ノイズ)までモデル化してしまう
  - R^2は上がるが、仮説モデル説明という科学的意味は失われる

そもそも論で

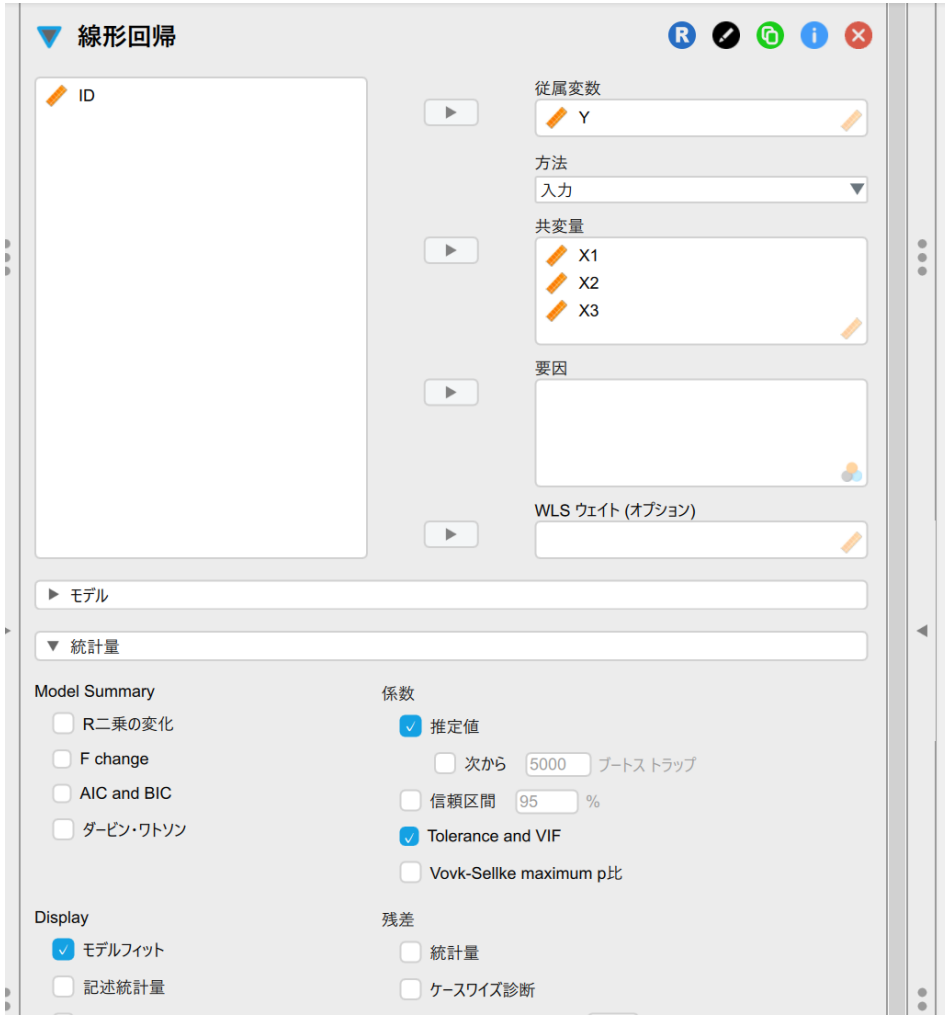
- 重回帰にそんなに説明変数が必要か
  - モデルとして、説明変数の影響を説明できなければ意味が無い
  - 説明変数が多いということはモデルが複雑になる
  - 説明変数が多いと、測定データのみに当てはまるモデルになりやすい
- 基本的にはシンプルに説明できるモデルの作成が望ましい

データはdata12_02

重回帰分析実行の実際

重回帰分析(JASP)

data12_02



結果

線形回帰

モデルの概要 - Y

モデル	R	R^2	調整済み R^2	RMSE
M_0	0.000	0.000	0.000	17.006
M_1	0.951	0.904	0.893	5.550

注 M_1 includes X1, X2, X3

分散分析

モデル		平方和	df	平均平方	F	p
M_1	回帰	7585.8	3	2528.60	82.08	< .001
	残差	801.0	26	30.81		
	合計	8386.8	29			

注 M_1 includes X1, X2, X3

注 意味のある情報を表示できないため、切片モデルは省略されています。

係数

							共線性統計	
モデル		非標準化	標準誤差	標準化	t	p	許容度	VIF
M_0	(Intercept)	190.569	3.105		61.378	< .001		
M_1	(Intercept)	25.289	15.290		1.654	.110		
	X1	1.667	0.166	0.698	10.040	< .001	0.760	1.316
	X2	0.812	0.092	0.690	8.819	< .001	0.601	1.665
	X3	−8.470	1.368	−0.429	−6.190	< .001	0.765	1.308

重回帰分析(jamovi)

モデル適合度指標

モデル	R	R^2	調整済R^2
1	0.951	0.904	0.893

注. Models estimated using sample size of N=30

オムニバスANOVA検定

	2乗和	自由度	2乗平均	F	p
X1	3105	1	3105.3	100.8	< .001
X2	2396	1	2396.0	77.8	< .001
X3	1180	1	1180.4	38.3	< .001
残差	801	26	30.8		

注. タイプ3の2乗和を使用しています

[3]

モデル係数 - Y

予測変数	推定値	標準誤差	t	p
切片	25.289	15.2904	1.65	0.110
X1	1.667	0.1661	10.04	< .001
X2	0.812	0.0921	8.82	< .001
X3	−8.470	1.3683	−6.19	< .001

前提チェック

共線性統計量

	VIF	トレランス
X1	1.32	0.760
X2	1.66	0.601
X3	1.31	0.765

[3]

jamoviの重回帰の注意点

- 調整済みR^2がオプション指定
  - 「モデル適合度」で指定
- 分散分析表もオプション指定
  - 「モデル係数」で指定

VIFの指定

- JASP、jamoviともにオプション
  - JASP 「統計量」のTolerance and VIF
  - jamovi 「前提チェック」の共線性統計量

重回帰の係数表(JASP)

係数

モデル		非標準化	標準誤差	標準化	t	p	共線性統計	
							許容度	VIF
M_0	(Intercept)	190.569	3.105		61.378	< .001		
M_1	(Intercept)	25.289	15.290		1.654	.110		
	X1	1.667	0.166	0.698	10.040	< .001	0.760	1.316
	X2	0.812	0.092	0.690	8.819	< .001	0.601	1.665
	X3	−8.470	1.368	−0.429	−6.190	< .001	0.765	1.308

全体のモデル適合とマルチコは独立

data12_03

モデルの概要 - Y

モデル	R	R^2	調整済み R^2	RMSE
M_0	0.000	0.000	0.000	51.531
M_1	0.986	0.972	0.969	9.081

注 M_1 includes X1, X2, X3

分散分析

モデル		平方和	df	平均平方	F	p
M_1	回帰	74865	3	24954.97	302.6	< .001
	残差	2144	26	82.46		
	合計	77009	29			

注 M_1 includes X1, X2, X3

注 意味のある情報を表示できないため、切片モデルは省略されてい

係数 ▼

モデル		非標準化	標準誤差	標準化	t	p	共線性統計	
							許容度	VIF
M_0	(Intercept)	198.958	9.408		21.147	< .001		
M_1	(Intercept)	−2.938	7.468		−0.393	.697		
	X1	2.755	0.452	0.906	6.096	< .001	0.048	20.626
	X2	0.243	0.408	0.089	0.596	.557	0.048	20.800
	X3	−2.944	1.723	−0.057	−1.709	.099	0.961	1.040

ノンパラメトリックな回帰分析

ロジスティック回帰

ロジスティック回帰とは？

- ロジスティック回帰は、結果が 0または1 のときに使う回帰モデル
  - 2値データだと、0、1に置き換えられる
  - 確率として扱える
- 独立変数は「数値に変換できれば」利用可能
  - ということで平均値を使わないからノンパラメトリックな手法
- 名義変数はそのままでは使えない → ダミー変数が必要

ロジスティック回帰とは？

- 確率は 0～1 の範囲
  - 線形回帰は －∞～＋∞ を取る
- 一度「オッズ」に変換して、さらに自然対数を使って確率を －∞～＋∞ に引き延ばす
  - 線形回帰の形で扱える
  - オッズ:ある事象が生起する確率と生起しない確率の比

ロジスティック回帰とは？

- 結果は偏回帰係数よりも、オッズ比を見る
- オッズ比
  - 他の独立変数が変化しないときに、その独立変数が1単位変化すると、従属変数が1になる確率が何倍になるか

オッズ比で係数を解釈

- オッズ比の解釈
  - オッズ比 ＞ 1　確率が上がる
  - オッズ比 ＜ 1　確率が下がる

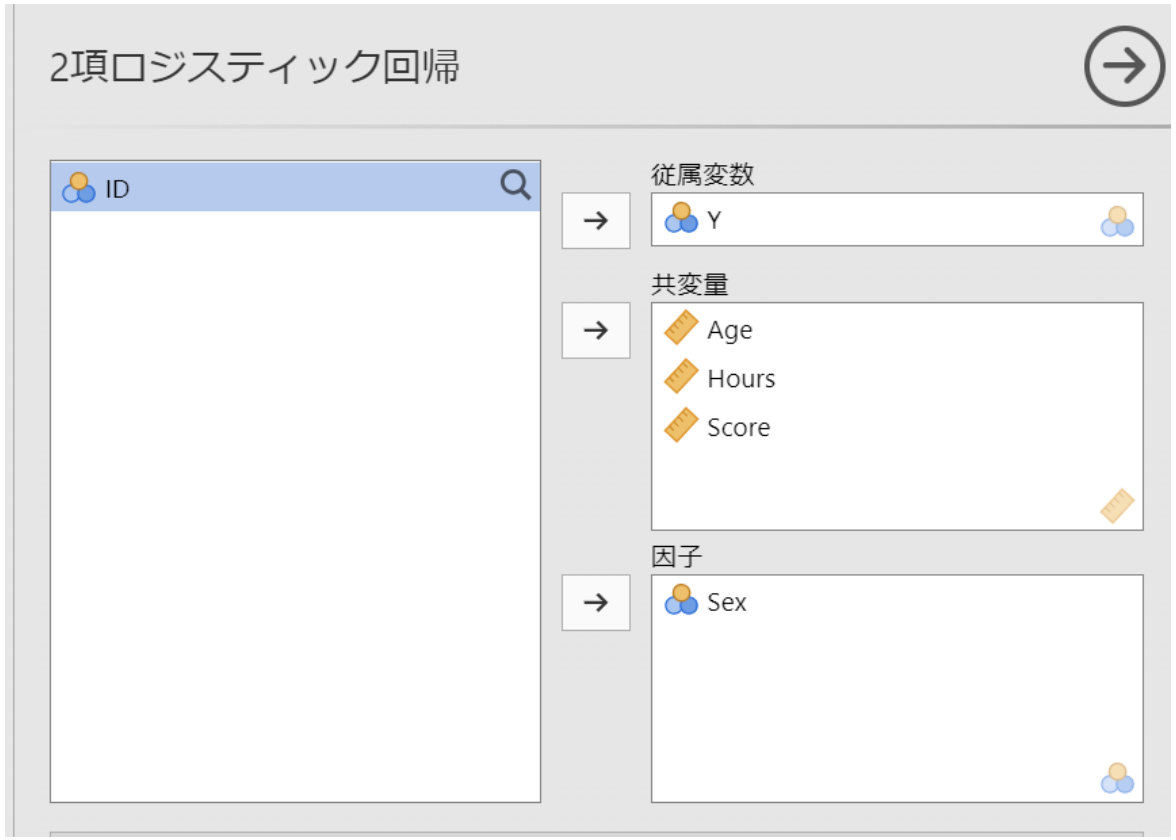
  - オッズ比が1だと変化がないことになる

ロジスティック回帰分析の出力の見方

1. 疑似決定係数(R^2)や適合指標でモデルの適合度を見る
2. 係数表で各独立変数の係数が統計的に有意かどうか判断する
3. 予測の適合度の指標を見てモデルが有効かどうか判断する

ロジスティック回帰の指定(JASP)

ロジスティック回帰の指定(jamovi)



モデルの概要

- 逸脱度(deviance)　AIC(Akaike's Information Criterion) BIC(Bayesian Information Criterion)→ 小さい方が当てはまりがいい
- 疑似決定係数→　尤度から計算　1に近い方がいい

モデルの概要 - Y

モデル	デビアンス	AIC	BIC	df	ΔX²	p	McFadden R²	Nagelkerke R²	Tjur R²	コックス＆スネルR ²
M_0	51.80	53.796	55.485	39			0.000		0.000	
M_1	39.10	49.096	57.540	35	12.700	.013	0.245	0.375	0.285	0.272

注 M_1 includes Sex, Score, Hours, Age

モデル適合度指標

モデル	逸脱度	AIC	BIC	$R^2_{マク}$	$R^2_{コックス}$	$R^2_{ナゲ}$	R^2_T
1	39.1	49.1	57.5	0.245	0.272	0.375	0.285

ノート. Models estimated using sample size of N=40

data12_04

係数

係数

モデル		推定	標準誤差	オッズ比	z	Wald Test			95% 信頼区間 オッズ比スケール)	
						Wald統計	df	p	下限	上限
M_0	(Intercept)	−0.619	0.331	0.538	−1.867	3.487	1	.062	0.281	1.031
M_1	(Intercept)	−16.033	8.030	1.089×10^{-7}	−1.997	3.987	1	.046	0.000	0.744
	Sex (2)	−2.445	1.048	0.087	−2.332	5.440	1	.020	0.011	0.677
	Score	0.180	0.120	1.198	1.497	2.240	1	.134	0.946	1.516
	Hours	0.617	0.463	1.854	1.333	1.778	1	.182	0.748	4.592
	Age	0.090	0.125	1.095	0.725	0.525	1	.469	0.857	1.398

注 Y水準の「1」がクラス1としてコード化されています。

モデル係数 - Y

予測変数	推定値	標準誤差	Z	p	オッズ比	95%信頼区間	
						下限	上限
切片	−16.0332	8.030	−1.997	.046	1.09e−7	1.59e−14	0.744
Sex:							
2 − 1	−2.4447	1.048	−2.332	.020	0.0867	0.0111	0.677
Age	0.0905	0.125	0.725	.469	1.0947	0.8572	1.398
Hours	0.6171	0.463	1.333	.182	1.8536	0.7483	4.592
Score	0.1803	0.120	1.497	.134	1.1976	0.9457	1.516

ノート. 推定値は 「Y = 1」 vs. 「Y = 0」 の対数オッズです

共線性診断

- 考え方は重回帰と同じ

多重共線性診断

	許容度	VIF
Sex	0.634	1.577
Score	0.661	1.512
Hours	0.843	1.186
Age	0.904	1.107

前提チェック

共線性統計量

	VIF	トレランス
Sex	1.58	0.634
Age	1.11	0.904
Hours	1.19	0.843
Score	1.51	0.661

[3]

予測の適合度について

- 重回帰と違って残差が出ない
  - 従属変数は0、1だから
- 実測と予測のクロス表(混合行列:confusion matrix)を作成して、的中度、感度、特異度や精度を求めて検討

予測の適合度

- AUC(Area of Under the Curve　ROC曲線下面積)
  →0.5～1.0　高い方がいい
- Accuracy(正確度)
  →的中率
- Sensitivity(感度)
- Specificity(特異度)
- Precision(精度)
  →陽性反応的中度

パフォーマンス診断 ▼

混同行列

観測された	予測 0	予測 1	% Correct
0	23	3	88.46
1	7	7	50.00
Overall % Correct			75.00

注 The cut-off value is set to 0.5

パフォーマンス・メトリクス

	値
Accuracy	0.750
AUC	0.786
Sensitivity	0.500
Specificity	0.885
Precision	0.700

予測

分類表 – ...

観測度数	予測値 0	予測値 1	正判別%
0	23	3	88.5
1	7	7	50.0

ノート. カットオフ値 = 0.5

予測指標

精度	特異度	感度	AUC
0.750	0.885	0.500	0.786

ノート. カットオフ値 = 0.5

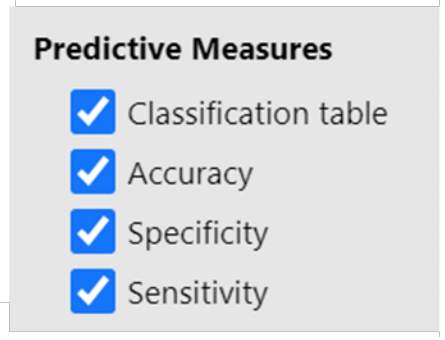
注意点

- jamoviの「予測指標」の「精度」は「正確度(的中率)」
  - 英語版はAccuracyでPrecisionではない
  - Accuracy(正確度)→的中率
  - Precision(精度)→陽性反応的中度

予測指標

精度	特異度	感度	AUC
0.750	0.885	0.500	0.786

ノート. カットオフ値 = 0.5



Predictive Measures

Accuracy	Specificity	Sensitivity	AUC
0.750	0.885	0.500	0.786

Note. The cut-off value is set to 0.5

おまけ

- jamoviが間違っているわけではない
  - JIS Z 8101「統計ー用語及び記号」では、総合精度＝accuracy
  - JIS Z 8103 「計測用語」では、精度＝accuracy
- 統計分野では正確度＝accuracy、精度＝precisionとしていることが多い

ちょっとだけ

ダミー変数の説明

名義変数がそのまま使えない理由

- 名義変数(例:A/B/C)は大小関係がない
  - そのまま投入すると誤ったモデルになる
  - そもそも文字データは分析できない
- 0/1 に変換して扱う(ダミー変数)
  - 例:喫煙(あり=1、なし=0)

複数カテゴリと基準カテゴリの考え方

- カテゴリ数−1　だけダミー化する
  - 例:3カテゴリ(A/B/C)の場合:ダミーは 2 個だけ作る
    - 基準カテゴリ(入れないカテゴリ)との比較を行う
    - 基準カテゴリ = C
    - D1:A=1, その他=0
    - D2:B=1, その他=0

なぜ全部入れてはいけないのか

- 全部入れると情報が重複する
  - 多重共線性が発生する
    - D1:A=1,その他=0
    - D2:B=1,その他=0
    - D3:C=1,その他=0

  →D1「0」、D2「0」ならD3は必ず「1」
  - 多重共線性の発生(ダミートラップ)

まとめ

- 回帰分析の本質は、「データに合わせる」ものではなく、理論的仮説を数式化し検証する方法
  - 何を説明したいのか(目的変数)を明確にする
  - 説明変数は仮説から吟味する
  - オーバーフィッティングに注意
- 分析の目的は説明すること
  - シンプルな方が理解がしやすい