

保健統計学実習

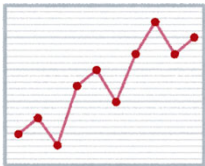
第3日目

- 第7回 重回帰分析
第8回 ロジスティック回帰、検査データの解析
第9回 調査データ解析(2):(web調査ツールを使用した)調査票作成、データ入力

滋賀医科大学NCD疫学研究センター
医療統計学部門

原田 亜紀子
(aharada@belle.shiga-med.ac.jp)

第7回 重回帰分析



3

講義・演習スケジュール

1 R, EZRの使い方、データセットの読み込み、頻度集計、記述統計、相関 2 EZRのコード保存, R-studio commander 3 エクセルの基礎(1)	8/29(木)
4 仮説検定の基礎, 2群の比較(t検定, Wilcoxon検定) 5 カイニ乗検定、マクネマー検定 6 調査データ解析(1): 調査票作成、データ入力	8/30(金)
7 重回帰分析 8 ロジスティック回帰、検査データの解析 9 調査データ解析(2):(web調査ツールを使用した)調査票作成、データ入力	9/2(月)
10 分散分析 11 サンプルサイズ 12 調査データ解析(3): 解析用データの作成	9/5(木)
13 主成分分析、因子分析、クラスター分析 14 解析実習・まとめ(復習・課題の時間)	9/6(金)

2

解析の手順 線形モデルのあてはめとモデル診断

なぜ線形モデルを使うのか？

- データ削減
- 予測式を算出したい
- パラメータ値の解釈(従属変数に対する独立変数の影響)

決定係数、残差の検討を丁寧に行う

仮定は正しいか？

- $y = x + \text{誤差}$
- 誤差の変動: 平均は0, 分散は等しい, 独立, ほぼ正規分布

偏重回帰係数の有意性を重視してモデルを作る

回帰診断

- 残差の検討と影響度分析

4

線形回帰

- xに対するyの線形回帰
- y: 応答, 従属変数
- x: 説明変数, 独立変数

- 予測 $\hat{y} = a + bx$
- 残差 $y - \hat{y}$

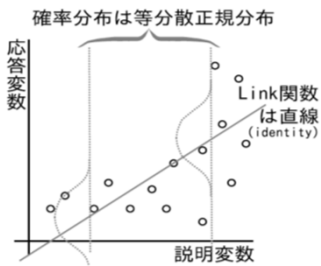
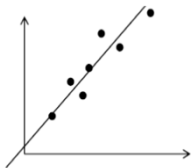


Fig.1 : (General) liner model

yという結果に対して、原因と考えられるxがどのように影響しているかを検討する手法
(例) 握力 = a + b₁×(年齢) + b₂×(性別) + b₃×(体重)

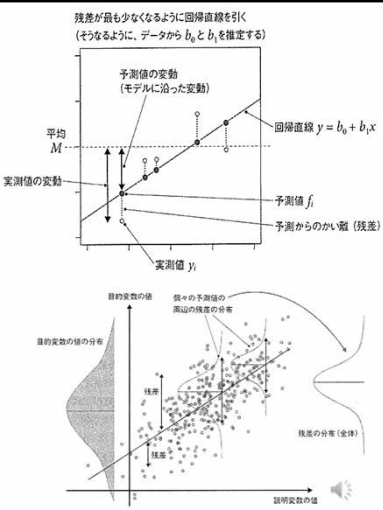
- 1. モデルのあてはまり
- 2. 変数選択によるモデル構築
- 3. モデルの解釈・評価

回帰と最小二乗法



残差が最小となるよう、係数を推定する
 $y = b_0 + b_1x_{1n} + b_px_{pn}$

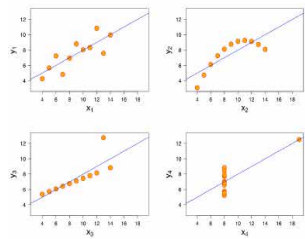
実測値(y)	予測値(\hat{y})	誤差 e_j
y_1	$b_0 + b_1x_{11} + \dots + b_px_{p1}$	$y_1 - \hat{y}_1$
y_2	$b_0 + b_1x_{12} + \dots + b_px_{p2}$	$y_2 - \hat{y}_2$
y_n	$b_0 + b_1x_{1n} + \dots + b_px_{pn}$	$y_n - \hat{y}_n$



線形モデルフィッティング
Anscombe's quartet

データセット
Ans1.csv~Ans4.csv
Ans1.sav~Ans4.sav

- 単純な記述統計量はほぼ同じだが、分布が大きく異なり、グラフにすると大きく異なって見える4つのデータセット。
- これらは、統計学者Francis Anscombeによって、データを分析する前にグラフ化することの重要性と、外れ値やその他の影響力のある観測値が統計的性質に及ぼす影響の両方を示すために、1973年に作成されたものです。

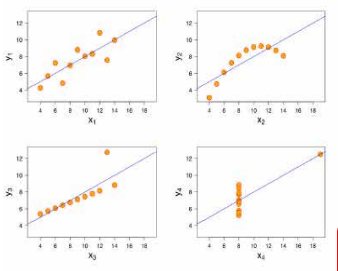


	I		II		III		IV	
	X ₁	Y ₁	X ₂	Y ₂	X ₃	Y ₃	X ₄	Y ₄
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
AVG	9	7.50	9	7.50	9	7.50	9	7.50
VAR	11	4.127	11	4.128	11	4.123	11	4.123
CORREL		0.816		0.816		0.816		0.817

線形モデルフィッティング Anscombe's quartet

データセット
Ans1.csv~Ans4.csv
Ans1.sav~Ans4.sav

形状は全く異なるが、要約統計では類似している4つのデータセット



	I	II	III	IV				
X _i	Y _i	X ₂	Y ₂	X ₃	Y ₃	X ₄	Y ₄	
10	8.04	10	9.14	10	7.46	8	6.58	
8	6.95	8	8.14	8	6.77	8	5.76	
13	7.58	13	8.74	13	12.74	8	7.71	
9	8.81	9	8.77	9	7.11	8	8.84	
11	8.33	11	9.26	11	7.81	8	8.47	
14	9.96	14	8.10	14	8.84	8	7.04	
6	7.24	6	6.13	6	6.08	8	5.25	
4	4.26	4	3.10	4	5.39	19	12.5	
12	10.84	12	9.13	12	8.15	8	5.56	
7	4.82	7	7.26	7	6.42	8	7.91	
5	5.68	5	4.74	5	5.73	8	6.89	
AVG	9	7.50	9	7.50	9	7.50	7.50	
VAR	11	4.127	11	4.128	11	4.123	11	4.123
CORREL	0.816	0.816	0.816	0.817				

モデル診断をやってみよう

線形回帰(単回帰、重回帰)

Call:
lm(formula = y ~ x, data = ANS1)

Residuals:

	Min	1Q	Median	3Q	Max
	-1.92127	-0.45577	-0.04136	0.70941	1.83882

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0001	1.1247	2.667	0.02573 *
x	0.5001	0.1179	4.241	0.00217 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6685, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF, p-value: 0.00217

分散分析
F値 17.99 P値 0.00217

```
> multreg.table
回帰係数推定値 95%信頼区間下限 95%信頼区間上限 標準偏差 t統計量
(Intercept) 3.0000909 0.4557369 5.5444449 1.1247468 2.667348
x 0.5000909 0.2333701 0.7668117 0.1179055 4.241455
P値
(Intercept) 0.025734051
x 0.002169629
```

回帰分析を行う

Rコマンド

```
##SPSSデータセットのインポート##
ANS1 <- readSPSS(
  file = "D:/大学/滋賀医大/講義/保健統計実習/ANS1.csv",
  rownames = FALSE, stringsAsFactors = TRUE, tolower = FALSE
)
```

統計分析のメニューから「線形回帰(単回帰、重回帰)」を選択する。

偏回帰係数、標準偏回帰係数

偏回帰係数

- 偏回帰係数は重回帰モデルにおける独立変数の係数である
- 他の独立変数の影響を除外した回帰係数となる

標準偏回帰係数(β)

- 平均0、分散1に標準化した単位に依存しない係数(変数間の比較が行えるようにする)
- 各独立変数が従属変数にどのくらい影響しているかを評価できる

標準偏回帰係数

- 標準偏回帰係数(A_i) a_i $Sx:x$ の標準偏差, $SDy:y$ の標準偏差

$$A_i = a_i \times \frac{SD_x}{SD_y}$$
$$= 0.50 \times (3.31 / 2.03) = 0.815$$

- 標準化したデータセットで重回帰
 - `z <- scale(Ans1)`
 - `z <- data.frame(z)`データセットZで重回帰分析を実施

回帰係数推定値

95%信頼区間下限

95%信頼区間上限

標準誤差

t統計量

P値

(Intercept)

-1.673724e-17

-0.4151695

0.4151695

0.1835281

-9.119714e-17

1.000000000

x

8.164205e-01

0.3809871

1.2518540

0.1924859

4.241455e+00

0.002169629

13

R, R², 調整済みR²

重相関係数(R)

- 重回帰式から得られる予測値と実測値の相関係数
- 1に近いほどあてはまりがよい
- 変数の数が多いと1に近づく(変数の数の影響をうける)

決定係数(R²)

- 重回帰モデルの適合性を評価する指標
- 変数の数が多いと1に近づく(変数の数の影響をうける)

自由度調整済み重相関係数・決定係数

- 独立変数の数、nを補正した指標

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

15

data=ans1

モデルの要約^b

モデル	R	R ² 調整済み	R ² 調整済み	推定値の標準誤差	Durbin-Watson
1	.816 ^a	.667	.629	1.23660	3.212

分散分析^a

モデル		平方和	自由度	平均平方	F 値	有意確率
1	回帰	27.510	1	27.510	17.990	.002 ^b
	残差	13.763	9	1.529		
	合計	41.273	10			

a. 従属変数 y

b. 予測値: (定数)、x。

係数^a

モデル		非標準化係数	標準化係数	t 値	有意確率	B の 95.0% 信頼区間	ゼロ次	相関係数	部分	共線性の統計量	VIF
		B	標準誤差	ベータ		下限	上限			許容度	
1	(定数)	3.000	1.125	2.667	.026	.456	5.544				
	x	.500	.118	.816	.424	.233	.767	.816	.816	1.000	1.000

a. 従属変数 y

Scatter plot of y1 vs x1

4

lm(y ~ x)

ID3, 9, 10あたりが予測から乖離している

残差

- 予測値の大小によって残差の分布に違いがないかを見る
- 0を中心に上下に均等に分布しているかを確認する

Normal Q-Q

- 残差が正規分布であれば直線になる

Scale-Location

- 残差をその標準偏差推定値で割ったもの。
- 値が±3を超えている場合は異常値である。

Residuals vs Leverage

- LEVERAGE
 - 説明変数空間における外れ値
- Cook's D
 - 問題のオブザベーションを除外したときのリグレッションの変化
 - 4/nを超える場合は注意

Cook' Dでは、ID3, 9の影響度が大きい

Variance inflation factor (VIF)

従属変数y, 独立変数x₁, x₂, x₃, x₄で構築された重回帰モデルがあったとして、
x₁を従属変数、残りのx₂, x₃, x₄を独立変数とした
重回帰モデルを構築し、重相関係数を求める

VIF=1 / (1 - R²) (分母は“許容度”とよばれる)

VIF ≧ 10 となるような変数は除いた方がよい

17

data=ans2

モデルの要約^b

モデル	R	R ² 乗	調整済みR ² 乗	推定値の標準誤差	Durbin-Watson
1	.816 ^a	.666	.629	1.23721	2.188

a. 予測値: (定数), x₂
b. 従属変数 y

分散分析^a

モデル		平方和	自由度	平均平方	F 値	有意確率
1	回帰	27.500	1	27.500	17.966	.002 ^b
	残差	13.776	9	1.531		
	合計	41.276	10			

a. 従属変数 y
b. 予測値: (定数), x₂

係数^a

モデル		非標準化係数 B	標準誤差	標準化係数 ベータ	t 値	有意確率	B の 95.0% 信頼区間 下限	上限
1	(定数)	3.001	1.125		2.667	.026	.455	5.547
	x ₂	.500	.118	.816	4.239	.002	.233	.767

a. 従属変数 y

19

参考: 影響度・残差関連の指標

MODEL Option or Statistic	Formula	
PRED (Y _i)	X _i b	予測値
RES (r _i)	Y _i - Y _i	残差
H (h _i)	x _i (X'X) ⁻¹ x _i	デコ比: 説明変数空間での外れ度、何個の 回帰係数を決定しているか<1
STDP	√h _i σ ²	
STDI	√(1+h _i)σ ²	
STDR	√(1-h _i)σ ²	
LCL	Ŷ _i - t _α STDI	
LCLM	Ŷ _i - t _α STDP	
UCL	Ŷ _i + t _α STDI	
UCLM	Ŷ _i + t _α STDP	
STUDENT	t _i	標準化した残差
RSTUDENT	STDR _i t _i	標準化した残差 (分散を当該obsを除いて計算)
COOKD	1 / (1 - STUDENT ²) * STDP ² / STDR ²	Cook's D (当該obsを除いたときの あてはめ結果の全体としての変化)
COVRATIO	det((σ ² * (X _(i) X _(i)) ⁻¹)) / det((σ ² * (X'X) ⁻¹))	
DFFITs	(Y _i - Ŷ _i) / (σ _(i) √h _i)	これを2乗しpで割り分散を置き換えたものがCook's D
DFBETAS _j	b _j - b _{j(i)} / σ _(i) √((X'X) _{jj}) ⁻¹	当該obsを除いた時の個々の回帰係数の変化
PRESS(pred _r)	t _i / 1 - h _i	

18

lm(y ~ x)

```
Call:
lm(formula = y ~ x, data = ANS2)

Residuals:
    Min       1Q   Median       3Q      Max
-1.9009 -0.7608  0.1291  0.9491  1.2891

Coefficients:
(Intercept)   3.001    1.125    2.667  0.02576 *
x             0.500    0.118    4.239  0.00218 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002178
```

20

data=ans3

モデルの要約^b

モデル	R	R2乗	調整済みR2乗	推定値の標準誤差	Durbin-Watson
1	.816 ^a	.666	.629	1.23631	2.144

a. 予測値: (定数)、 x_0
b. 従属変数 y

分散分析^a

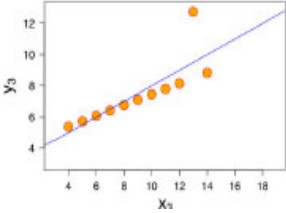
モデル		平方和	自由度	平均平方	F値	有意確率
1	回帰	27.470	1	27.470	17.972	.002 ^b
	残差	13.756	9	1.528		
	合計	41.226	10			

a. 従属変数 y
b. 予測値: (定数)、 x_0

係数^a

モデル		非標準化係数 B	標準化係数 標準誤差 ベータ	t値	有意確率	Bの95.0%信頼区間 下限 上限	ゼロ次 相関 偏	部分 相関	共線性の統計量 許容度 VIF
1	(定数)	3.002	1.124	2.670	.026	.459 5.546			
	x	.500	.118	.816	.423	.233 .766	.816	.816	1.000 1.000

a. 従属変数 y



data=ans4

モデルの要約^b

モデル	R	R2乗	調整済みR2乗	推定値の標準誤差	Durbin-Watson
1	.817 ^a	.667	.630	1.23570	1.662

a. 予測値: (定数)、 x_0
b. 従属変数 y

分散分析^a

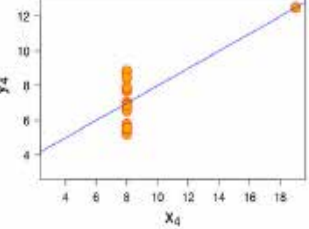
モデル		平方和	自由度	平均平方	F値	有意確率
1	回帰	27.490	1	27.490	18.003	.002 ^b
	残差	13.742	9	1.527		
	合計	41.232	10			

a. 従属変数 y
b. 予測値: (定数)、 x_0

係数^a

モデル		非標準化係数 B	標準化係数 標準誤差 ベータ	t値	有意確率	Bの95.0%信頼区間 下限 上限	ゼロ次 相関 偏	部分 相関	共線性の統計量 許容度 VIF
1	(定数)	3.002	1.124	2.671	.026	.459 5.544			
	x	.500	.118	.817	.423	.233 .766	.817	.817	1.000 1.000

a. 従属変数 y



```
Call:
lm(formula = y ~ x, data = ANS3)

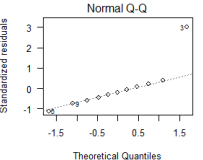
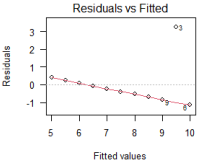
Residuals:
    Min       1Q   Median       3Q      Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

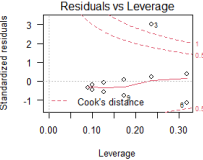
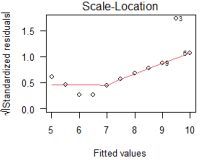
Coefficients:
(Intercept)  3.0025  1.1245  2.670  0.02562 *
x            0.4997  0.1179  4.239  0.00218 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002176
```

lm(y ~ x)





```
Call:
lm(formula = y ~ x, data = ANS4)

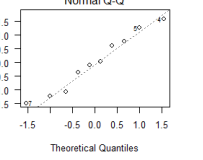
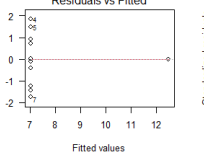
Residuals:
    Min       1Q   Median       3Q      Max
-1.751 -0.831  0.000  0.809  1.839

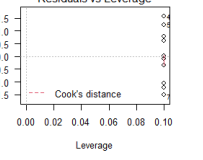
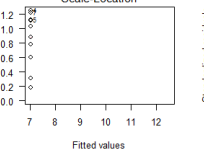
Coefficients:
(Intercept)  3.0017  1.1239  2.671  0.02559 *
x            0.4999  0.1178  4.243  0.00216 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
F-statistic: 18 on 1 and 9 DF, p-value: 0.002165
```

lm(y ~ x)





ans1~ans4

- 4つの回帰について、これだけ異なるデータの集まりだが、同じような直線が引けてしまう
 - グラフを見ずに、平均、回帰係数などの数値だけみるとほぼ同じ
- 残差に系統的なパターンが残されているなら、さらにモデル化すべき

25

1. モデルのあてはまり

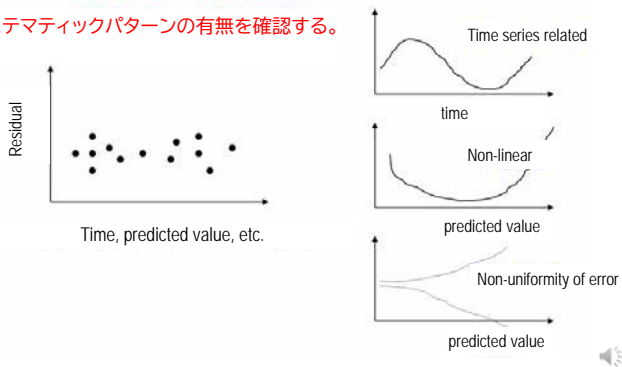
2. 変数選択によるモデル構築

3. モデルの解釈・評価

27

残差分析

システマティックパターンの有無を確認する。



26

変数選択

相関の高い変数をモデルに含めると係数が不安定になる(多重共線性)。

<対策>

① 変数の選択

ALL(全変数を一度に入力)
FORWARD(前方、増加)
BACKWARD(後方、減少)
STEPWISE(増加/減少)を使用する。

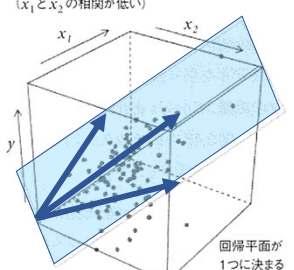
② 合成変数の作成

③ 事前情報の利用(ベイズ型)

28

多重共線性

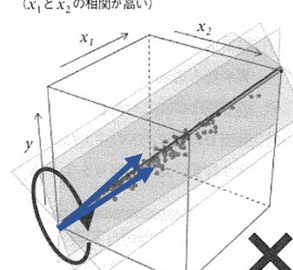
■ 通常の回帰分析
(x_1 と x_2 の相関が低い)



回帰平面が1つに決まる

$$y = b_0 + b_1x_1 + b_2x_2$$

■ 多重共線性がある状態
(x_1 と x_2 の相関が高い)



データが直線のまわりに分布
⇒ 回帰平面が安定しない

図3.46 多重共線性

29

線形回帰(単回帰、重回帰)

モデル名を入力: RegModel.5
複数の変数はCtrlキーを押しながらクリック。
目的変数 (1つ選択)
説明変数 (1つ以上選択)

Call:
lm(formula = y ~ x1 + x2 + x3, data = TempDF)

Residuals:

Min 1Q Median 3Q Max
-2.4259 -0.7411 0.1107 0.6243 2.4522

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0003582	0.1077994	0.003	0.997
x1	1.0127592	0.1134334	8.928	2.98e-14 ***
x2	1.0043676	0.1057964	9.493	1.82e-15 ***
x3	0.8505119	0.1024469	8.302	6.50e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.053 on 96 degrees of freedom
Multiple R-squared: 0.7171, Adjusted R-squared: 0.7083
F-statistic: 81.13 on 3 and 96 DF, p-value: < 2.2e-16

31

演習 変数選択

データセット
select.csv
select.sav

“select.csv”データは、 $y=x_1+x_2+x_3$ +誤差(x_1, x_2, x_3 は独立)というモデルで作られています。

- x_4 は x_1 と高い相関がある
- x_5 は x_2 と高い相関がある
- x_6 - x_{15} は互いに独立であり、 y とも独立である。

変数選択を適用して、どのような結果が得られるか確認してください。

■ x_1 - x_{15} を用いてモデルを作成する。相関の高い変数を含む場合、どのような変数が選択されるのか

30

線形回帰(単回帰、重回帰)

モデル名を入力: RegModel.5
複数の変数はCtrlキーを押しながらクリック。
目的変数 (1つ選択)
説明変数 (1つ以上選択)

Call:
lm(formula = y ~ x1 + x2 + x3, data = TempDF)

Residuals:

Min 1Q Median 3Q Max
-2.4259 -0.7411 0.1107 0.6243 2.4522

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0003582	0.1077994	0.003	0.997
x1	1.0127592	0.1134334	8.928	2.98e-14 ***
x2	1.0043676	0.1057964	9.493	1.82e-15 ***
x3	0.8505119	0.1024469	8.302	6.50e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.053 on 96 degrees of freedom
Multiple R-squared: 0.7171, Adjusted R-squared: 0.7083
F-statistic: 81.13 on 3 and 96 DF, p-value: < 2.2e-16

32

重回帰分析 注意点

1. 目的の再確認

1) 予測式を算出したい
→ 決定係数や残差の検討を詳細に

2) パラメータ値の解釈(従属変数に対する独立変数の影響)
→ 偏回帰係数の有意性を重視してモデルを構築する
→ 仮説に合わせた変数の投入(強制投入)

2. 標本数と独立変数の数

独立変数1つに対し $n \geq 20 \sim 30$ 程度が目安

3. 外れ値のチェック

4. 独立変数間の相関

5. 正規分布からのズレが大きい場合は変数変換(対数変換など)

33

一般線形モデル
一般化線形モデル

Development of linear models

階層ベイズモデル

より柔軟な統計モデリング

一般化線形混合モデル

個人差や部位差などのランダム効果への対応

一般化線形モデル

正規分布以外の確率分布を扱う

一般線形モデル

Logistic Poisson

Linear regression Analysis of Variance

最小二乗法

Statistical Inference MCMC

最尤法

35

第8回
ロジスティック回帰

34

一般線形モデル→一般化線形モデル

GLMは正規分布だけでなく、指数族(二項、ガンマ、ポアソンなど)の分布も扱うことができます！

確率分布は正規分布

Response variable

Explanatory variables

Link Function : identity

Fig.1 : 一般線形モデル

確率分布はポアソン分布

Response variable

Explanatory variables

Link function: log

正規分布を超えて拡張することを検討する

Fig.2 : 一般化線形モデル(例:ポアソン)

36

GLMは以下の3つの要素で構成される

1. 確率分布の指数族

指数族(二項、ガンマ、ポアソンなど)

2. 線形予測量

$\eta = X\beta$

線形予測量(η)は独立変数の情報をモデルに取り込む量である。
 η は未知パラメータ β の線形結合(したがって「線形」)で表され、線形結合の係数は独立変数の行列Xとして表される。

3. リンク関数

ある式を線形に変換する関数。

37

回帰モデルの拡張

- 0-1データに対するロジスティック回帰
- 頻度データに対するポアソン回帰
- 生存時間データに対するCox回帰
- 一般化推定方程式(Generalized Estimating Equation)
 - 反復測定データに対するGEE
- 個人差、階層、繰り返しを扱うことができるモデル(混合モデル)
 - 線型:
 - 一般化線型:
 - 非線型:

39

Link function

- 式を線形に変換する関数

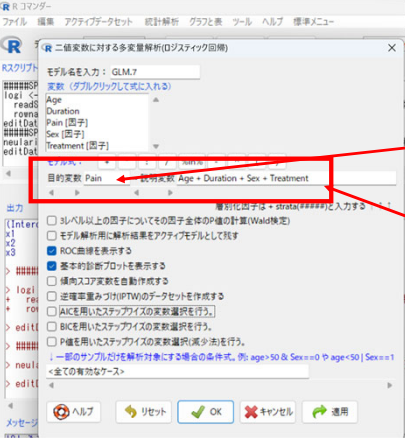
Distribution	Support	Discrete variable	Continuous variable	Probability Distribution	Mean function
Normal	real: $(-\infty, +\infty)$				μ
Exponential	real: $(0, +\infty)$				μ
Gamma	real: $(0, +\infty)$				μ
Inverse Gaussian	real: $(0, +\infty)$				μ
Poisson	integer: $0, 1, 2, \dots$	Bernoulli		binomial	$\mu(1-\mu)$
		Binomial		binomial	$\mu(1-\mu)$
		Poisson		poisson	μ
		Gamma		gamma	μ^2
		Normal		gaussian	identity

Diagram: glm() の family (circled) points to $\text{frequently used link functions}$ (circled). The link functions listed are: logit, logit, log, log, identity.

38

ロジスティック回帰

40



• 式を作成

Pain(あり・なし)

年齢、期間、治療

41

Logistic regression

```
Coefficients:
(Intercept) 15.574390 8.591008 2.383 0.01813 *
Age -0.262093 0.097012 -2.702 0.00690 **
Duration 0.005859 0.032992 0.178 0.85905
Sex[T.F] 1.832202 0.796206 2.301 0.02138 *
Treatment[T.B] 3.708542 1.140577 3.251 0.00115 **
Treatment[T.A] 3.181690 1.016021 3.132 0.00174 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 81.503  on 59  degrees of freedom
Residual deviance: 48.736  on 54  degrees of freedom
AIC: 60.736

> odds
      odds
Age      0.769
Duration 1.010
Sex[T.F] 6.250
Treatment[T.B] 40.800
Treatment[T.A] 24.100

Number of Fisher Scoring iterations: 5
```

$\ln\left(\frac{p}{1-p}\right) = a + bX$

$\frac{p}{1-p} = e^{a+bX}$

odds

オッズ比 95%信頼区間下限 95%信頼区間上限 P値

(Intercept)	5810000.000	14.200	2.37e+12	0.01810
Age	0.769	0.636	9.31e-01	0.00690
Duration	1.010	0.943	1.07e+00	0.85900
Sex[T.F]	6.250	1.310	2.97e+01	0.02140
Treatment[T.B]	40.800	4.360	3.81e+02	0.00115
Treatment[T.A]	24.100	3.290	1.76e+02	0.00174

$e^{5.1817} = 24.0974$

$e^{3.7085} = 40.7942$

43

Logistic regression

ロジスティック回帰では、従属変数はオッズの自然対数であるロジットである、

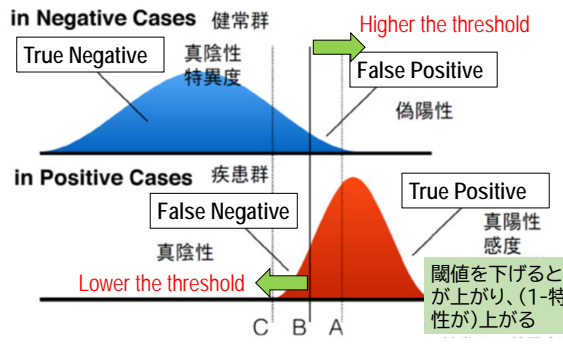
$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

ロジットとはオッズの対数であり、オッズはPの関数である。ロジスティック回帰では、次のようになる。

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = a + bX$$
$$\frac{P}{1-P} = e^{a+bX}$$
$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

42

ROC(Receiver Operating Characteristic)



in Negative Cases 健康群

True Negative 真陰性

False Positive 偽陽性

Higher the threshold

in Positive Cases 疾患群

False Negative 偽陰性

True Positive 真陽性

閾値を下げると感度が上がり、(1-特異性)が上がる

44

感度・特異度

		Disease		Predictive Value	
		⊕	⊖		
Test	⊕	A True Positive (TP)	B False Positive (FP)	Positive Predictive Value (PPV) $\frac{TP}{TP + FP} = \frac{A}{A + B}$	Total Positive Results (A + B)
	⊖	C False Negative (FN)	D True Negative (TN)	Negative Predictive Value (NPV) $\frac{TN}{FN + TN} = \frac{D}{C + D}$	Total Negative Results (C + D)
Sensitivity & Specificity		Sensitivity $\frac{TP}{TP + FN} = \frac{A}{A + C}$	Specificity $\frac{TN}{FP + TN} = \frac{D}{B + D}$		
		All diseased patients (A + C)	All non-diseased patients (B + D)		

疾病を有する患者において真陽性となる割合

↑

病気を持っていない患者さんの真の陰性結果の割合を示しています。

45

ROC(Receiver Operating Characteristic)

47

陽性（陰性）的中度

		Disease		Predictive Value	
		⊕	⊖		
Test	⊕	A True Positive (TP)	B False Positive (FP)	Positive Predictive Value (PPV) $\frac{TP}{TP + FP} = \frac{A}{A + B}$	Total Positive Results (A + B)
	⊖	C False Negative (FN)	D True Negative (TN)	Negative Predictive Value (NPV) $\frac{TN}{FN + TN} = \frac{D}{C + D}$	Total Negative Results (C + D)
Sensitivity & Specificity		Sensitivity $\frac{TP}{TP + FN} = \frac{A}{A + C}$	Specificity $\frac{TN}{FP + TN} = \frac{D}{B + D}$		
		All diseased patients (A + C)	All non-diseased patients (B + D)		

陽性の場合、病気ありの割合

←

陰性で、病気でない割合

46

ROC(Receiver Operating Characteristic)

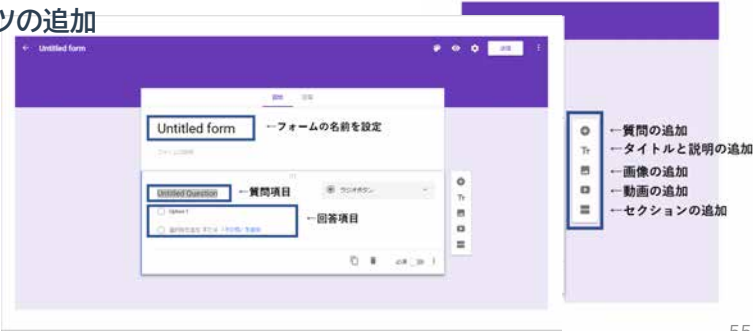
48

第9回
調査データ解析(2):Web調査ツールを使用した調査表作成・データ入力 データセット比較

53

Googleフォームの作り方

- 1. Googleアカウントを作成
- 2. テンプレートの選択
- 3. パーツの追加



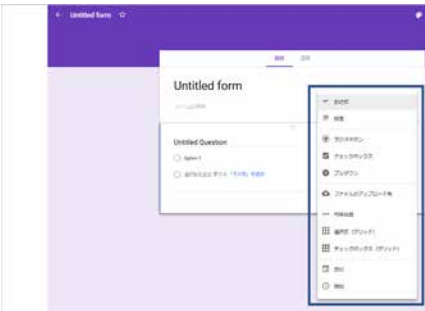
55

Googleフォームとは

- Googleフォームとは、Googleのサービスの1つとして提供されているフォーム作成ツール
- 使いやすさや無料で使える点などから、アンケートフォームや問い合わせフォーム、キャンペーンへの申し込みフォームなど、様々な用途で利用されている
- フォームを簡単に作成できるだけでなく、集計や分析をアシストする機能もある

54

4. 設問 / 選択肢の作成

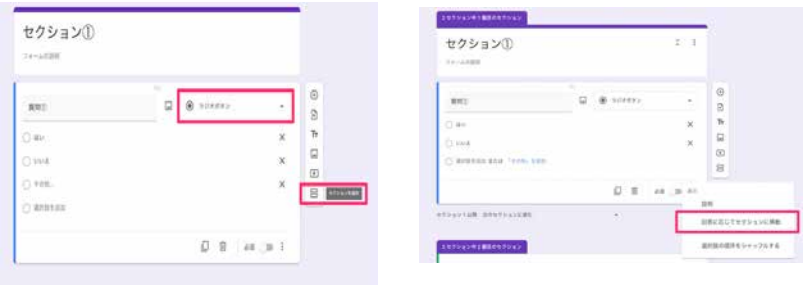


- 記述式…一文程度の短い文章を記載してもらうとき
- 段落式…段落をわけるとような長い文章を記載してもらうとき
- ラジオボタン…複数の回答から1つの回答を選択してもらうとき
- チェックボックス…複数の回答から1つ以上の回答を選択してもらうとき
- プルダウン…プルダウン形式で回答を選択してもらうとき
- 均等目盛…例えば1～5などの段階にわけて評価を知りたいとき

56

<条件分岐>

- 条件分岐をするには、まず「ラジオボタン」「プルダウン」のどちらかの形式で質問を作成しましょう。作成できたら、【セクションを追加】をクリック



61

調査項目例

人格特性：日本語版 Ten Item Personality Inventory (TIPI-J)

- 10問7段階評価(1～7点)
- 外向性、協調性、勤勉性、神経症傾向、開放性を評価
- 得点が高いほど各項目が強い

クロノタイプ(朝方夜型)

・19問

【クロノタイプ判定】

[70～86点]明らかな朝型 [59～69点]ほぼ朝型 [42～58点]中間型 [31～41点]ほぼ夜型 [16～30点]明らかな夜型

アテネ睡眠調査票

・8問(0～3)

・[1～3点]・・・睡眠がとれています [4～5点]・・・不眠症の疑いが少しあります [6点以上]・・・不眠症の可能性が高いです

<スプレッドシートへエクスポート>

- スプレッドシートと連携させてグラフを調整することも可能



Googleフォームに回答がくるたびに、回答がスプレッドシートにリアルタイムで反映されていきます

62

フォームを作成し回答しあう

- 条件
 - 数値情報を調査する項目を設定(仮想でも良い)
 - 自由記載の欄を作成
 - 必ず入力してもらう



64

Google Formの調査データ のデータセット化

65

睡眠調査データ

- 睡眠調査データを配布します
 - データ(1)とデータ(2)があります
 - データ(1)を原本とします。データ(2)は同じデータを入力したファイルです
(注: 早めにデータ回収を打ち切り、重複送信をそのままにしてあるという設定。
二つはデータが一致しない可能性がある)。
- この二つのデータが一致しているか、確認してもらいます
 - いくつか入力間違いも作っています。

67

- 睡眠調査データを比較(compare)してみよう
 - どの部分の入力が間違っているか特定しよう



66

COMPARE

例題
ae.sav
ae2.sav
Sleep.q.csv
Sleep.q2.csv

データセットを比較するには？
調査後にデータを入力した場合、入力間違いをしている可能性があります。

<見つける方法>

- データクリーニングで判別する場合もある
(レンジチェック、外れ値などのチェック)
- 2回入力(あるいは他の人が入力したもの)と比較
⇒臨床試験などではこのような方法をとります
⇒ダブルエントリーといいます。
- この二つのデータセットを比較する方法を行います。
 - R studioを使用します。
 - 「arsenal」というパッケージを使用します。



68

R-studio

- 「ae」と「ae2」というデータセットを読み込む
- 「arsenal」というパッケージを導入

```
ae <-  
read_sav('https://raw.githubusercontent.com/harabou/Biostat_Kyoto_pref/main/  
data/%2303/compare/ae.csv')  
  
ae2 <-  
read_sav('https://raw.githubusercontent.com/harabou/Biostat_Kyoto_pref/main/  
data/%2303/compare/ae2.csv')  
  
install.packages("arsenal")  
library(arsenal)  
comparedf(ae, ae2)  
summary(comparedf(ae, ae2))
```

69

Table: Differences detected by variable

var.x	var.y	n	NAs
ID	ID	0	0
ae	ae	4	0
AESTD	AESTD	0	0
SER	SER	0	0
GRADE	GRADE	0	0

Table: Differences detected

var.x	var.y	..row.names..	values.x	values.y	row.x	row.y
ae	ae		5	下痢	5	5
ae	ae		6	下痢	6	6
ae	ae		11	発熱	11	11
ae	ae		14	疲労	14	14

71

COMPARE 手順

Table: Summary of data.frames version arg ncol nrow


x . 5 86

y ae2 5 86

[データ]—[データセットの比較]

Table: Summary of overall comparison


statistic	value
Number of by-variables	0
Number of non-by variables in common	5
Number of variables in x but not y	5
Number of variables compared	0
Number of variables in y but not x	0
Number of variables compared with some values unequal	1
Number of variables compared with all values equal	4
Number of observations in common	86
Number of observations in x but not y	0
Number of observations in y but not x	0
Number of observations with some compared variables unequal	4
Number of observations with all compared variables equal	82
Number of values unequal	4



72

COMPARE 結果

原本
Aさんの入力



Bさんの入力

ID	ae	AESTD	SER	GRADE	CasesCompare	ID	ae	AESTD	SER	GRADE
A-001	頭痛	4/24/2006	非重篤	軽度	0	A-001	頭痛	4/24/2007	非重篤	軽度
A-002	発熱	12/31/2006	非重篤	軽度	0	A-002	発熱	12/31/2006	非重篤	軽度
A-003	発熱	2/14/2007	非重篤	軽度	0	A-003	発熱	2/14/2007	非重篤	軽度
A-004	腹痛	2/9/2007	非重篤	軽度	0	A-004	腹痛	2/9/2007	非重篤	軽度
A-005	下痢	1/20/2006	非重篤	軽度	1	A-005	腹痛	1/20/2006	非重篤	軽度
A-005	下痢	12/18/2006	非重篤	軽度	1	A-005	腹痛	12/18/2006	非重篤	軽度
A-006	発熱	8/1/2006	非重篤	軽度	0	A-006	発熱	8/1/2006	非重篤	軽度
A-007	頭痛	4/11/2006	非重篤	軽度	0	A-007	頭痛	4/11/2006	非重篤	軽度
A-008	下痢	3/4/2006	非重篤	軽度	0	A-008	下痢	3/4/2006	非重篤	軽度
A-008	発熱	3/4/2006	非重篤	軽度	0	A-008	発熱	3/4/2006	非重篤	軽度
A-008	発熱	7/6/2006	非重篤	軽度	1	A-008	腹痛	7/6/2006	非重篤	軽度
A-008	腹痛	3/4/2006	非重篤	軽度	0	A-008	腹痛	3/4/2006	非重篤	軽度
A-009	頭痛	8/10/2006	非重篤	軽度	0	A-009	頭痛	8/10/2006	非重篤	軽度
A-010	疲労	7/1/2006	非重篤	軽度	1	A-010	疲労腹痛	7/1/2006	非重篤	軽度

原情報（紙の調査票、診療録情報など）から手入力する場合、シングルエントリーでは間違っている可能性があるため、ダブルエントリーで正確な入力情報を得る

72

第3日目 課題

1) 各自の調査票

- ・調査票を作成
- ・リンク先を入力
- ・コード表の作成

2) 睡眠調査データ(配布)

これまでの演習でやってきたことを生かして、集計を行う

- ・基本集計 : EZR「記述統計」

余力があれば: 分析: クロノタイプ(朝型・夜型)と睡眠スコア